SOLID STATE PHYSICS

PART I

Transport Properties of Solids

M. S. Dresselhaus

6.732
Fall, 2001

Oded Rabin – Head TA; Room 13-3025
Marcie Black – TA assistant; Room 13-3041
Yu-Ming Lin – TA assistant; Room 13-3037
Laura Doughty – Support; Room 13-3005
Lectures: MWF 9-10
Room 13-4101
Recitation: F 11-12
Room 38-136
10 problem sets
3 quizzes

- Part I
  Transport
- Part II
  Optical
- Part III
  Magnetism
- Part IV
  Superconductivity

Contents

Chapter 1

Review of Energy Dispersion Relations in Solids

References:

- Ashcroft and Mermin, *Solid State Physics*, Holt, Rinehart and Winston, 1976, Chapters 8, 9, 10, 11.
- Bassani and Parravicini, *Electronic States and Optical Transitions in Solids*, Pergamon, 1975, Chapter 3.
- Kittel, *Introduction to Solid State Physics*, Wiley, 1986, pp. 228-239.
- Mott & Jones – *The Theory of the Properties of Metals and Alloys*, Dover, 1958 pp. 56–85.
- Omar, *Elementary Solid State Physics*, Addison–Wesley, 1975, pp. 189–210.
- Ziman, *Principles of the Theory of Solids*, Cambridge, 1972, Chapter 3.

1.1 Introduction

The transport properties of solids are closely related to the energy dispersion relations $E(\vec{k})$ in these materials and in particular to the behavior of $E(\vec{k})$ near the Fermi level. Conversely, the analysis of transport measurements provides a great deal of information on $E(\vec{k})$. Although transport measurements do not generally provide the most sensitive tool for studying $E(\vec{k})$, such measurements are fundamental to solid state physics because they can be carried out on nearly all materials and therefore provide a valuable tool for characterizing materials. To provide the necessary background for the discussion of transport properties, we give here a brief review of the energy dispersion relations $E(\vec{k})$ in solids. In this connection, we consider in Chapter 1 the two limiting cases of weak and tight binding. In Chapter 2 we will discuss $E(\vec{k})$ for real solids including prototype metals, semiconductors, semimetals and insulators.

1.2 One Electron $E(\vec{k})$ in Solids

1.2.1 Weak Binding or Nearly Free Electron Approximation

In the weak binding approximation, we assume that the periodic potential $V(\vec{r}) = V(\vec{r}+\vec{R_n})$ is sufficiently weak so that the electrons behave almost as if they were free and the effect of the periodic potential can be handled in perturbation theory (see Appendix A). In this formulation $V(\vec{r})$ can be an *arbitrary* periodic potential. The weak binding approximation has achieved some success in describing the valence electrons in metals. For the core electrons, however, the potential energy is comparable with the kinetic energy so that core electrons are tightly bound and the weak binding approximation is not applicable. In the weak binding approximation we solve the Schrödinger equation in the limit of a very weak periodic potential

$$\mathcal{H}\psi = E\psi. \tag{1.1}$$

Using time–independent perturbation theory (see Appendix A) we write

$$E(\vec{k}) = E^{(0)}(\vec{k}) + E^{(1)}(\vec{k}) + E^{(2)}(\vec{k}) + ... \tag{1.2}$$

and take the unperturbed solution to correspond to $V(\vec{r}) = 0$ so that $E^{(0)}(\vec{k})$ is the plane wave solution

$$E^{(0)}(\vec{k}) = \frac{\hbar^2 k^2}{2m}. \tag{1.3}$$

The corresponding normalized eigenfunctions are the plane wave states

$$\psi_{\vec{k}}^{(0)}(\vec{r}) = \frac{e^{i\vec{k}\cdot\vec{r}}}{\Omega^{1/2}} \tag{1.4}$$

in which Ω is the volume of the crystal.

The first order correction to the energy $E^{(1)}(\vec{k})$ is the diagonal matrix element of the perturbation potential taken between the unperturbed states:

$$\begin{aligned} E^{(1)}(\vec{k}) &= \langle \psi_{\vec{k}}^{(0)} \mid V(\vec{r}) \mid \psi_{\vec{k}}^{(0)} \rangle = \tfrac{1}{\Omega}\textstyle\int_\Omega e^{-i\vec{k}\cdot\vec{r}} V(\vec{r}) e^{i\vec{k}\cdot\vec{r}} d^3r \\ &= \tfrac{1}{\Omega_0}\textstyle\int_{\Omega_0} V(\vec{r}) d^3r = \overline{V(\vec{r})} \end{aligned} \tag{1.5}$$

where $\overline{V(\vec{r})}$ is independent of $\vec{k}$, and Ω_0 is the volume of the unit cell. Thus, in first order perturbation theory, we merely add a constant energy $\overline{V(\vec{r})}$ to the free particle energy, and that constant term is exactly the mean potential energy seen by the electron, averaged over the unit cell. The terms of interest arise in second order perturbation theory and are

$$E^{(2)}(\vec{k}) = \sum_{\vec{k}'}{}' \frac{|\langle \vec{k}'|V(\vec{r})|\vec{k}\rangle|^2}{E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}')} \tag{1.6}$$

where the prime on the summation indicates that $\vec{k}' \neq \vec{k}$. We next compute the matrix element $\langle \vec{k}'|V(\vec{r})|\vec{k}\rangle$ as follows:

$$\begin{aligned} \langle \vec{k}'|V(\vec{r})|\vec{k}\rangle &= \textstyle\int_\Omega \psi_{\vec{k}'}^{(0)*} V(\vec{r}) \psi_{\vec{k}}^{(0)} d^3r \\ &= \tfrac{1}{\Omega}\textstyle\int_\Omega e^{-i(\vec{k}'-\vec{k})\cdot\vec{r}} V(\vec{r}) d^3r \\ &= \tfrac{1}{\Omega}\textstyle\int_\Omega e^{i\vec{q}\cdot\vec{r}} V(\vec{r}) d^3r \end{aligned} \tag{1.7}$$

where $\vec{q}$ is the difference wave vector $\vec{q} = \vec{k} - \vec{k}'$ and the integration is over the whole crystal. We now exploit the periodicity of $V(\vec{r})$. Let $\vec{r} = \vec{r}\,' + \vec{R}_n$ where $\vec{r}\,'$ is an arbitrary vector in a unit cell and $\vec{R}_n$ is a lattice vector. Then since $V(\vec{r}) = V(\vec{r}\,')$

$$\langle \vec{k}'|V(\vec{r})|\vec{k}\rangle = \frac{1}{\Omega}\sum_n \int_{\Omega_0} e^{i\vec{q}\cdot(\vec{r}\,'+\vec{R}_n)} V(\vec{r}\,')d^3r' \tag{1.8}$$

where the sum is over unit cells and the integration is over the volume of one unit cell. Then

$$\langle \vec{k}'|V(\vec{r})|\vec{k}\rangle = \frac{1}{\Omega}\sum_n e^{i\vec{q}\cdot\vec{R}_n} \int_{\Omega_0} e^{i\vec{q}\cdot\vec{r}\,'} V(\vec{r}\,')d^3r'. \tag{1.9}$$

Writing the following expressions for the lattice vectors $\vec{R}_n$ and for the wave vector $\vec{q}$

$$\begin{aligned} \vec{R}_n &= \textstyle\sum_{j=1}^3 n_j \vec{a}_j \\ \vec{q} &= \textstyle\sum_{j=1}^3 \alpha_j \vec{b}_j \end{aligned} \tag{1.10}$$

where n_j is an integer, then the lattice sum $\sum_n e^{i\vec{q}\cdot\vec{R}_n}$ can be carried out exactly to yield

$$\sum_n e^{i\vec{q}\cdot\vec{R}_n} = \left[\prod_{j=1}^3 \frac{1-e^{2\pi i N_j \alpha_j}}{1-e^{2\pi i \alpha_j}}\right] \tag{1.11}$$

where $N = N_1N_2N_3$ is the total number of unit cells in the crystal and α_j is a real number. This sum fluctuates wildly as $\vec{q}$ varies and is appreciable only if

$$\vec{q} = \sum_{j=1}^3 m_j \vec{b}_j \tag{1.12}$$

where m_j is an integer and $\vec{b}_j$ is a primitive vector in reciprocal space, so that $\vec{q}$ must be a reciprocal lattice vector. Hence we have

$$\sum_n e^{i\vec{q}\cdot\vec{R}_n} = N\delta_{\vec{q},\vec{G}} \tag{1.13}$$

since $\vec{b}_j \cdot \vec{R}_n = 2\pi l_{jn}$ where l_{jn} is an integer.

This discussion shows that the matrix element $\langle \vec{k}'|V(\vec{r})|\vec{k}\rangle$ is only important when $\vec{q} = \vec{G}$ is a reciprocal lattice vector $= \vec{k} - \vec{k}'$ from which we conclude that the periodic potential $V(\vec{r})$ only connects wave vectors $\vec{k}$ and $\vec{k}'$ separated by a reciprocal lattice vector. We note that this is the same relation that determines the Brillouin zone boundary. The matrix element is then

$$\langle \vec{k}'|V(\vec{r})|\vec{k}\rangle = \frac{N}{\Omega}\int_{\Omega_0} e^{i\vec{G}\cdot\vec{r}\,'} V(\vec{r}\,')d^3r\,'\delta_{\vec{k}'-\vec{k},\vec{G}} \tag{1.14}$$

where

$$\frac{N}{\Omega} = \frac{1}{\Omega_0} \tag{1.15}$$

and the integration is over the unit cell. We introduce $V_{\vec{G}}$ = Fourier coefficient of $V(\vec{r})$ where

$$V_{\vec{G}} = \frac{1}{\Omega_0}\int_{\Omega_0} e^{i\vec{G}\cdot\vec{r}\,'} V(\vec{r}\,')d^3r\,' \tag{1.16}$$

so that

$$\langle \vec{k}'|V(\vec{r})|\vec{k}\rangle = \delta_{\vec{k}-\vec{k}',\vec{G}}\, V_{\vec{G}}. \tag{1.17}$$

We can now use this matrix element to calculate the 2^{nd} order change in the energy

$$E^{(2)}(\vec{k}) = \sum_{\vec{G}} \frac{|V_{\vec{G}}|^2}{k^2-(k')^2}\left(\frac{2m}{\hbar^2}\right) = \frac{2m}{\hbar^2}\sum_{\vec{G}} \frac{|V_{\vec{G}}|^2}{k^2-(\vec{G}+\vec{k})^2}. \tag{1.18}$$

We observe that when $k^2 = (\vec{G}+\vec{k})^2$ the denominator vanishes and $E^{(2)}(\vec{k})$ can become very large. This condition is identical with the Laue diffraction condition. Thus, at a Brillouin zone boundary, the weak perturbing potential has a very large effect and therefore non–degenerate perturbation theory will not work in this case.

For $\vec{k}$ values near a Brillouin zone boundary, we must then use degenerate perturbation theory (see Appendix A). Since the matrix elements coupling the plane wave states $\vec{k}$ and $\vec{k}+\vec{G}$ do not vanish, *first order degenerate* perturbation theory is sufficient and leads to the determinantal equation

$$\begin{vmatrix} E^{(0)}(\vec{k}) + E^{(1)}(\vec{k}) - E & \langle \vec{k}+\vec{G}|V(\vec{r})|\vec{k}\rangle \\ \langle \vec{k}|V(\vec{r})|\vec{k}+\vec{G}\rangle & E^{(0)}(\vec{k}+\vec{G}) + E^{(1)}(\vec{k}+\vec{G}) - E \end{vmatrix} = 0 \tag{1.19}$$

in which

$$\begin{aligned} E^{(0)}(\vec{k}) &= \tfrac{\hbar^2 k^2}{2m} \\ E^{(0)}(\vec{k}+\vec{G}) &= \tfrac{\hbar^2(\vec{k}+\vec{G})^2}{2m} \end{aligned} \tag{1.20}$$

and

$$\begin{aligned} E^{(1)}(\vec{k}) &= \langle \vec{k}|V(\vec{r})|\vec{k}\rangle = \overline{V(\vec{r})} = V_0 \\ E^{(1)}(\vec{k}+\vec{G}) &= \langle \vec{k}+\vec{G}|V(\vec{r})|\vec{k}+\vec{G}\rangle = V_0. \end{aligned} \tag{1.21}$$

Solution of this determinantal equation (Eq. 1.19) yields:

$$[E - V_0 - E^{(0)}(\vec{k})][E - V_0 - E^{(0)}(\vec{k}+\vec{G})] - |V_{\vec{G}}|^2 = 0, \tag{1.22}$$

or equivalently

$$E^2 - E[2V_0 + E^{(0)}(\vec{k}) + E^{(0)}(\vec{k}+\vec{G})] + [V_0 + E^{(0)}(\vec{k})][V_0 + E^{(0)}(\vec{k}+\vec{G})] - |V_{\vec{G}}|^2 = 0. \tag{1.23}$$

Solution of the quadratic equation (Eq. 1.23) yields

$$E^{\pm} = V_0 + \frac{1}{2}[E^{(0)}(\vec{k}) + E^{(0)}(\vec{k}+\vec{G})] \pm \sqrt{\frac{1}{4}[E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})]^2 + |V_{\vec{G}}|^2} \tag{1.24}$$

and we come out with two solutions for the two strongly coupled states. It is of interest to look at these two solutions in two limiting cases:

case (i) $|V_{\vec{G}}| \ll \frac{1}{2}|[E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})]|$

In this case we can expand the square root expression in Eq. 1.24 for small $|V_{\vec{G}}|$ to obtain:

$$\begin{aligned} E(\vec{k}) &= V_0 + \tfrac{1}{2}[E^{(0)}(\vec{k}) + E^{(0)}(\vec{k}+\vec{G})] \\ &\pm \tfrac{1}{2}[E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})] \cdot [1 + \tfrac{2|V_{\vec{G}}|^2}{[E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})]^2} + \ldots] \end{aligned} \tag{1.25}$$

which simplifies to the two solutions:

$$E^-(\vec{k}) = V_0 + E^{(0)}(\vec{k}) + \frac{|V_{\vec{G}}|^2}{E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})} \tag{1.26}$$

$$E^+(\vec{k}) = V_0 + E^{(0)}(\vec{k}+\vec{G}) + \frac{|V_{\vec{G}}|^2}{E^{(0)}(\vec{k}+\vec{G}) - E^{(0)}(\vec{k})} \tag{1.27}$$

and we recover the result Eq. 1.18 obtained before using non–degenerate perturbation theory. This result in Eq. 1.18 is valid far from the Brillouin zone boundary, but near the zone boundary the more complete expression of Eq. 1.24 must be used.

case (ii) $|V_{\vec{G}}| \gg \frac{1}{2}|[E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})]|$

Sufficiently close to the Brillouin zone boundary

$$|E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})| \ll |V_{\vec{G}}| \tag{1.28}$$

so that we can expand $E(\vec{k})$ as given by Eq. 1.24 to obtain

$$E^{\pm}(\vec{k}) = \frac{1}{2}[E^{(0)}(\vec{k}) + E^{(0)}(\vec{k}+\vec{G})] + V_0 \pm \left[|V_{\vec{G}}| + \frac{1}{8}\frac{[E^{(0)}(\vec{k}) - E^{(0)}(\vec{k}+\vec{G})]^2}{|V_{\vec{G}}|} + \ldots\right] \tag{1.29}$$

$$\cong \frac{1}{2}[E^{(0)}(\vec{k}) + E^{(0)}(\vec{k}+\vec{G})] + V_0 \pm |V_{\vec{G}}|, \tag{1.30}$$

so that at the Brillouin zone boundary $E^+(\vec{k})$ is elevated by $|V_{\vec{G}}|$, while $E^-(\vec{k})$ is depressed by $|V_{\vec{G}}|$ and the band gap that is formed is $2|V_{\vec{G}}|$, where $\vec{G}$ is the reciprocal lattice vector for which $E(\vec{k}_{B.Z.}) = E(\vec{k}_{B.Z.} + \vec{G})$ and

$$V_{\vec{G}} = \frac{1}{\Omega_0}\int_{\Omega_0} e^{i\vec{G}\cdot\vec{r}} V(\vec{r}) d^3r. \tag{1.31}$$

From this discussion it is clear that every Fourier component of the periodic potential gives rise to a specific band gap. We see further that the *band gap* represents a range of energy values for which there is no solution to the eigenvalue problem of Eq. 1.19 for real k (see Fig. 1.1). In the band gap we assign an imaginary value to the wave vector which can be interpreted as a highly damped and non–propagating wave.

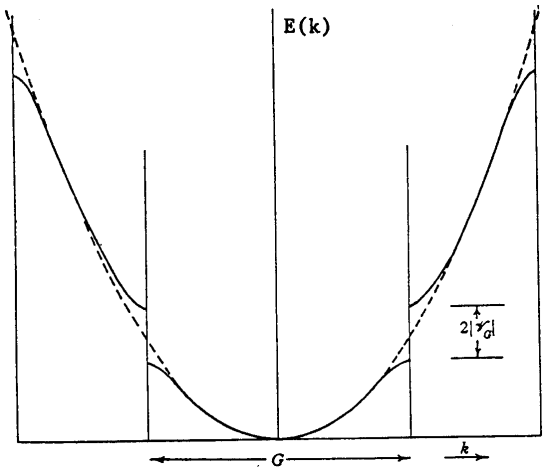


Figure 1.1: One dimensional electron energy bands for the nearly free electron model shown in the extended Brillouin zone scheme. The dashed curve corresponds to the case of free electrons and solid curves to the case where a weak periodic potential is present.

We note that the larger the value of $\vec{G}$, the smaller the value of $V_{\vec{G}}$, so that higher Fourier components give rise to smaller band gaps. Near these energy discontinuities, the wave functions become linear combinations of the unperturbed states

$$\begin{aligned} \psi_{\vec{k}} &= \alpha_1 \psi^{(0)}_{\vec{k}} + \beta_1 \psi^{(0)}_{\vec{k}+\vec{G}} \\ \psi_{\vec{k}+\vec{G}} &= \alpha_2 \psi^{(0)}_{\vec{k}} + \beta_2 \psi^{(0)}_{\vec{k}+\vec{G}} \end{aligned} \tag{1.32}$$

and at the zone boundary itself, instead of traveling waves $e^{i\vec{k}\cdot\vec{r}}$, the wave functions become standing waves $\cos\vec{k}\cdot\vec{r}$ and $\sin\vec{k}\cdot\vec{r}$. We note that the $\cos(\vec{k}\cdot\vec{r})$ solution corresponds to a maximum in the charge density at the lattice sites and therefore corresponds to an energy minimum (the lower level). Likewise, the $\sin(\vec{k}\cdot\vec{r})$ solution corresponds to a minimum in the charge density and therefore corresponds to a maximum in the energy, thus forming the upper level.

In constructing $E(\vec{k})$ for the reduced zone scheme we make use of the periodicity of $E(\vec{k})$ in reciprocal space

$$E(\vec{k}+\vec{G}) = E(\vec{k}). \tag{1.33}$$

The reduced zone scheme more clearly illustrates the formation of energy bands (labeled (1) and (2) in Fig. 1.2), band gaps E_g and band widths (defined in Fig. 1.2 as the range of energy between E_{min} and E_{max} for a given energy band).

We now discuss the connection between the $E(\vec{k})$ relations shown above and the transport properties of solids, which can be illustrated by considering the case of a semiconductor. An intrinsic semiconductor at temperature $T = 0$ has no carriers so that the Fermi level runs right through the band gap. On the diagram of Fig. 1.2, this would mean that the Fermi level might run between bands (1) and (2), so that band (1) is completely occupied

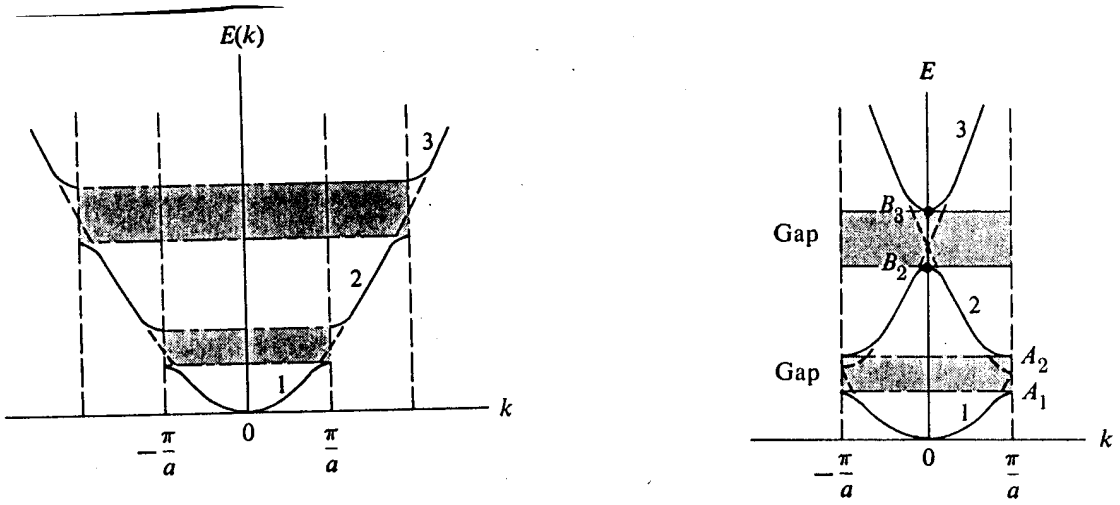


Figure 1.2: (a) One dimensional electron energy bands for the nearly free electron model shown in the extended Brillouin zone scheme for the three bands of lowest energy. (b) The same $E(\vec{k})$ as in (a) but now shown on the reduced zone scheme. The shaded areas denote the band gaps and the white areas the band states.

and band (2) is completely empty. One further property of the semiconductor is that the band gap E_g be small enough so that at some temperature (e.g., room temperature) there will be reasonable numbers of thermally excited carriers, perhaps $10^{15}/\text{cm}^3$. The doping with donor (electron donating) impurities will raise the Fermi level and doping with acceptor (electron extracting) impurities will lower the Fermi level. Neglecting for the moment the effect of impurities on the $E(\vec{k})$ relations for the perfectly periodic crystal, let us consider what happens when we raise the Fermi level into the bands. If we know the shape of the $E(\vec{k})$ curve, we are in a position to estimate the velocity of the electrons and also the so–called *effective mass* of the electrons. From the diagram in Fig. 1.2 we see that the conduction bands tend to fill up electron states starting at their energy extrema.

Since the energy bands have zero slope about their extrema, we can write $E(\vec{k})$ as a quadratic form in $\vec{k}$. It is convenient to write the proportionality in terms of the quantity called the effective mass m^*

$$E(\vec{k}) = E(0) + \frac{\hbar^2 k^2}{2m^*} \tag{1.34}$$

so that m^* is defined by

$$\frac{1}{m^*} \equiv \frac{\partial^2 E(\vec{k})}{\hbar^2 \partial k^2} \tag{1.35}$$

and we can say in some approximate way that an electron in a solid moves as if it were a free electron but with an effective mass m^* rather than a free electron mass. The larger the band curvature, the smaller the effective mass. The mean velocity of the electron is also

found from $E(\vec{k})$, according to the relation

$$\vec{v}_k = \frac{1}{\hbar}\frac{\partial E(\vec{k})}{\partial \vec{k}}. \tag{1.36}$$

For this reason the energy dispersion relations $E(\vec{k})$ are very important in the determination of the transport properties for carriers in solids.

1.2.2 Tight Binding Approximation

In the tight binding approximation a number of assumptions are made and these are different from the assumptions that are made for the weak binding approximation. The assumptions for the tight binding approximation are:

1. The energy eigenvalues and eigenfunctions are known for an electron in an isolated atom.

2. When the atoms are brought together to form a solid they remain sufficiently far apart so that each electron can be assigned to a particular atomic site. This assumption is not valid for valence electrons in metals and for this reason, these valence electrons are best treated by the weak binding approximation.

3. The periodic potential is approximated by a superposition of atomic potentials.

4. Perturbation theory can be used to treat the difference between the actual potential and the atomic potential.

Thus both the weak and tight binding approximations are based on perturbation theory. For the weak binding approximation the unperturbed state is the free electron plane–wave state while for the tight binding approximation the unperturbed state is the atomic state. In the case of the weak binding approximation, the perturbation Hamiltonian is the weak periodic potential itself, while for the tight binding case, the perturbation is the *difference* between the periodic potential and the atomic potential around which the electron is localized.

We review here the major features of the tight binding approximation. Let $\phi(\vec{r}-\vec{R}_n)$ represent the atomic wave function for an atom at a lattice position denoted by $\vec{R}_n$, which is measured with respect to the origin. The Schrödinger equation for an electron in an isolated atom is then:

$$\Big[-\frac{\hbar^2}{2m}\nabla^2 + U(\vec{r}-\vec{R}_n) - E^{(0)}\Big]\phi(\vec{r}-\vec{R}_n) = 0 \tag{1.37}$$

where $U(\vec{r}-\vec{R}_n)$ is the atomic potential and $E^{(0)}$ is the atomic eigenvalue (see Fig. 1.3). We now assume that the atoms are brought together to form the crystal for which $V(\vec{r})$ is the periodic potential, and $\psi(\vec{r})$ and $E(\vec{k})$ are, respectively, the wave function and energy eigenvalue for the electron in the crystal:

$$\Big[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r}) - E\Big]\psi(\vec{r}) = 0. \tag{1.38}$$

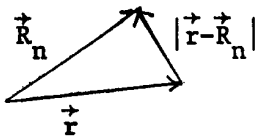


Figure 1.3: Definition of the vectors used in the tight binding approximation.

In the tight binding approximation we write $V(\vec{r})$ as a sum of atomic potentials:

$$V(\vec{r}) \simeq \sum_n U(\vec{r} - \vec{R}_n). \tag{1.39}$$

If the interaction between neighboring atoms is ignored, then each state has a degeneracy of N = number of atoms in the crystal. However, the interaction between the atoms lifts this degeneracy.

The energy eigenvalues $E(\vec{k})$ in the tight binding approximation for a non–degenerate s–state is simply given by

$$E(\vec{k}) = \frac{\langle \vec{k}|\mathcal{H}|\vec{k}\rangle}{\langle \vec{k}|\vec{k}\rangle}. \tag{1.40}$$

The normalization factor in the denominator $\langle \vec{k}|\vec{k}\rangle$ is inserted because the wave functions $\psi_{\vec{k}}(\vec{r})$ in the tight binding approximation are usually not normalized. The Hamiltonian in the tight binding approximation is written as

$$\mathcal{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r}) = \Big\{ -\frac{\hbar^2}{2m}\nabla^2 + [V(\vec{r}) - U(\vec{r} - \vec{R}_n)] + U(\vec{r} - \vec{R}_n) \Big\} \tag{1.41}$$

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}' \tag{1.42}$$

in which $\mathcal{H}_0$ is the atomic Hamiltonian at site n

$$\mathcal{H}_0 = -\frac{\hbar^2}{2m}\nabla^2 + U(\vec{r} - \vec{R}_n) \tag{1.43}$$

and $\mathcal{H}'$ is the difference between the actual periodic potential and the atomic potential at lattice site n

$$\mathcal{H}' = V(\vec{r}) - U(\vec{r} - \vec{R}_n). \tag{1.44}$$

We construct the wave functions for the unperturbed problem as a linear combination of atomic functions $\phi_j(\vec{r} - \vec{R}_n)$ labeled by quantum number j

$$\psi_j(\vec{r}) = \sum_{n=1}^{N} C_{j,n}\phi_j(\vec{r} - \vec{R}_n) \tag{1.45}$$

and so that $\psi_j(\vec{r})$ is an eigenstate of a Hamiltonian satisfying the periodic potential of the lattice. In this treatment we assume that the tight binding wave–functions ψ_j can be

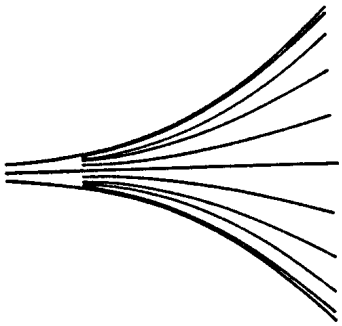


Figure 1.4: The relation between atomic states and the broadening due to the presence of neighboring atoms. As the interatomic distance decreases (going to the right in the diagram), the level broadening increases so that a band of levels occurs at atomic separations characteristic of solids.

identified with a *single* atomic state ϕ_j; this approximation must be relaxed in dealing with degenerate levels. According to Bloch's theorem $\psi_j(\vec{r})$ must satisfy the relation:

$$\psi_j(\vec{r}+\vec{R}_m) = e^{i\vec{k}\cdot\vec{R}_m}\psi_j(\vec{r}) \tag{1.46}$$

where $\vec{R}_m$ is an arbitrary lattice vector. This restriction imposes a special form on the coefficients $C_{j,n}$.

Substitution of the expansion in atomic functions $\psi_j(\vec{r})$ from Eq. 1.45 into the left side of Eq. 1.46 yields:

$$\begin{aligned}\psi_j(\vec{r}+\vec{R}_m) &= \textstyle\sum_n C_{j,n}\ \phi_j(\vec{r}-\vec{R}_n+\vec{R}_m) \\ &= \textstyle\sum_Q C_{j,Q+m}\ \phi_j(\vec{r}-\vec{R}_Q) \\ &= \textstyle\sum_n C_{j,n+m}\ \phi(\vec{r}-\vec{R}_n)\end{aligned} \tag{1.47}$$

where we have utilized the substitution $\vec{R}_Q = \vec{R}_n - \vec{R}_m$ and the fact that Q is a dummy index. Now for the right side of the Bloch theorem (Eq. 1.46) we have

$$e^{i\vec{k}\cdot\vec{R}_m}\psi_j(\vec{r}) = \sum_n C_{j,n}e^{i\vec{k}\cdot\vec{R}_m}\phi_j(\vec{r}-\vec{R}_n). \tag{1.48}$$

The coefficients $C_{j,n}$ which relate the actual wave function $\psi_j(\vec{r})$ to the atomic functions $\phi_j(\vec{r}-\vec{R}_n)$ are therefore not arbitrary but must thus satisfy:

$$C_{j,n+m} = e^{i\vec{k}\cdot\vec{R}_m}C_{j,n} \tag{1.49}$$

which can be accomplished by setting:

$$C_{j,n} = \xi_j e^{i\vec{k}\cdot\vec{R}_n} \tag{1.50}$$

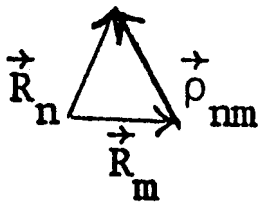


Figure 1.5: Definition of $\vec{\rho}_{nm}$ denoting the distance between atoms at $\vec{R}_m$ and $\vec{R}_n$.

where the new coefficient ξ_j is independent of n. We therefore obtain:

$$\psi_{j,\vec{k}}(\vec{r}) = \xi_j \sum_n e^{i\vec{k}\cdot\vec{R}_n} \phi_j(\vec{r} - \vec{R}_n) \tag{1.51}$$

where j is an index labeling the particular atomic state of degeneracy N and $\vec{k}$ is the quantum number for the translation operator and labels the Bloch state $\psi_{j,\vec{k}}(\vec{r})$.

For simplicity, we will limit the present discussion of the tight binding approximation to s–bands (non–degenerate atomic states) and therefore we can suppress the j index on the wave functions. (The treatment for p–bands is similar to what we will do here, but more complicated because of the degeneracy of the atomic states.) To find matrix elements of the Hamiltonian we write

$$\langle\vec{k}'|\mathcal{H}|\vec{k}\rangle = |\xi|^2 \sum_{n,m} e^{i(\vec{k}\cdot\vec{R}_n - \vec{k}'\cdot\vec{R}_m)} \int_\Omega \phi^*(\vec{r} - \vec{R}_m)\mathcal{H}\phi(\vec{r} - \vec{R}_n) d^3r \tag{1.52}$$

in which the integration is carried out throughout the volume of the crystal. Since $\mathcal{H}$ is a function which is periodic in the lattice, the only significant distance (see Fig. 1.5) is

$$(\vec{R}_n - \vec{R}_m) = \vec{\rho}_{nm}. \tag{1.53}$$

We then write the integral in Eq. 1.52 as:

$$\langle\vec{k}'|\mathcal{H}|\vec{k}\rangle = |\xi|^2 \sum_{\vec{R}_m} e^{i(\vec{k}-\vec{k}')\cdot\vec{R}_m} \sum_{\vec{\rho}_{nm}} e^{i\vec{k}\vec{\rho}_{nm}} \mathcal{H}_{mn}(\vec{\rho}_{nm}) \tag{1.54}$$

where we have written the matrix element $\mathcal{H}_{mn}(\vec{\rho}_{nm})$ as

$$\mathcal{H}_{mn}(\vec{\rho}_{nm}) = \int_\Omega \phi^*(\vec{r} - \vec{R}_m)\mathcal{H}\phi(\vec{r} - \vec{R}_m - \vec{\rho}_{nm}) d^3r = \int_\Omega \phi^*(\vec{r}\,')\mathcal{H}\phi(\vec{r}\,' - \vec{\rho}_{nm}) d^3r'. \tag{1.55}$$

We note here that the integral in Eq. 1.55 depends only on $\vec{\rho}_{nm}$ and not on $\vec{R}_m$. According to Eq. 1.13, the first sum in Eq. 1.54 is

$$\sum_{\vec{R}_m} e^{i(\vec{k}-\vec{k}')\cdot\vec{R}_m} = \delta_{\vec{k}',\vec{k}+\vec{G}} N \tag{1.56}$$

where $\vec{G}$ is a reciprocal lattice vector. It is convenient to restrict the $\vec{k}$ vectors to lie within the first Brillouin zone (i.e., we limit ourselves to reduced wave vectors). This is consistent with the manner of counting states with the periodic boundary conditions on a crystal of dimension d on a side

$$k_i d = 2\pi m_i \qquad \text{for each direction } i \tag{1.57}$$

where m_i is an integer in the range $1 \leq m_i < N_i$ where $N_i \approx N^{1/3}$. From Eq. 1.57 we have

$$k_i = \frac{2\pi m_i}{d}. \tag{1.58}$$

The maximum value that a particular m_i can assume is N_i and the maximum value for k_i is $2\pi/a$ at the Brillouin zone boundary since $N_i/d = 1/a$. With this restriction, $\vec{k}$ and $\vec{k}'$ must both lie within the 1^{st} B.Z. and thus cannot differ by any reciprocal lattice vector other than $\vec{G} = 0$. We thus obtain the following form for the matrix element of $\mathcal{H}$ (and also for the matrix elements of $\mathcal{H}_0$ and $\mathcal{H}'$):

$$\langle \vec{k}'|\mathcal{H}|\vec{k}\rangle = |\xi|^2 N \delta_{\vec{k},\vec{k}'} \sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}} \mathcal{H}_{mn}(\vec{\rho}_{nm}) \tag{1.59}$$

yielding the result

$$E(\vec{k}) = \frac{\langle \vec{k}|\mathcal{H}|\vec{k}\rangle}{\langle \vec{k}|\vec{k}\rangle} = \frac{\sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}} \mathcal{H}_{mn}(\vec{\rho}_{nm})}{\sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}} \mathcal{S}_{mn}(\vec{\rho}_{nm})} \tag{1.60}$$

in which

$$\langle \vec{k}'|\vec{k}\rangle = |\xi|^2 \delta_{\vec{k},\vec{k}'} N \sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}} \mathcal{S}_{mn}(\vec{\rho}_{nm}) \tag{1.61}$$

where the matrix element $\mathcal{S}_{mn}(\vec{\rho}_{nm})$ measures the overlap of atomic functions on different sites

$$\mathcal{S}_{mn}(\vec{\rho}_{nm}) = \int_{\Omega} \phi^*(\vec{r})\phi(\vec{r} - \vec{\rho}_{nm}) d^3r. \tag{1.62}$$

The overlap integral $\mathcal{S}_{mn}(\vec{\rho}_{nm})$ will be nearly 1 when $\vec{\rho}_{nm} = 0$ and will fall off rapidly as $\vec{\rho}_{nm}$ increases, which exemplifies the spirit of the tight binding approximation. By selecting $\vec{k}$ vectors to lie within the first Brillouin zone, the orthogonality condition on the $\psi_{\vec{k}}(\vec{r})$ is automatically satisfied. Writing $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}'$ yields:

$$\begin{aligned} \mathcal{H}_{mn} &= \int_{\Omega} \phi^*(\vec{r} - \vec{R}_m) \left[-\tfrac{\hbar^2}{2m}\nabla^2 + U(\vec{r} - \vec{R}_n) \right] \phi(\vec{r} - \vec{R}_n) d^3r \\ &+ \int_{\Omega} \phi^*(\vec{r} - \vec{R}_m)[V(\vec{r}) - U(\vec{r} - \vec{R}_n)]\phi(\vec{r} - \vec{R}_n) d^3r \end{aligned} \tag{1.63}$$

or

$$\mathcal{H}_{mn} = E^{(0)} \mathcal{S}_{mn}(\vec{\rho}_{nm}) + \mathcal{H}'_{mn}(\vec{\rho}_{nm}) \tag{1.64}$$

which results in the general expression for the tight binding approximation:

$$E(\vec{k}) = E^{(0)} + \frac{\sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}} \mathcal{H}'_{mn}(\vec{\rho}_{nm})}{\sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}} \mathcal{S}_{mn}(\vec{\rho}_{nm})}. \tag{1.65}$$

In the spirit of the tight binding approximation, the second term in Eq. 1.65 is assumed to be small, which is a good approximation if the overlap of the atomic wave functions is small. We classify the sum over $\vec{\rho}_{nm}$ according to the distance between site m and site n: (i) zero distance, (ii) the nearest neighbor distance, (iii) the next nearest neighbor distance, etc.

$$\sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}}\mathcal{H}'_{mn}(\vec{\rho}_{nm}) = \mathcal{H}'_{nn}(0) + \sum_{\vec{\rho}_1} e^{i\vec{k}\cdot\vec{\rho}_{nm}}\mathcal{H}'_{mn}(\vec{\rho}_{nm}) + ... \tag{1.66}$$

The zeroth neighbor term $\mathcal{H}'_{nn}(0)$ in Eq. 1.66 results in a constant additive energy, independent of $\vec{k}$. The sum over nearest neighbor distances $\vec{\rho}_1$ gives rise to a $\vec{k}$–dependent perturbation, and hence is of particular interest in calculating the band structure. The terms $\mathcal{H}'_{nn}(0)$ and the sum over the nearest neighbor terms in Eq. 1.66 are of comparable magnitude, as can be seen by the following argument. In the integral

$$\mathcal{H}'_{nn}(0) = \int \phi^*(\vec{r}-\vec{R}_n)[V - U(\vec{r}-\vec{R}_n)]\phi(\vec{r}-\vec{R}_n)d^3r \tag{1.67}$$

we note that $|\phi(\vec{r}-\vec{R}_n)|^2$ has an appreciable amplitude only in the vicinity of the site $\vec{R}_n$. But at site $\vec{R}_n$, the potential energy term $[V - U(\vec{r}-\vec{R}_n)] = \mathcal{H}'$ is a small term, so that $\mathcal{H}'_{nn}(0)$ represents the product of a small term times a large term. On the other hand, the integral $\mathcal{H}'_{mn}(\vec{\rho}_{nm})$ taken over nearest neighbor distances has a factor $[V - U(\vec{r}-\vec{R}_n)]$ which is large near the mth site; however, in this case the wave functions $\phi^*(\vec{r}-\vec{R}_m)$ and $\phi(\vec{r}-\vec{R}_n)$ are on different atomic sites and have only a small overlap on nearest neighbor sites. Therefore $\mathcal{H}'_{mn}(\vec{\rho}_{nm})$ over nearest neighbor sites also results in the product of a large quantity times a small quantity.

In treating the denominator in the perturbation term of Eq. 1.65, we must sum

$$\sum_{\vec{\rho}_{nm}} e^{i\vec{k}\cdot\vec{\rho}_{nm}}\mathcal{S}_{mn}(\vec{\rho}_{nm}) = \mathcal{S}_{nn}(0) + \sum_{\vec{\rho}_1} e^{i\vec{k}\cdot\vec{\rho}_{nm}}\mathcal{S}_{mn}(\vec{\rho}_{nm}) + ... \tag{1.68}$$

In this case the leading term $\mathcal{S}_{nn}(0)$ is approximately unity and the overlap integral $\mathcal{S}_{mn}(\vec{\rho}_{nm})$ over nearest neighbor sites is small, and can be neglected to lowest order in comparison with unity. The nearest neighbor term in Eq. 1.68 is of comparable magnitude to the next nearest neighbor terms arising from $\mathcal{H}_{mn}(\vec{\rho}_{nm})$ in Eq. 1.66.

We will here make *several explicit evaluations* of $E(\vec{k})$ in the tight–binding limit to show how this method incorporates the crystal symmetry. For illustrative purposes we will give results for the simple cubic lattice (SC), the body centered cubic (BCC) and face centered cubic lattice (FCC). We shall assume here that the overlap of atomic potentials on neighboring sites is sufficiently weak so that only nearest neighbor terms need be considered in the sum on $\mathcal{H}'_{mn}$ and only the leading term in the sum of $\mathcal{S}_{mn}$.

For the simple cubic structure there are 6 terms in the nearest neighbor sum on $\mathcal{H}'_{mn}$ with $\vec{\rho}_1$ vectors given by:

$$\vec{\rho}_1 = a(\pm 1, 0, 0), \;\; a(0, \pm 1, 0), \;\; a(0, 0, \pm 1). \tag{1.69}$$

By symmetry $\mathcal{H}'_{mn}(\vec{\rho}_1)$ is the same for all of the $\vec{\rho}_1$ vectors so that

$$E(\vec{k}) = E^{(0)} + \mathcal{H}'_{nn}(0) + 2\mathcal{H}'_{mn}(\vec{\rho}_1)[\cos k_x a + \cos k_y a + \cos k_z a] + ... \tag{1.70}$$

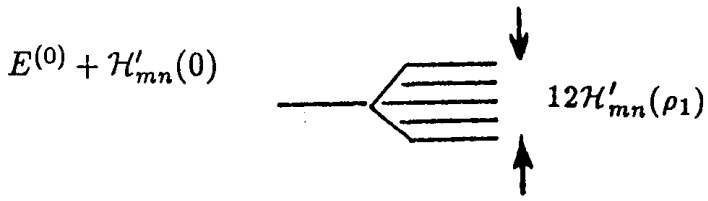


Figure 1.6: The relation between the atomic levels and the broadened level in the tight binding approximation.

where $\vec{\rho}_1$ = nearest neighbor separation and k_x, k_y, k_z are components of wave vector $\vec{k}$ in the first Brillouin zone.

This dispersion relation $E(\vec{k})$ clearly satisfies three properties which characterize energy eigenvalues in typical periodic structures:

1. Periodicity in $\vec{k}$ space under translation by a reciprocal lattice vector $\vec{k} \to \vec{k} + \vec{G}$,
2. $E(\vec{k})$ is an even function of $\vec{k}$ (i.e., $E(k) = E(-k)$)
3. $\partial E/\partial k = 0$ at the Brillouin zone boundary

In the above expression (Eq. 1.70) for $E(\vec{k})$, the maximum value for the term in brackets is $\pm$ 3. Therefore for a simple cubic lattice in the tight binding approximation we obtain a bandwidth of 12 $\mathcal{H}'_{mn}(\rho_1)$ from nearest neighbor interactions as shown in Fig. 1.6. Because of the different locations of the nearest neighbor atoms in the case of the BCC and FCC lattices, the expression for $E(\vec{k})$ will be different for the various cubic lattices. Thus the form of the tight binding approximation explicitly takes account of the crystal structure. These results are summarized below.

simple cubic

$$E(\vec{k}) = \text{ const} + 2\mathcal{H}'_{mn}(\vec{\rho}_1)[\cos k_x a + \cos k_y a + \cos k_z a] + ... \tag{1.71}$$

body centered cubic

The eight $\vec{\rho}_1$ vectors for the nearest neighbor distances in the BCC structure are $(\pm a/2, \pm a/2, \pm a/2)$ so that there are 8 exponential terms which combine in pairs such as:

$$\left[\exp\frac{ik_x a}{2}\exp\frac{ik_y a}{2}\exp\frac{ik_z a}{2} + \exp\frac{-ik_x a}{2}\exp\frac{ik_y a}{2}\exp\frac{ik_z a}{2}\right] \tag{1.72}$$

to yield

$$2\cos(\frac{k_x a}{2})\exp\frac{ik_y a}{2}\exp\frac{ik_z a}{2}. \tag{1.73}$$

We thus obtain for the BCC structure:

$$E(\vec{k}) = \text{ const} + 8\mathcal{H}'_{mn}(\vec{\rho}_1)\cos(\frac{k_x a}{2})\cos(\frac{k_y a}{2})\cos(\frac{k_z a}{2}) + ... \tag{1.74}$$

where $\mathcal{H}'_{mn}(\vec{\rho}_1)$ is the matrix element of the perturbation Hamiltonian taken between nearest neighbor atomic orbitals.

face centered cubic

For the FCC structure there are 12 nearest neighbor distances $\vec{\rho}_1$: $(0, \pm\frac{a}{2}, \pm\frac{a}{2})$, $(\pm\frac{a}{2}, \pm\frac{a}{2}, 0)$, $(\pm\frac{a}{2}, 0, \pm\frac{a}{2})$, so that the twelve exponential terms combine in groups of 4 to yield:

$$\exp\tfrac{ik_xa}{2}\exp\tfrac{ik_ya}{2} + \exp\tfrac{ik_xa}{2}\exp\tfrac{-ik_ya}{2} + \exp\tfrac{-ik_xa}{2}\exp\tfrac{ik_ya}{2} + \exp\tfrac{-ik_xa}{2}\exp\tfrac{-ik_ya}{2} = 4\cos(\tfrac{k_xa}{2})\cos(\tfrac{k_ya}{2}), \tag{1.75}$$

thus resulting in the energy dispersion relation

$$E(\vec{k}) = \text{ const}+4\mathcal{H}'_{mn}(\vec{\rho}_1)\left[\cos(\frac{k_ya}{2})\cos(\frac{k_za}{2}) + \cos(\frac{k_xa}{2})\cos(\frac{k_za}{2}) + \cos(\frac{k_xa}{2})\cos(\frac{k_ya}{2})\right]+... \tag{1.76}$$

We note that $E(\vec{k})$ for the FCC is different from that for the SC or BCC structures. The tight-binding approximation has symmetry considerations built into its formulation through the symmetrical arrangement of the atoms in the lattice. The situation is quite different in the weak binding approximation where symmetry enters into the form of $V(\vec{r})$ and determines which Fourier components $V_{\vec{G}}$ will be important in creating band gaps.

1.2.3 Weak and Tight Binding Approximations

We will now make some general statements about bandwidths and forbidden band gaps which follow from either the tight binding or weak binding approximations. With increasing energy, the bandwidth tends to increase. On the tight–binding picture, the higher atomic states are less closely bound to the nucleus, and the resulting increased overlap of the wave functions results in a larger value for $\mathcal{H}'_{mn}(\vec{\rho}_1)$ in the case of the higher atomic states: that is, for silicon, which has 4 valence electrons in the $n = 3$ shell, the overlap integral $\mathcal{H}'_{mn}(\vec{\rho}_1)$ will be smaller than for germanium which is isoelectronic to silicon but has instead 4 valence electrons in the $n = 4$ atomic shell. On the weak–binding picture, the same result follows, since for higher energies, the electrons are more nearly free; therefore, there are more allowed energy ranges available, or equivalently, the energy range of the forbidden states is smaller. Also in the weak–binding approximation the band gap of $2|V_{\vec{G}}|$ tends to decrease as $\vec{G}$ increases, because of the oscillatory character of $e^{-i\vec{G}\cdot\vec{r}}$ in

$$V_{\vec{G}} = \frac{1}{\Omega_0}\int_{\Omega_0} e^{-i\vec{G}\cdot\vec{r}}V(\vec{r})d^3r. \tag{1.77}$$

From the point of view of the tight–binding approximation, the increasing bandwidth with increasing energy (see Fig. 1.7) is also equivalent to a decrease in the forbidden band gap. At the same time, the atomic states at higher energies become more closely spaced, so that the increased bandwidth eventually results in band overlaps. When band overlaps occur, the tight-binding approximation as given above must be generalized to treat coupled or interacting bands using degenerate perturbation theory (see Appendix A).

1.2.4 Tight Binding Approximation with 2 Atoms/Unit Cell

We present here a simple example of the tight binding approximation for a simplified version of polyacetylene which has two carbon atoms (with their appended hydrogens) per unit

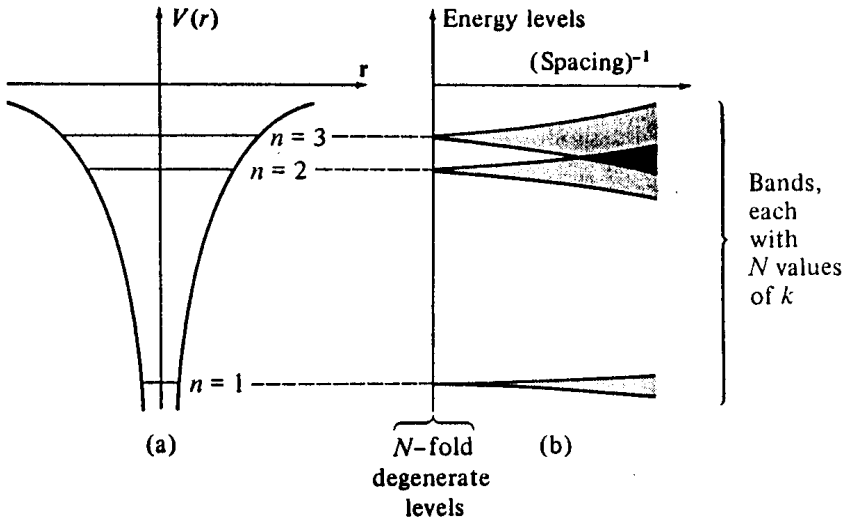


Figure 1.7: Schematic diagram of the increased bandwidth and decreased band gap in the tight binding approximation as the interatomic separation decreases.

cell. In Fig. 1.8 we show, within the box defined by the dotted lines, the unit cell for *trans*-polyacetylene $(CH)_x$. This unit cell of an infinite one-dimensional chain contains two inequivalent carbon atoms, A and B. There is one π-electron per carbon atom, thus giving rise to two π-energy bands in the first Brillouin zone. These two bands are called bonding π-bands for the valence band, and anti-bonding π-bands for the conduction band.

The lattice unit vector and the reciprocal lattice unit vector of this one-dimensional polyacetylene chain are given by $\vec{a}_1 = (a, 0, 0)$ and $\vec{b}_1 = (2\pi/a, 0, 0)$, respectively. The Brillouin zone is the line segment $-\pi/a < k < \pi/a$. The Bloch orbitals consisting of A and B atoms are given by

$$\psi_j(r) = \frac{1}{\sqrt{N}} \sum_{R_\alpha} e^{ikR_\alpha} \phi_j(r - R_\alpha), \quad (\alpha = \mathrm{A}, \mathrm{B}) \tag{1.78}$$

where the summation is taken over the atom site coordinate R_α for the A or B carbon atoms in the solid.

To solve for the energy eigenvalues and wavefunctions we need to solve the general equation:

$$\mathcal{H}\psi = E\mathcal{S}\psi \tag{1.79}$$

where $\mathcal{H}$ is the $n \times n$ tight binding matrix Hamiltonian for the n coupled bands ($n = 2$ in the case of polyacetylene) and $\mathcal{S}$ is the corresponding $n \times n$ overlap integral matrix. To obtain a solution to this matrix equation, we require that the determinant $|\mathcal{H} - E\mathcal{S}|$ vanish. This approach is easily generalized to periodic structures with more than 2 atoms per unit cell.

The (2×2) matrix Hamiltonian, $\mathcal{H}_{\alpha\beta}$, $(\alpha, \beta = A, B)$ is obtained by substituting Eq. (1.78) into

$$\mathcal{H}_{jj'}(\vec{k}) = \langle \psi_j | \mathcal{H} | \psi_{j'} \rangle, \quad \mathcal{S}_{jj'}(\vec{k}) = \langle \psi_j | \psi_{j'} \rangle \qquad (j, j' = 1, 2), \tag{1.80}$$

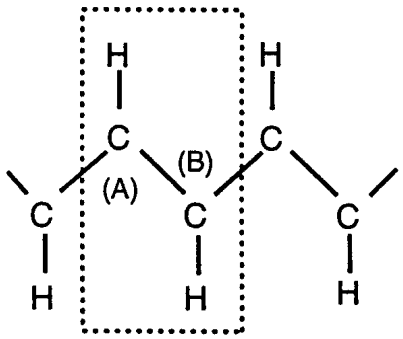


Figure 1.8: The unit cell of *trans*-polyacetylene bounded by a box defined by the dotted lines, and showing two inequivalent carbon atoms, A and B, in the unit cell.

where the integrals over the Bloch orbitals, $\mathcal{H}_{jj'}(\vec{k})$ and $\mathcal{S}_{jj'}(\vec{k})$, are called transfer integral matrices and overlap integral matrices, respectively. When $\alpha = \beta =$ A, we obtain the diagonal matrix element

$$\begin{aligned}
\mathcal{H}_{AA}(r) &= \frac{1}{N}\sum_{R,R'} e^{ik(R-R')}\langle\phi_A(r-R')|\mathcal{H}|\phi_A(r-R)\rangle \\
&= \frac{1}{N}\sum_{R'=R} E_{2p} + \frac{1}{N}\sum_{R'=R\pm a} e^{\pm ika}\langle\phi_A(r-R')|\mathcal{H}|\phi_A(r-R)\rangle \\
&\quad +(\text{terms equal to or more distant than } R' = R \pm 2a) \\
&= E_{2p} + (\text{terms equal to or more distant than } R' = R \pm a).
\end{aligned} \tag{1.81}$$

In Eq. (1.81) the main contribution to the matrix element $\mathcal{H}_{AA}$ comes from $R' = R$, and this gives the orbital energy of the $2p$ level, E_{2p}. We note that E_{2p} is not simply the atomic energy value for the free atom, because the Hamiltonian $\mathcal{H}$ also includes a crystal potential contribution. The next order contribution to $\mathcal{H}_{AA}$ in Eq. (1.81) comes from terms in $R' = R \pm a$, which are here neglected for simplicity. Similarly, $\mathcal{H}_{BB}$ also gives E_{2p} to the same order of approximation.

Next let us consider the off-diagonal matrix element $\mathcal{H}_{AB}(r)$ which explicitly couples the A unit to the B unit. The largest contribution to $\mathcal{H}_{AB}(r)$ arises when atoms A and B are nearest neighbors. Thus in the summation over R', we only consider the terms with $R' = R \pm a/2$ as a first approximation and neglect more distant terms to obtain

$$\begin{aligned}
\mathcal{H}_{AB}(r) &= \frac{1}{N}\sum_{R}\Big\{e^{-ika/2}\langle\phi_A(r-R)|\mathcal{H}|\phi_B(r-R-a/2)\rangle \\
&\quad +e^{ika/2}\langle\phi_A(r-R)|\mathcal{H}|\phi_B(r-R+a/2)\rangle\Big\} \\
&= 2t\cos(ka/2)
\end{aligned} \tag{1.82}$$

where t is the transfer integral appearing in Eq. (1.82) and is denoted by

$$t = \langle\phi_A(r-R)|\mathcal{H}|\phi_B(r-R\pm a/2)\rangle. \tag{1.83}$$

Here we have assumed that all the π bonding orbitals are of equal length (1.5Å bonds). In the real $(CH)_x$ compound, bond alternation occurs, in which the bonding between adjacent carbon atoms alternates between single bonds (1.7Å) and double bonds (1.3Å). With this bond alternation, the two matrix elements between atomic wavefunctions in Eq. (1.82) are not equal. Although the distortion of the lattice lowers the total energy, the electronic energy always decreases more than the lattice energy in a one-dimensional material. This distortion deforms the lattice by a process called the Peierls instability. This instability arises for example when a distortion is introduced into a system containing a previously degenerate system with 2 equivalent atoms per unit cell. The distortion making the atoms inequivalent increases the unit cell by a factor of 2 and decreases the reciprocal lattice by a factor of 2. If the energy band was formally half filled, a band gap is introduced by the Peierls instability at the Fermi level, which lowers the total energy of the system. It is stressed that t has a negative value. The matrix element $\mathcal{H}_{BA}(r)$ is obtained from $\mathcal{H}_{AB}(r)$ through the Hermitian conjugation relation $\mathcal{H}_{BA} = \mathcal{H}^*_{AB}$, but since $\mathcal{H}_{AB}$ is real, we obtain $\mathcal{H}_{BA} = \mathcal{H}_{AB}$.

The overlap matrix $\mathcal{S}_{ij}$ can be calculated by a similar method as was used for $\mathcal{H}_{ij}$, except that the intra-atomic integral $\mathcal{S}_{ij}$ yields a unit matrix in the limit of large interatomic distances, if we assume that the atomic wavefunction is normalized so that $\mathcal{S}_{AA} = \mathcal{S}_{BB} = 1$. It is assumed that for polyacetylene the $\mathcal{S}_{AA}$ and $\mathcal{S}_{BB}$ matrix elements are still approximately unity. For the off-diagonal matrix element for polyacetylene we have $\mathcal{S}_{AB} = \mathcal{S}_{BA} = 2s\cos(ka/2)$, where s is an overlap integral between the nearest A and B atoms,

$$s = \langle \phi_A(r - R) | \phi_B(r - R \pm a/2) \rangle. \tag{1.84}$$

The secular equation for the $2p_z$ orbital of CH_x is obtained by setting the determinant of $|\mathcal{H} - E\mathcal{S}|$ to zero to obtain

$$\begin{aligned} & \begin{vmatrix} E_{2p} - E & 2(t - sE)\cos(ka/2) \\ 2(t - sE)\cos(ka/2) & E_{2p} - E \end{vmatrix} \\ = \; & (E_{2p} - E)^2 - 4(t - sE)^2\cos^2(ka/2) \\ = \; & 0 \end{aligned} \tag{1.85}$$

yielding the eigenvalues of the energy dispersion relations of Eq. (1.85)

$$E_{\pm}(\vec{k}) = \frac{E_{2p} \pm 2t\cos(ka/2)}{1 \pm 2s\cos(ka/2)}, \;\; (-\frac{\pi}{a} < k < \frac{\pi}{a}) \tag{1.86}$$

in which the + sign is associated with the bonding π-band and the $-$ sign is associated with the antibonding π^*-band, as shown in Fig. 1.9. Here it is noted that by setting E_{2p} to zero (thereby defining the origin of the energy), the levels E_+ and E_- are degenerate at $ka = \pm\pi$. Figure 1.9 is constructed for $t < 0$ and $s > 0$. Since there are two π electrons per unit cell, each with a different spin orientation, both electrons occupy the bonding π energy band. The effect of the inter-atomic bonding is to lower the total energy below E_{2p}.

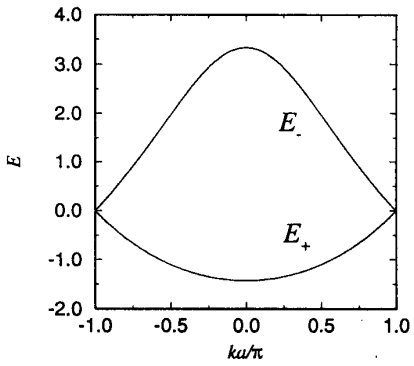


Figure 1.9: The energy dispersion relation $E_{\pm}(\vec{k})$ for polyacetylene $[(\mathrm{CH})_x]$, given by Eq. (1.86) with values for the parameters $t = -1$ and $s = 0.2$. Curves $E_{+}(\vec{k})$ and $E_{-}(\vec{k})$ are called bonding π and antibonding π^* energy bands, respectively, and the energy is plotted in units of t.

Chapter 2

Examples of Energy Bands in Solids

References

- J.C. Slater - Quantum Theory of Atoms and Molecules, Chapter 10.
- R.E. Peierls - Quantum Theory of Solids, Chapter 4
- F. Bassani and G. Pastori Paravicini - Electronic States and Optical Transitions in Solids, Chapter 4

2.1 General Issues

We present here some examples of energy bands which are representative of metals, semiconductors and insulators and point out some of the characteristic features in each case. Figure 2.1 distinguishes in a schematic way between insulators (a), metals (b), semimetals (c), a thermally excited semiconductor (d) for which at $T = 0$ all states in the valence band are occupied and all states in the conduction band are unoccupied, assuming no impurities or crystal defects. Finally in Fig. 2.1(e), we see a p-doped semiconductor which is deficient in electrons, not having sufficient electrons to fill the valence band completely as in (d).

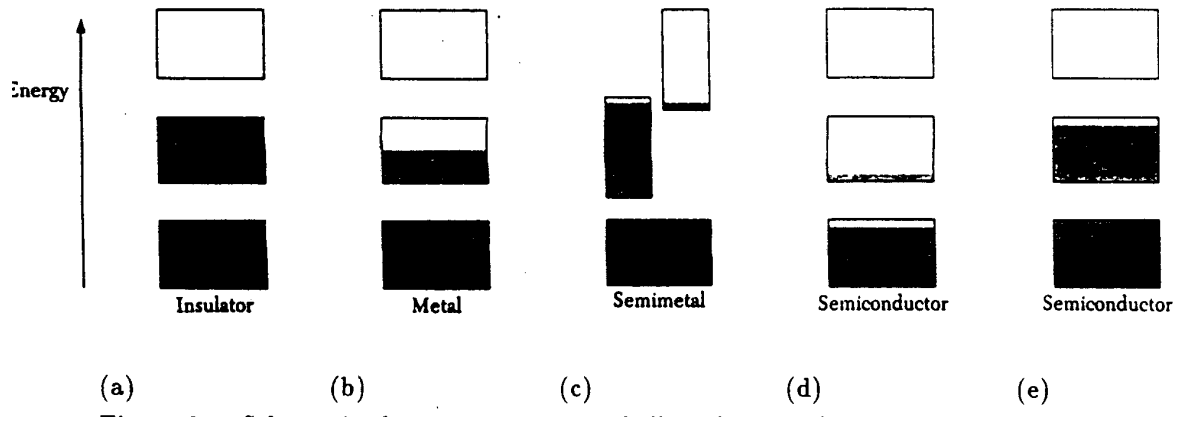
Figure 2.2 shows a schematic view of the electron dispersion relations for an insulator (a), while (c) shows dispersion relations for a metal. In the case of Fig. 2.2(b), we have a semimetal if the number of electrons equals the number of holes, but a metal otherwise.

In this chapter we examine a number of representative $E(\vec{k})$ diagrams for illustrative materials. For each of the $E(\vec{k})$ diagrams we consider the following questions:

1. Is the material a metal, a semiconductor (direct or indirect gap), semimetal or insulator?

2. To which atomic (molecular) levels do the bands on the band diagram correspond? Which bands are important in determining the electronic structure? What are the bandwidths, bandgaps?

3. What information does $E(\vec{k})$ diagram provide concerning the following questions:



(a) (b) (c) (d) (e)

Figure 2.1: Schematic electron occupancy of allowed energy bands for an insulator, metal, semimetal and semiconductor. The vertical extent of the boxes indicates the allowed energy regions: the shaded areas indicate the regions filled with electrons. In a semimetal (such as bismuth) one band is almost filled and another band is nearly empty at a temperature of absolute zero. A pure semiconductor (such as silicon) becomes an insulator at $T = 0$. Panel (d) shows a semiconductor at a finite temperature, with carriers that are thermally excited. Panel (e) shows a p-doped semiconductor that is electron-deficient, as, for example, because of the introduction of acceptor impurities.

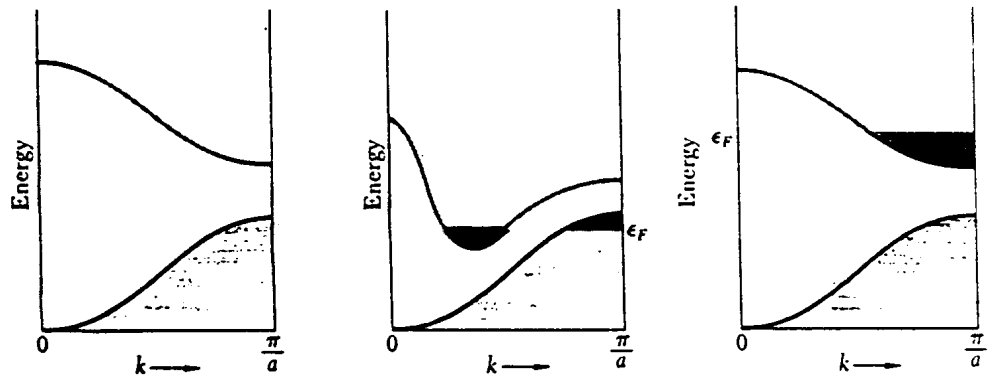


Figure 2.2: Occupied states and band structures giving (a) an insulator, (b) a metal or a semimetal because of band overlap, and (c) a metal because of partial occupation of an electron band. In (b) the band overlap need not occur along the same direction of the wave vector in the Brillouin zone.

(a) Where are the carriers in the Brillouin zone?

(b) Are the carriers electrons or holes?

(c) Are there many or few carriers?

(d) How many carrier pockets of each type are there in the Brillouin zone?

(e) What is the shape of the Fermi surface?

(f) Are the carrier velocities high or low?

(g) Are the carrier mobilities for each carrier pocket high or low?

4. What information is provided concerning the optical properties?

(a) Where in the Brillouin Zone is the threshold for optical transitions?

(b) At what photon energy does the optical threshold occur?

(c) For semiconductors, does the threshold correspond to a direct gap or an indirect gap (phonon-assisted) transition?

2.2 Metals

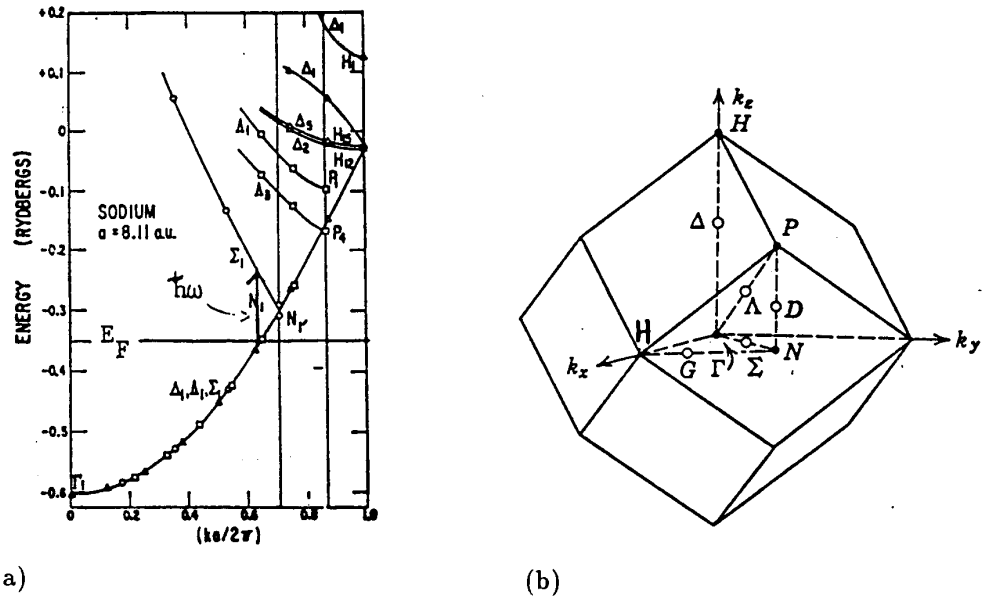
2.2.1 Alkali Metals–e.g., Sodium

For the alkali metals the valence electrons are nearly free and the weak binding approximation describes these electrons quite well. The Fermi surface is nearly spherical and the band gaps are small. The crystal structure for the alkali metals is body centered cubic (BCC) and the $E(\vec{k})$ diagram is drawn starting with the bottom of the half–filled conduction band. For example, the $E(\vec{k})$ diagram in Fig. 2.3 for sodium begins at ~ -0.6 Rydberg and represents the $3s$ conduction band. The filled valence bands lie much lower in energy and are not shown in Fig. 2.3.

For the case of sodium, the $3s$ conduction band is very nearly free electron–like and the $E(\vec{k})$ relations are closely isotropic. Thus the $E(\vec{k})$ relations along the $\Delta(100)$, $\Sigma(110)$ and $\Lambda(111)$ directions are essentially coincident and can be so plotted, as shown on the figure. For these metals, the Fermi level is determined so that the $3s$ band is exactly half–occupied, since the Brillouin zone is large enough to accommodate 2 electrons per unit cell. Thus the radius of the Fermi surface k_F satisfies the relation

$$\frac{4}{3}\pi k_F^3 = \frac{1}{2}V_{\text{B.Z.}} = \frac{1}{2}(2)(\frac{2\pi}{a})^3, \text{ or } \quad \frac{k_F\, a}{2\pi} \sim 0.63, \tag{2.1}$$

where $V_{\text{B.Z.}}$ and a are, respectively, the volume of the Brillouin zone and the lattice constant. For the alkali metals, the effective mass m^* is nearly equal to the free electron mass m and the Fermi surface is nearly spherical and never comes close to the Brillouin zone boundary. The zone boundary for the Σ, Λ and Δ directions are indicated in the $E(\vec{k})$ diagram of Fig. 2.3 by vertical lines. For the alkali metals the band gaps are very small compared to the band widths and the $E(\vec{k})$ relations are parabolic ($E = \hbar^2k^2/2m^*$) almost up to the Brillouin zone boundaries. By comparing $E(\vec{k})$ for Na with the BCC empty lattice bands (see Fig. 2.4) for which the potential $V(r) = 0$, we can see the effect of the very weak periodic potential in partially lifting the band degeneracy at the various high symmetry points in the Brillouin zone. The threshold for optical transitions corresponds to photons having



(a) (b)

Figure 2.3: (a) Energy dispersion relations for the nearly free electron metal sodium which has an atomic configuration $1s^2 2s^2 2p^6 3s$. (b) The Brillouin zone for the BCC lattice showing the high symmetry points and axes. Sodium can be considered as a prototype alkali metal crystalline solid.

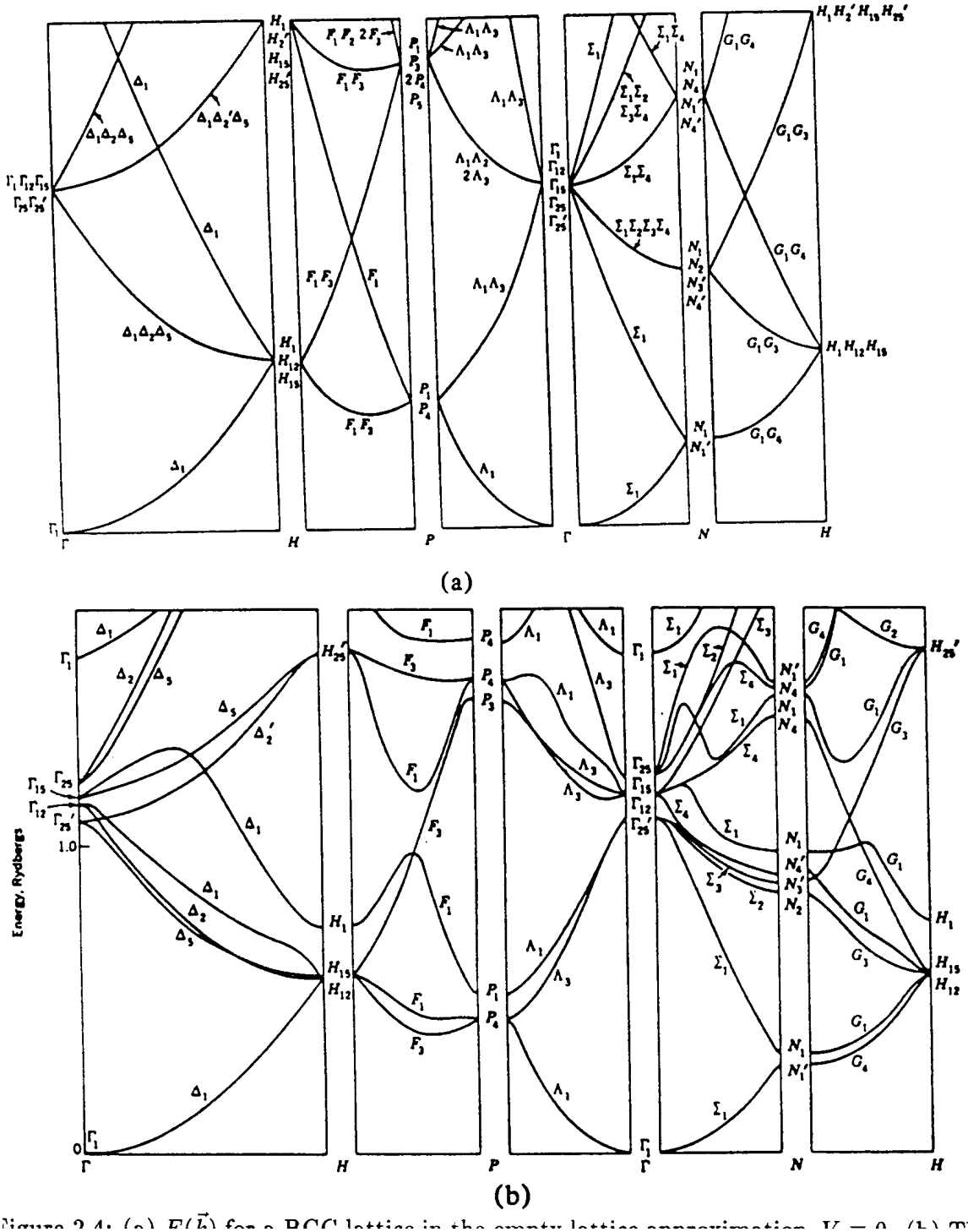
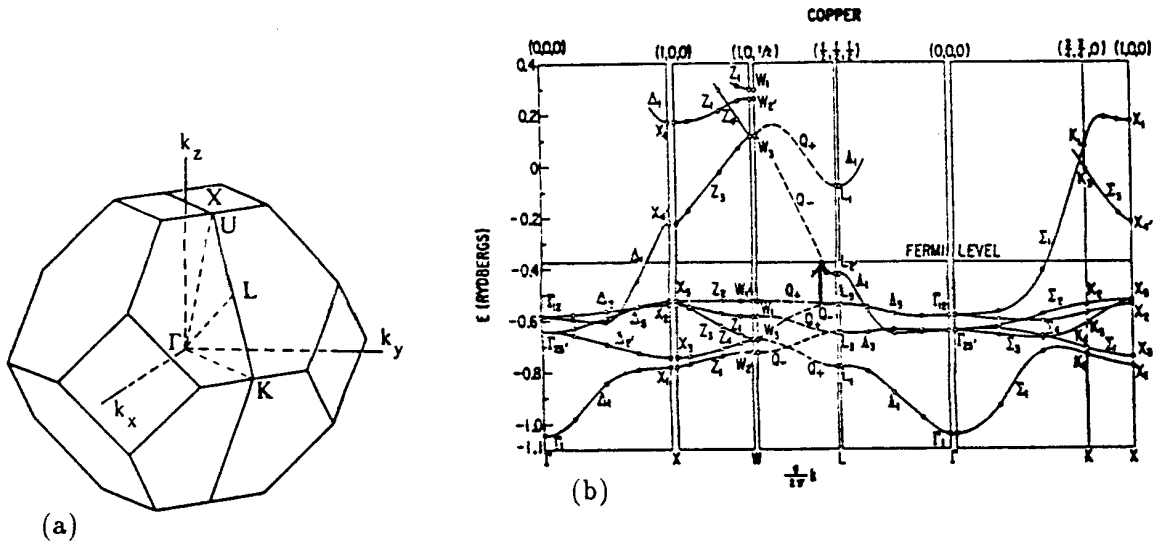


Figure 2.4: (a) $E(\vec{k})$ for a BCC lattice in the empty lattice approximation, $V \equiv 0$. (b) The same but for sodium, showing the effect of a weak periodic potential in lifting accidental band degeneracies at $k = 0$ and at the zone boundaries (high symmetry points) in the Brillouin zone. Note that the splittings are quite different for the various bands and at different high symmetry points.



(a) (b)

Figure 2.5: (a) Brillouin zone for a FCC lattice showing high symmetry points. (b) The calculated energy bands for copper along the various symmetry axes of the FCC Brillouin zone shown in (a).

sufficient energy to take an electron from an occupied state at k_F to an unoccupied state at k_F since the wave vector for photons is very small compared with k_F and wave vector conservation (i.e., crystal momentum conservation) is required for optical transitions. The threshold for optical transitions is indicated by $\hbar\omega$ in Fig. 2.3. Because of the low density of initial and final states for a given energy separation, we would expect optical interband transitions for alkali metals to be very weak and this is in agreement with experimental observations for all the alkali metals. The notation a.u. in Fig. 2.3 stands for atomic units and expresses lattice constants in units of Bohr radii. The electron energy is given in Rydbergs where 1 Rydberg = 13.6 eV, the ionization energy of a hydrogen atom.

2.2.2 Noble Metals

The noble metals are copper, silver and gold and they crystallize in a face centered cubic (FCC) structure; the usual notation for the high symmetry points in the FCC Brillouin zone are shown on the diagram in Fig. 2.5(a). As in the case of the alkali metals, the noble metals have one valence electron/atom and therefore one electron per primitive unit cell. However, the free electron picture does not work so well for the noble metals, as you can see by looking at the energy band diagram for copper given in Fig. 2.5(b).

In the case of copper, the bands near the Fermi level are derived from the $4s$ and $3d$ atomic levels. The so-called $4s$ and $3d$ bands accommodate a total of 12 electrons, while the number of available electrons is 11. Therefore the Fermi level must cross these bands. Consequently copper is metallic. In Fig. 2.5(b) we see that the $3d$ bands are relatively flat and show little dependence on wave vector $\vec{k}$. We can trace the $3d$ bands by starting at $\vec{k} = 0$ with the $\Gamma_{25'}$ and Γ_{12} levels. On the other hand, the $4s$ band has a strong k–dependence

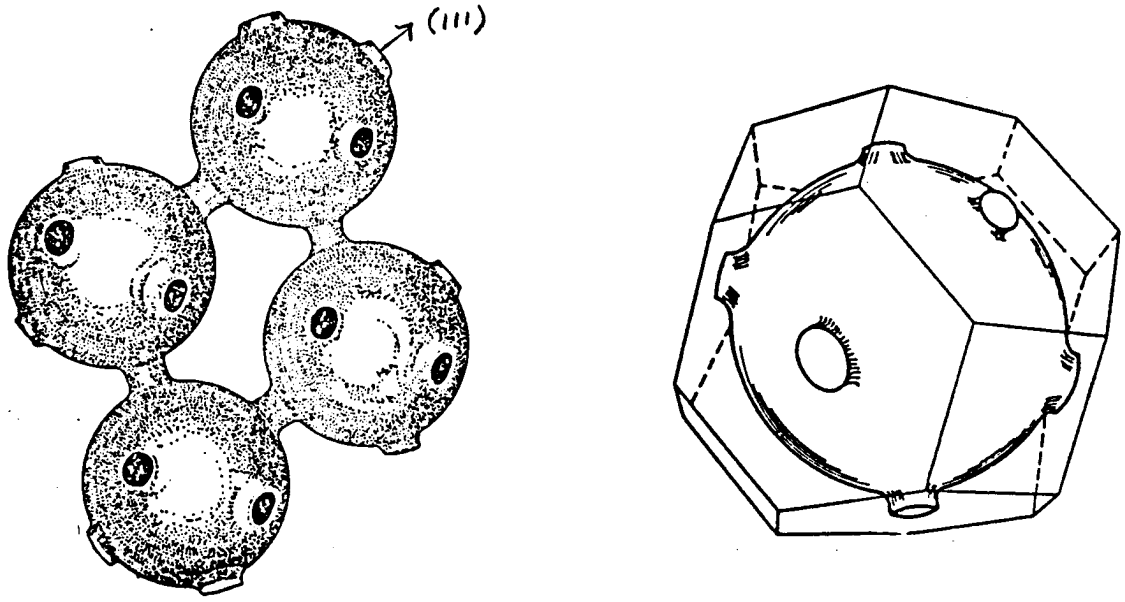


Figure 2.6: (a) The copper Fermi surface in the extended zone scheme. (b) A sketch of the Fermi surface of copper inscribed within the FCC Brillouin zone.

and large curvature. This band can be traced by starting at $\vec{k} = 0$ with the Γ_1 level. About halfway between Γ and X, the $4s$ level approaches the $3d$ levels and *mixing* or *hybridization* occurs. As we further approach the X–point, we can again pick up the $4s$ band (beyond where the interaction with the $3d$ bands occurs) because of its high curvature. This $4s$ band eventually crosses the Fermi level before reaching the Brillouin Zone boundary at the X point. A similar mixing or hybridization between $4s$ and $3d$ bands occurs in going from Γ to L, except that in this case the $4s$ band reaches the Brillouin Zone boundary *before* crossing the Fermi level.

Of particular significance for the transport properties of copper is the band gap that opens up at the L–point. In this case, the band gap is between the $L_{2'}$ level below the Fermi level E_F and the L_1 level above E_F. Since this bandgap is comparable with the typical bandwidths in copper, we cannot expect the Fermi surface to be free electron–like. By looking at the energy bands $E(\vec{k})$ along the major high symmetry directions such as the (100), (110) and (111) directions, we can readily trace the origin of the copper Fermi surface [see Fig. 2.6(a)]. Here we see basically a spherical Fermi surface with necks pulled out in the (111) directions and making contact with the Brillouin zone boundary through these necks, thereby linking the Fermi surface in one zone to that in the next zone in the extended zone scheme. In the (100) direction, the cross section of the Fermi surface is nearly circular, indicative of the nearly parabolic $E(\vec{k})$ relation of the $4s$ band at the Fermi level in going from Γ to X. In contrast, in going from Γ to L, the $4s$ band never crosses the Fermi level. Instead the $4s$ level is depressed from the free electron parabolic curve as the Brillouin zone boundary is reached, thereby producing a higher density of states. Thus,

near the zone boundary, more electrons can be accommodated per unit energy range, or to say this another way, there will be increasingly more $\vec{k}$ vectors with approximately the same energy. This causes the constant energy surfaces to be pulled out in the direction of the Brillouin zone boundary [see Fig. 2.6(b)]. This "pulling out" effect follows both from the weak binding and tight binding approximations and the effect is more pronounced as the strength of the periodic potential (or $V_{\vec{G}}$) increases.

If the periodic potential is sufficiently strong so that the resulting bandgap at the zone boundary straddles the Fermi level, as occurs at the L–point in copper, the Fermi surface makes contact with the Brillouin zone boundary. The resulting Fermi surfaces are called *open surfaces* because the Fermi surfaces between neighboring Brillouin zones are connected, as seen in Fig. 2.6(a). The electrons associated with the necks are contained in the electron pocket shown in the $E(\vec{k})$ diagram away from the L–point in the LW direction which is $\perp$ to the $\{111\}$ direction. The copper Fermi surface shown in Fig. 2.6(a) bounds *electron* states. Hole pockets are formed in copper [see Fig. 2.6(a)] in the extended zone and constitute the unoccupied space between the electron surfaces. Direct evidence for hole pockets is provided by Fermi surface measurements to be described later in this course.

From the $E(\vec{k})$ diagram for copper [Fig. 2.5(b)] we see that the threshold for optical interband transitions occurs for photon energies sufficient to take an electron at constant $\vec{k}$–vector from a filled $3d$ level to an unoccupied state above the Fermi level. Such interband transitions can be made near the L–point in the Brillouin zone [as shown by the arrow on Fig. 2.5(b)]. Because of the high density of initial states in the d–band, these transitions will be quite intense. The occurrence of these interband transitions at ~ 2 eV gives rise to a large absorption of electromagnetic energy in this photon energy region. The reddish color of copper metal is thus due to a higher reflectivity for photons in the red (below the threshold for interband transitions) than for photons in the blue (above this threshold).

2.2.3 Polyvalent Metals

The simplest example of a polyvalent metal is aluminum with 3 electrons/atom and having a $3s^2 3p$ electronic configuration for the valence electrons. (As far as the number of electrons/atom is concerned, two electrons/atom completely fill a non-degenerate band—one for spin up, the other for spin down.) Because of the partial filling of the $3s^2 3p^6$ bands, aluminum is a metal. Aluminum crystallizes in the FCC structure so we can use the same notation as for the Brillouin zone in Fig. 2.5(a). The energy bands for aluminum (see Fig. 2.7) are very free electron–like. This follows from the small magnitudes of the band gaps relative to the band widths on the energy band diagram shown in Fig. 2.7. The lowest valence band shown in Fig. 2.7 is the $3s$ band which can be traced by starting at zero energy at the Γ point ($\vec{k} = 0$) and going out to X_4 at the X–point, to W_3 at the W–point, to L_2' at the L–point and back to Γ_1 at the Γ point ($\vec{k} = 0$). Since this band always lies below the Fermi level, it is completely filled, containing 2 electrons. The third valence electron partially occupies the second and third p–bands (which are more accurately described as hybridized $3p$–bands with some admixture of the $3s$ bands with which they interact). From Fig. 2.7 we can see that the second band is partly filled; the occupied states extend from the Brillouin zone boundary inward toward the center of the zone; this can be seen in going from the X point to Γ, on the curve labeled Δ_1. Since the second band states near the center of the Brillouin zone remain unoccupied, the volume enclosed by the Fermi surface

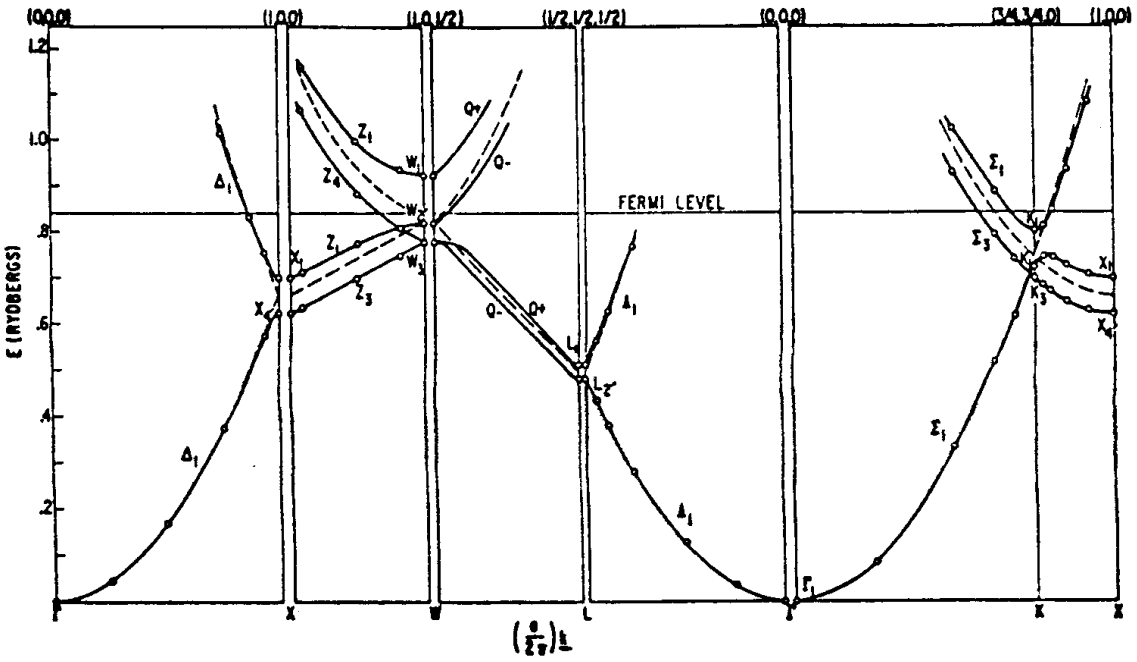


Figure 2.7: Electronic energy band diagram for aluminum which crystallizes in a FCC structure. The dashed lines correspond to the free electron model and the solid curves include the effect of the periodic potential $V(\vec{r})$.

in the second band is a hole pocket. Because $E(\vec{k})$ for the second band in the vicinity of E_F is free electron–like, the masses for the holes are approximately equal to the free electron mass.

The 3^{rd} zone electron pockets are small and are found around the K– and W–points as can be seen in Fig. 2.7. These pockets are $\vec{k}$ space volumes that enclose electron states, and because of the large curvature of $E(\vec{k})$, these electrons have relatively small masses. This diagram gives no evidence for any 4^{th} zone pieces of Fermi surface, and for this reason we can conclude that all the electrons are either in the second band or in the third band. The total electron concentration is sufficient to exactly fill a half of a Brillouin zone:

$$V_{e,2} + V_{e,3} = \frac{V_{BZ}}{2}. \tag{2.2}$$

With regard to the second zone, it is partially filled with electrons and the rest of the zone is empty (since holes correspond to the unfilled states):

$$V_{h,2} + V_{e,2} = V_{BZ}, \tag{2.3}$$

with the volume that is empty slightly exceeding the volume that is occupied. Therefore we focus attention on the more dominant second zone holes. Substitution of Eq. 2.2 into Eq. 2.3 then yields for the second zone holes and the third zone electrons

$$V_{h,2} - V_{e,3} = \frac{V_{BZ}}{2} \tag{2.4}$$

where the subscripts e, h on the volumes in $\vec{k}$ space refer to electrons and holes and the Brillouin zone (B.Z.) index is given for each of the carrier pockets. Because of the small masses and high mobility of the 3^{rd} zone electrons, they play a more important role in the transport properties of aluminum than would be expected from their small numbers.

From the $E(\vec{k})$ diagram in Fig. 2.7 we see that at the same $\vec{k}$–points (near the K– and W–points in the Brillouin zone) there are occupied $3s$ levels and unoccupied $3p$ levels separated by $\sim$ 1eV. From this we conclude that optical interband transitions should be observable in the 1eV photon energy range. Such interband transitions are in fact observed experimentally and are responsible for the departures from nearly perfect reflectivity of aluminum mirrors in the vicinity of 1 eV.

2.3 Semiconductors

Assume that we have a semiconductor at $T = 0$ K with no impurities. The Fermi level will then lie within a band gap. Under these conditions, there are no carriers, and no Fermi surface. We now illustrate the energy band structure for several representative semiconductors in the limit of $T = 0$K and no impurities. Semiconductors having no impurities or defects are called *intrinsic* semiconductors.

2.3.1 PbTe

In Fig. 2.8 we illustrate the energy bands for PbTe. This direct gap semiconductor [see Fig. 2.9(a)] is chosen initially for illustrative purposes because the energy bands in the va-

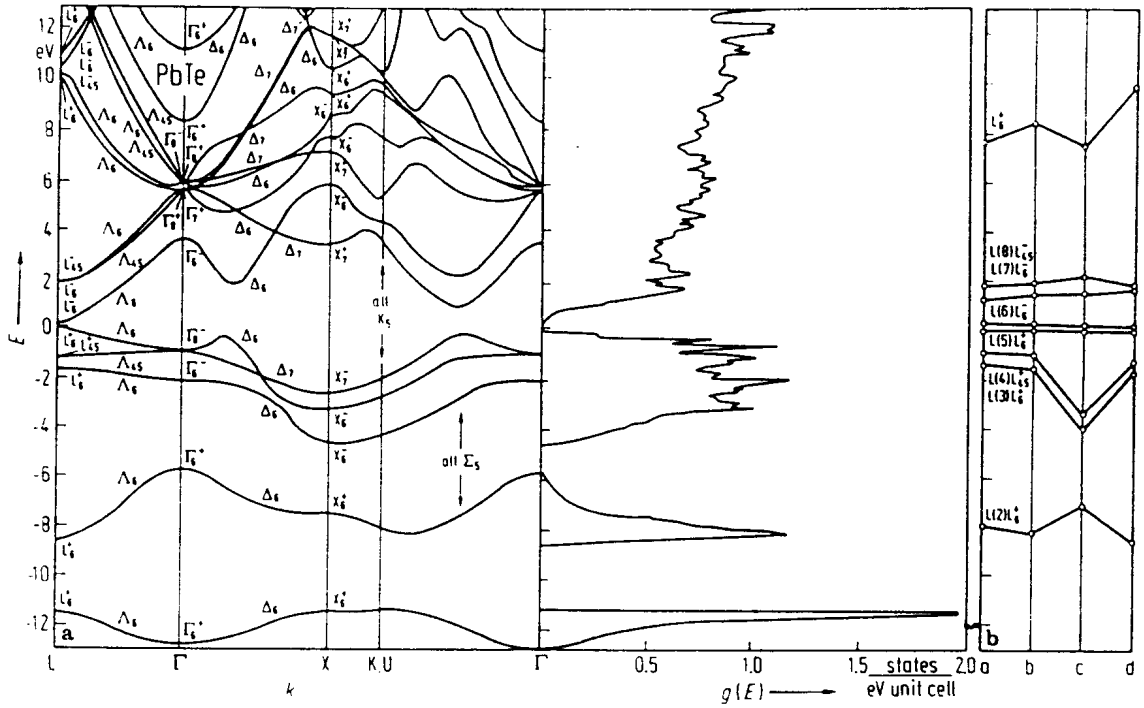


Figure 2.8: (a) Energy band structure and density of states for PbTe obtained from an empirical pseudopotential calculation. (b) Theoretical values for the L point bands calculated by different models (labeled a, b, c, d on the x-axis) (Ref. Landolt and Bornstein).

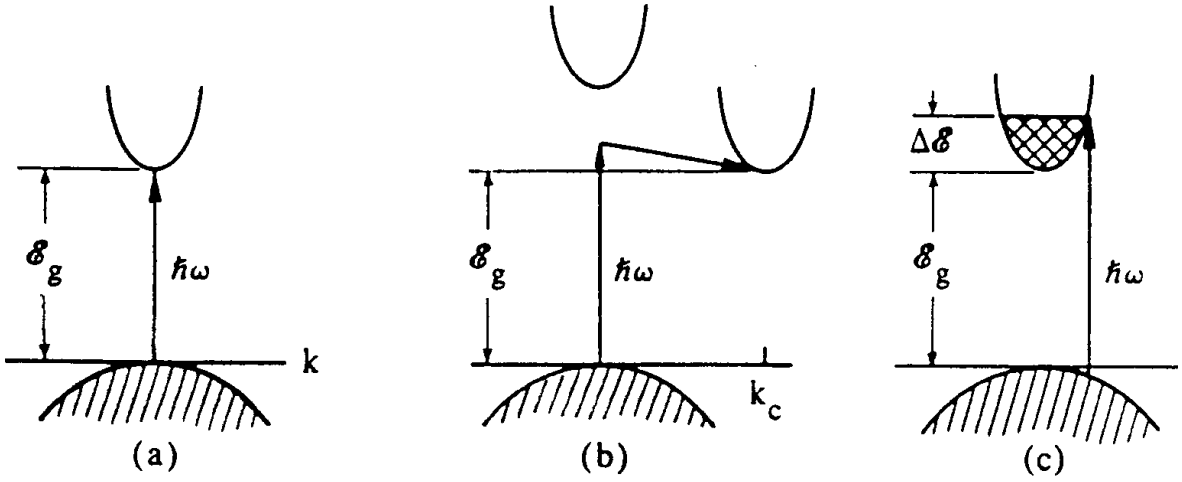


Figure 2.9: Optical absorption processes for (a) a direct band gap semiconductor, (b) an indirect band gap semiconductor, and (c) a direct band gap semiconductor with the conduction band filled to the level shown.

lence and conduction bands that are of particular interest are non-degenerate. Therefore, the energy states in PbTe near E_F are simpler to understand than for the more common semiconductors silicon and germanium, and many of the III–V and II–VI compound semiconductors.

In Fig. 2.8, we show the position of E_F for the idealized conditions of the intrinsic (no carriers at $T = 0$) semiconductor PbTe. From a diagram like this, we can obtain a great deal of information which could be useful for making semiconductor devices. For example, we can calculate effective masses from the band curvatures, and electron velocities from the slopes of the $E(\vec{k})$ dispersion relations shown in Fig. 2.8.

Suppose we add impurities (e.g., donor impurities) to PbTe. The donor impurities will raise the Fermi level and an electron pocket will eventually be formed in the L_6^- conduction band about the L–point. This electron pocket will have an ellipsoidal Fermi surface because the band curvature is different as we move away from the L point in the $L\Gamma$ direction as compared with the band curvature as we move away from L on the Brillouin zone boundary containing the L point (e.g., LW direction). Figure 2.8 shows $E(\vec{k})$ from L to Γ corresponding to the (111) direction. Since the effective masses

$$\frac{1}{m_{ij}^*} = \frac{1}{\hbar^2}\frac{\partial^2 E(\vec{k})}{\partial k_i \partial k_j} \tag{2.5}$$

for both the valence and conduction bands in the longitudinal $L\Gamma$ direction are heavier than in the LK and LW directions, the ellipsoids of revolution describing the carrier pockets are prolate for both holes and electrons. The L and Σ point room temperature band gaps are 0.311 eV and 0.360 eV, respectively. For the electrons, the effective mass parameters are $m_\perp = 0.053m_e$ and $m_\parallel = 0.620m_e$. The experimental hole effective masses at the L point are $m_\perp = 0.0246m_e$ and $m_\parallel = 0.236m_e$ and at the Σ point the hole effective mass values are $m_\perp = 0.124m_e$ and $m_\parallel = 1.24m_e$. Thus for the L-point carrier pockets, the semi-major axis of the constant energy surface along $L\Gamma$ will be longer than along LK. From the $E(\vec{k})$ diagram for PbTe in Fig. 2.8 one would expect that hole carriers could be thermally excited to a second band at the Σ point, which is indicated on the $E(\vec{k})$ diagram. At room temperature, these Σ point hole carriers contribute significantly to the transport properties.

Because of the small gap (0.311 eV) in PbTe at the L–point, the threshold for interband transitions will occur at infrared frequencies. PbTe crystals can be prepared either p–type or n–type, but never perfectly stoichiometrically (i.e., intrinsic PbTe has not been prepared). Therefore, at room temperature the Fermi level E_F often lies in either the valence or conduction band for actual PbTe crystals. Since optical transitions conserve wavevector, the interband transitions will occur at k_F [see Fig. 2.9(c)] and at a higher photon energy than the direct band gap. This increase in the threshold energy for interband transitions in degenerate semiconductors (where E_F lies within either the valence or conduction bands) is called the *Burstein shift*.

2.3.2 Germanium

We will next look at the $E(\vec{k})$ relations for: (1) the group IV semiconductors which crystallize in the diamond structure and (2) the closely related III–V compound semiconductors which crystallize in the zincblende structure (see Fig. 2.10 for a schematic diagram for this class of semiconductors). These semiconductors have degenerate valence bands at $\vec{k} = 0$

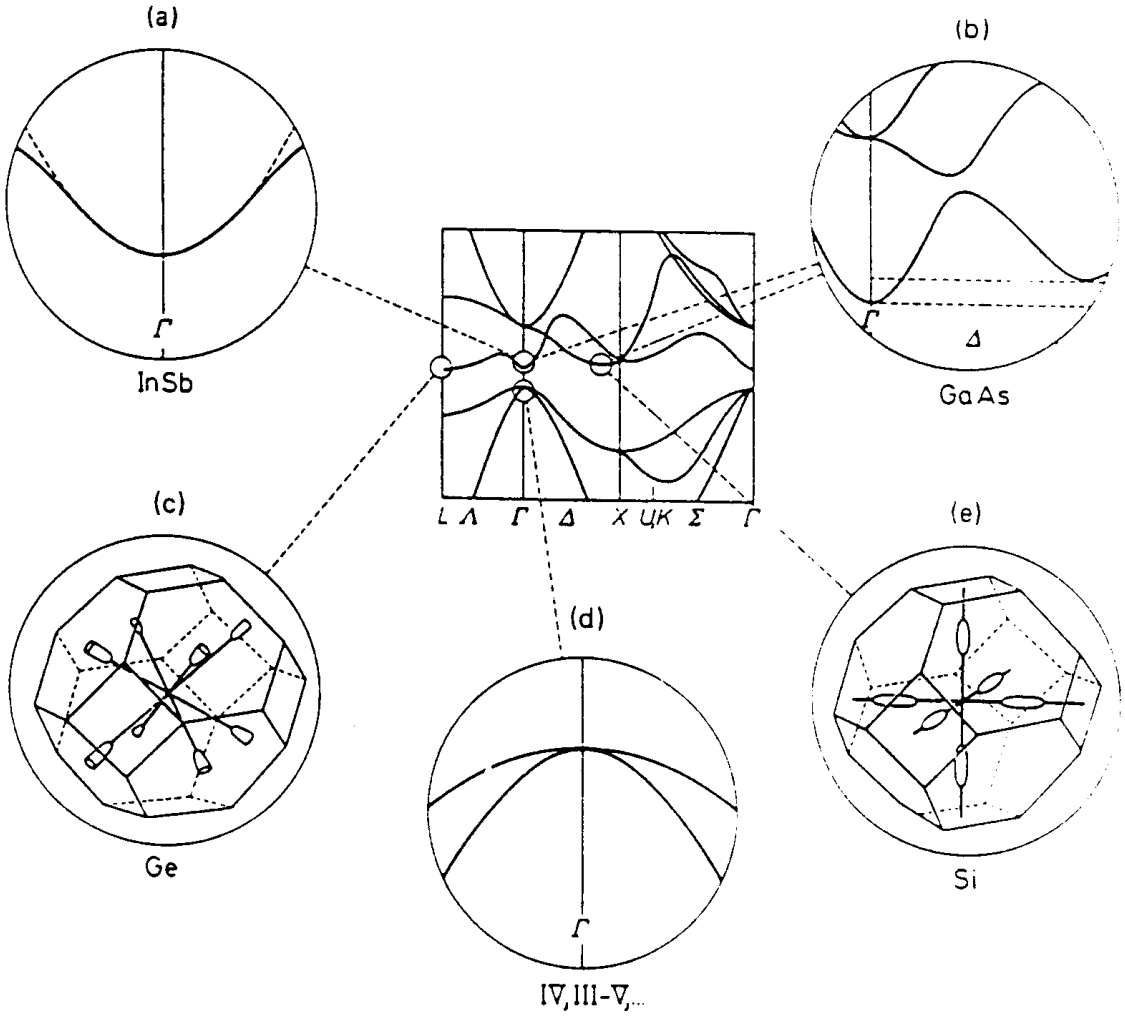
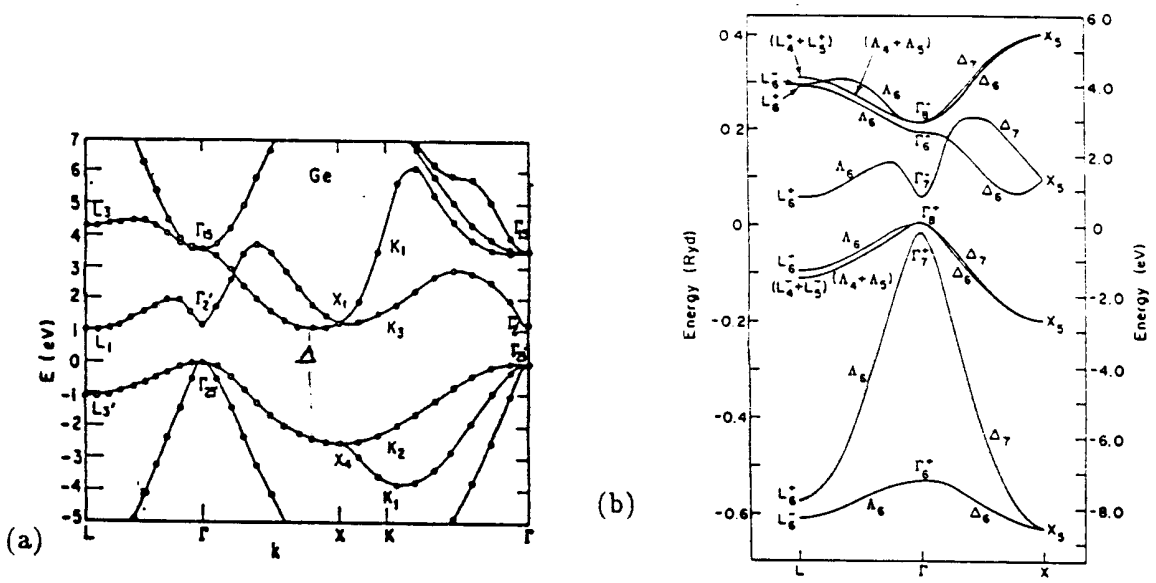


Figure 2.10: Important details of the band structure of typical group IV and III–V semiconductors.



(a) (b)

Figure 2.11: Electronic energy band structure of Ge (a) without spin-orbit interaction. (b) The electronic energy bands near $k = 0$ when the spin-orbit interaction is included.

[see Fig. 2.10(d)] and for this reason have more complicated $E(\vec{k})$ relations for carriers than is the case for the lead salts discussed in §2.3.1. The $E(\vec{k})$ diagram for germanium is shown in Fig. 2.11. Ge is a semiconductor with a bandgap occurring between the top of the valence band at $\Gamma_{25'}$, and the bottom of the lowest conduction band at L_1. Since the valence and conduction band extrema occur at *different* points in the Brillouin zone, Ge is an *indirect gap* semiconductor [see Fig. 2.9(b)]. Using the same arguments as were given in §2.3.1 for the Fermi surface of PbTe, we see that the constant energy surfaces for electrons in germanium are ellipsoids of revolution [see Fig. 2.10(c)]. As for the case of PbTe, the ellipsoids of revolution are elongated along ΓL which is the heavy mass direction in this case. Since the multiplicity of L–points is 8, we have 8 half–ellipsoids of this kind within the first Brillouin zone, just as for the case of PbTe. By translation of these half–ellipsoids by a reciprocal lattice vector we can form 4 full–ellipsoids. The $E(\vec{k})$ diagram for germanium (see Fig. 2.11) further shows that the next highest conduction band above the L point minimum is at the Γ–point ($\vec{k}$=0) and after that along the ΓX axis at a point commonly labeled as a Δ–point. Because of the degeneracy of the highest valence band, the Fermi surface for holes in germanium is more complicated than for electrons. The lowest direct band gap in germanium is at $\vec{k} = 0$ between the $\Gamma_{25'}$ valence band and the $\Gamma_{2'}$ conduction band. From the $E(\vec{k})$ diagram we note that the electron effective mass for the $\Gamma_{2'}$ conduction band is very small because of the high curvature of the $\Gamma_{2'}$ band about $\vec{k} = 0$, and this effective mass is isotropic so that the constant energy surfaces are spheres.

The optical properties for germanium show a very weak optical absorption for photon energies corresponding to the indirect gap (see Fig. 2.12). Since the valence and conduction band extrema occur at a different $\vec{k}$–point in the Brillouin zone, the indirect gap excitation requires a phonon to conserve crystal momentum. Hence the threshold for this indirect

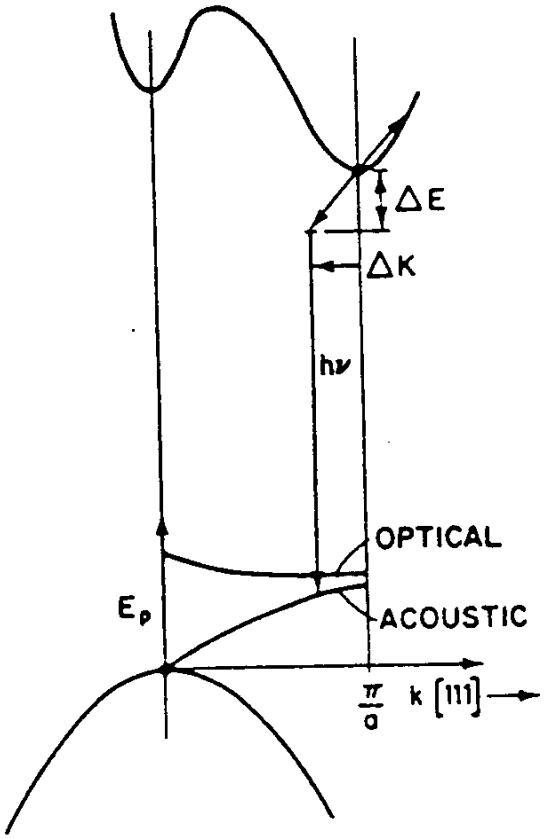


Figure 2.12: Illustration of the indirect emission of light due to carriers and phonons in Ge. [$h\nu$ is the photon energy; ΔE is the energy delivered to an electron; E_p is the energy delivered to the lattice (phonon energy)].

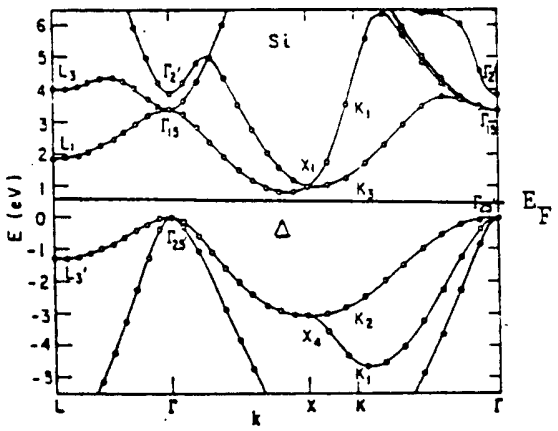


Figure 2.13: Electronic energy band structure of Si.

transition is

$$(\hbar\omega)_{\text{threshold}} = E_{L_1} - E_{\Gamma_{25'}} - E_{\text{phonon}}. \tag{2.6}$$

The optical absorption for germanium increases rapidly above the photon energy corresponding to the direct band gap $E_{\Gamma_{2'}} - E_{\Gamma_{25'}}$, because of the higher probability for the direct optical excitation process. However, the absorption here remains low compared with the absorption at yet higher photon energies because of the low density of states for the Γ–point transition, as seen from the $E(\vec{k})$ diagram. Very high optical absorption, however, occurs for photon energies corresponding to the energy separation between the $L_{3'}$ and L_1 bands which is approximately the same for a large range of $\vec{k}$ values, thereby giving rise to a very large joint density of states (the number of states with constant energy separation per unit energy range). A large *joint density of states* arising from the *tracking* of conduction and valence bands is found for germanium, silicon and the III–V compound semiconductors, and for this reason these materials tend to have high dielectric constants (to be discussed in Part II of this course which focuses on optical properties).

2.3.3 Silicon

From the energy band diagram for silicon shown in Fig. 2.13, we see that the energy bands of Si are quite similar to those for germanium. They do, however, differ in detail. For example, in the case of silicon, the electron pockets are formed around a Δ point located along the ΓX (100) direction. For silicon there are 6 electron pockets within the first Brillouin zone instead of the 8 half–pockets which occur in germanium. The constant energy surfaces are again ellipsoids of revolution with a heavy longitudinal mass and a light transverse effective mass [see Fig. 2.10(e)]. The second type of electron pocket that is energetically favored is about the L_1 point, but to fill electrons there, we would need to raise the Fermi energy by ~ 1 eV.

Silicon is of course the most important semiconductor for device applications and is

at the heart of semiconductor technology for transistors, integrated circuits, and many electronic devices. The optical properties of silicon also have many similarities to those in germanium, but show differences in detail. For Si, the indirect gap [see Fig. 2.9(b)] occurs at ~ 1 eV and is between the $\Gamma_{25'}$ valence band and the Δ conduction band extrema. Just as in the case for germanium, strong optical absorption occurs for large volumes of the Brillouin zone at energies comparable to the $L_{3'} \rightarrow L_1$ energy separation, because of the "tracking" of the valence and conduction bands. The density of electron states for Si covering a wide energy range is shown in Fig. 2.14 where the corresponding energy band diagram is also shown. Most of the features in the density of states can be identified with the band model.

2.3.4 III–V Compound Semiconductors

Another important class of semiconductors is the III–V compound semiconductors which crystallize in the zincblende structure; this structure is like the diamond structure except that the two atoms/unit cell are of a different chemical species. The III–V compounds also have many practical applications, such as semiconductor lasers for fast electronics, GaAs in light emitting diodes, and InSb for infrared detectors. In Fig. 2.15 the $E(\vec{k})$ diagram for GaAs is shown and we see that the electronic levels are very similar to those of Si and Ge. One exception is that the lowest conduction band for GaAs is at $\vec{k} = 0$ so that both valence and conduction band extrema are at $\vec{k}$=0. Thus GaAs is a *direct* gap semiconductor [see Fig. 2.9(a)], and for this reason GaAs shows a stronger and more sharply defined optical absorption threshold than Si or Ge. Figure 2.10(b) shows a schematic of the conduction bands for GaAs. Here we see that the lowest conduction band for GaAs has high curvature and therefore a small effective mass. This mass is isotropic so that the constant energy surface for electrons in GaAs is a sphere and there is just one such sphere in the Brillouin zone. The next lowest conduction band is at a Δ point and a significant carrier density can be excited into this Δ point pocket at high temperatures.

The constant energy surface for electrons in the direct gap semiconductor InSb shown in Fig. 2.16 is likewise a sphere, because InSb is also a direct gap semiconductor. InSb differs from GaAs in having a very small band gap (~ 0.2 eV), occurring in the infrared. Both direct and indirect band gap materials are found in the III–V compound semiconductor family. Except for optical phenomena close to the band gap, these compound semiconductors all exhibit very similar optical properties which are associated with the band-tracking phenomena discussed in §2.3.2.

2.3.5 "Zero Gap" Semiconductors – Gray Tin

It is also possible to have "zero gap" semiconductors. An example of such a material is gray tin which also crystallizes in the diamond structure. The energy band model for gray tin without spin–orbit interaction is shown in Fig. 2.17(a). On this diagram the zero gap occurs between the $\Gamma_{25'}$ valence band and the $\Gamma_{2'}$ conduction band, and the Fermi level runs right through this degeneracy point between these bands. Spin–orbit interaction (to be discussed later in this course) is very important for gray tin in the region of the $\vec{k} = 0$ band degeneracy and a detailed diagram of the energy bands near $\vec{k} = 0$ and including the effect of spin–orbit interaction is shown in Fig. 2.17(b). In gray tin the effective mass for the conduction band is much lighter than for the valence band as can be seen by the band curvatures shown in Fig. 2.17(b).

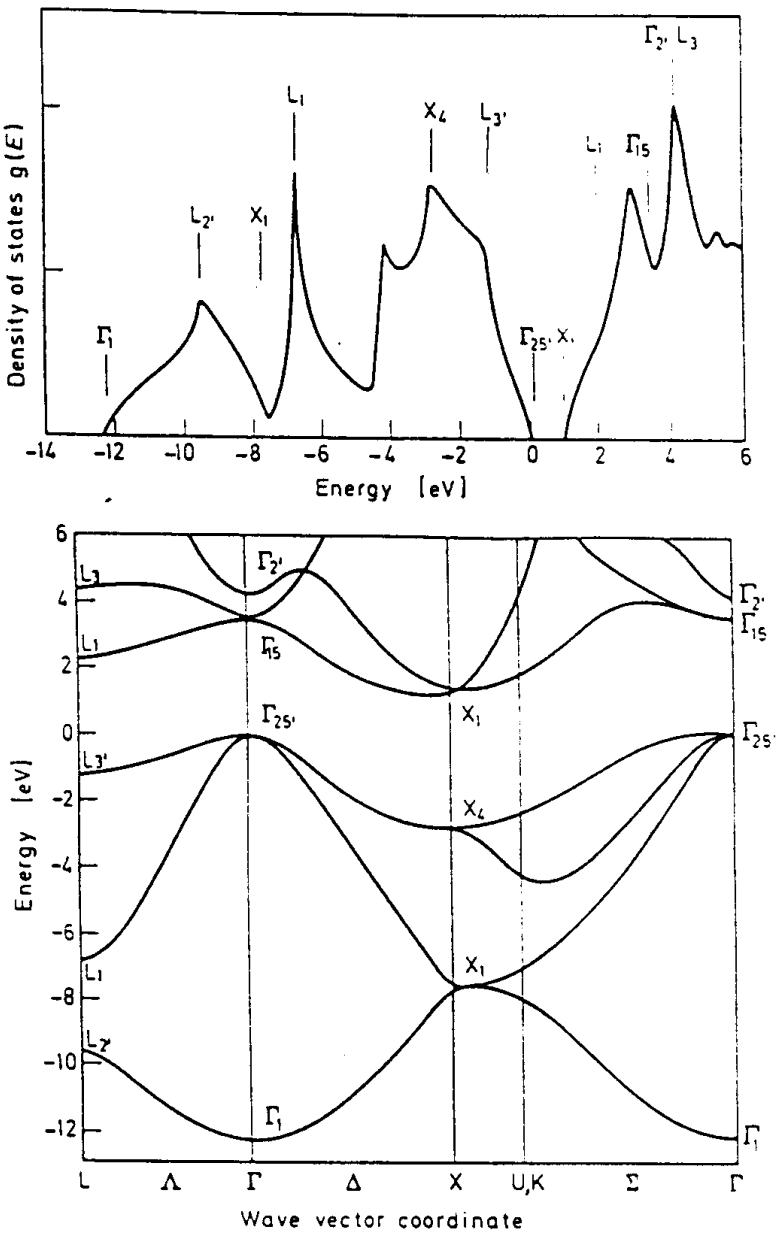


Figure 2.14: (a) Density of states in the valence and conduction bands of silicon, and (b) the corresponding $E(\vec{k})$ curves showing the symbols of the high symmetry points of the band structure.

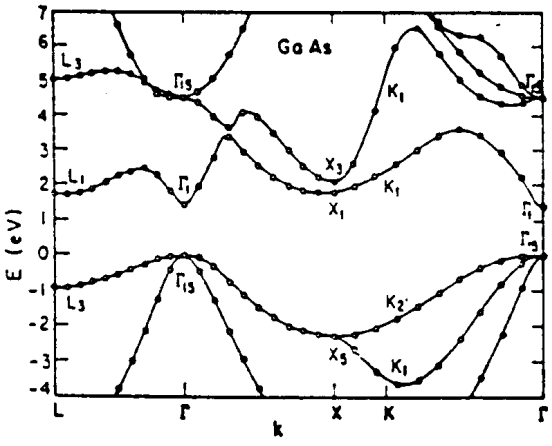


Figure 2.15: Electronic energy band structure of the III-V compound GaAs.

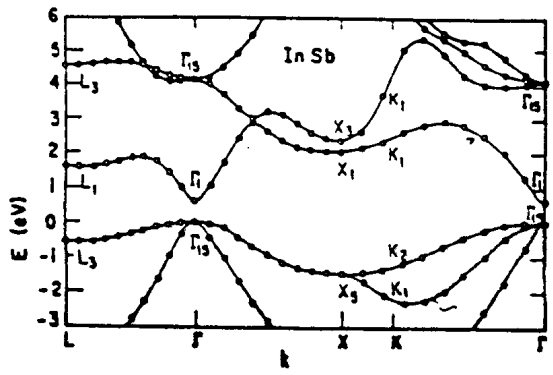


Figure 2.16: Electronic energy band structure of the III-V compound InSb.

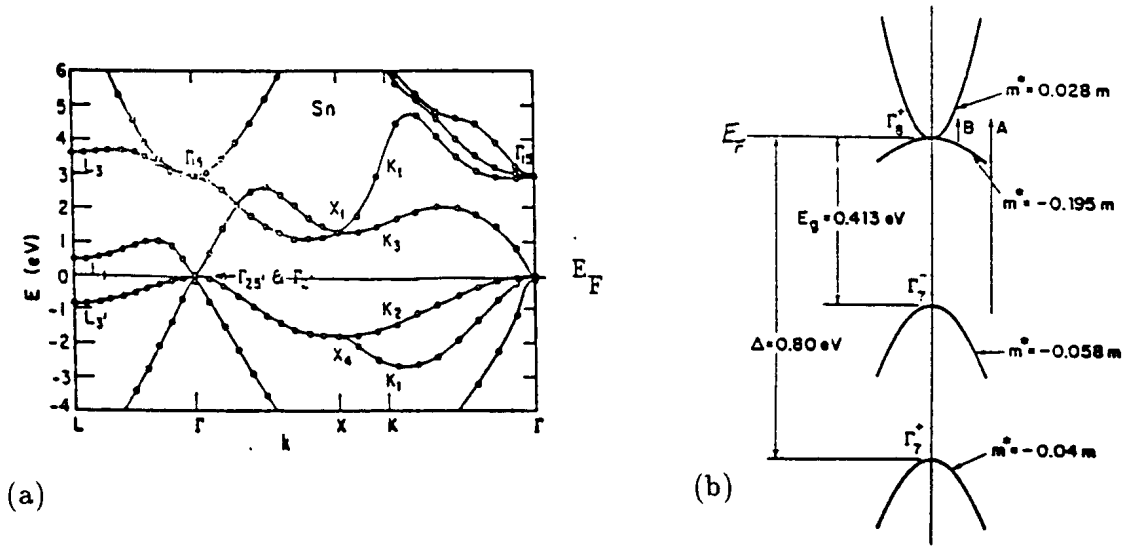


Figure 2.17: (a) Electronic energy band structure of gray Sn neglecting the spin-orbit interaction. (b) Detailed diagram of the energy bands of gray tin near $k = 0$ including the spin-orbit interaction. The Fermi level goes through the degenerate point between the filled valence band and the empty conduction band in the idealized model for gray tin at $T = 0$. The Γ_7^- hole band has the same symmetry as the conduction band for Ge when spin–orbit interaction is included, as shown in Fig. 2.11(b). The Γ_7^+ hole band has the same symmetry as the "split–off" valence band for Ge when spin–orbit interaction is included.

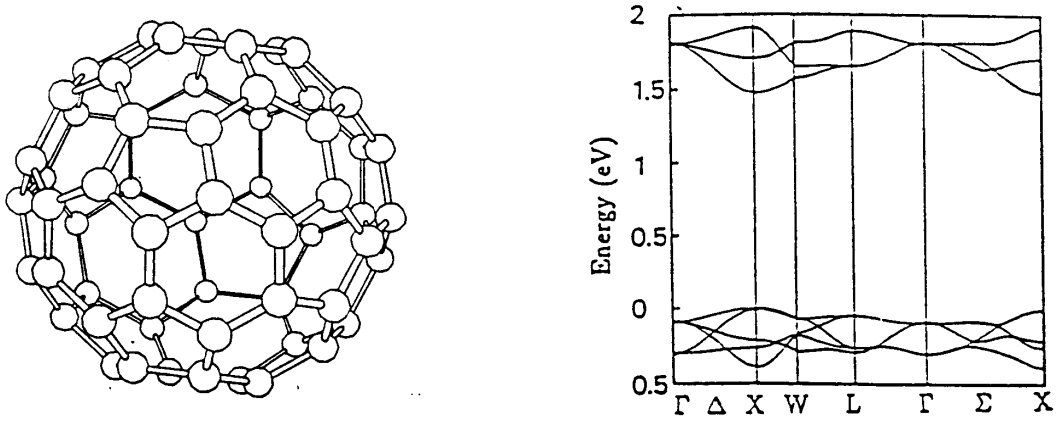


Figure 2.18: (a)Structure of the icosahedral C_{60} molecule, and (b) the calculated one-electron electronic energy band structure of FCC solid C_{60}. The Fermi energy lies between the occupied valence levels and the empty conduction levels.

Optical transitions in Fig. 2.17(b) labeled B occur in the far infrared from the upper valence band to the conduction band. In the near infrared, interband transitions labeled A are induced from the Γ_7^- valence band to the Γ_8^+ conduction band. We note that gray tin is classified as a zero gap semiconductor rather than a semimetal (see §2.4) because there are no band overlaps in a zero-gap semiconductor anywhere in the Brillouin zone. Because of the zero band gap in grey tin, impurities play a major role in determining the position of the Fermi level. Gray tin is normally prepared n-type which means that there are some electrons always present in the conduction band (for example a typical electron concentration would be $10^{15}/\text{cm}^3$ which amounts to less than 1 carrier/10^7 atoms).

2.3.6 Molecular Semiconductors – Fullerenes

Other examples of semiconductors are molecular solids such as C_{60} (see Fig. 2.18). For the case of solid C_{60}, we show in Fig. 2.18(a) a C_{60} molecule, which crystallizes in a FCC structure with four C_{60} molecules per conventional simple cubic unit cell. A small distortion of the bonds, lengthening the C–C bond lengths on the single bonds to 1.46Å and shortening the double bonds to 1.40Å, stabilizes a band gap of ∼1.5eV [see Fig. 2.18(b)]. In this semiconductor the energy bandwidths are very small compared with the band gaps, so that this material can be considered as an organic molecular semiconductor. The transport properties of C_{60} differ markedly from those for conventional group IV or III–V semiconductors.

2.4 Semimetals

Another type of material that commonly occurs in nature is the *semimetal*. Semimetals have exactly the correct number of electrons to completely fill an integral number of Brillouin zones. Nevertheless, in a semimetal the highest occupied Brillouin zone is not filled up completely, since some of the electrons find lower energy states in "higher" zones (see

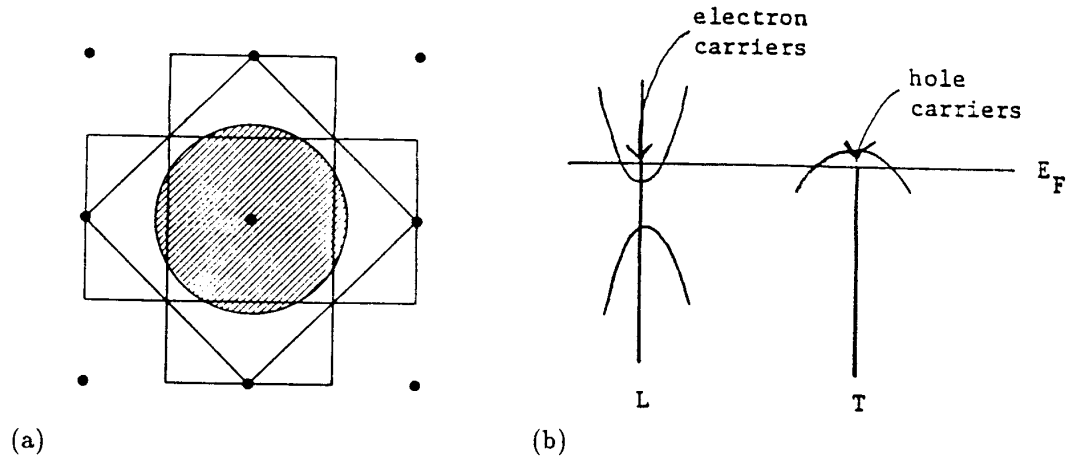


Figure 2.19: (a) Schematic diagram of a semimetal in two dimensions. (b) Schematic diagram of the energy bands $E(\vec{k})$ of bismuth showing electron pockets at the L point and hole pockets at the T point. The T point is the $\{111\}$ direction along which a stretching distortion occurs in real space and the L points refer to the 3 other equivalent $\{1\bar{1}\bar{1}\}$, $\{\bar{1}1\bar{1}\}$, and $\{\bar{1}\bar{1}1\}$ directions.

Fig. 2.2). For semimetals the number of electrons that spill over into a higher Brillouin zone is exactly equal to the number of holes that are left behind. This is illustrated schematically in Fig. 2.19(a) where a two-dimensional Brillouin zone is shown and a circular Fermi surface of equal area is inscribed. Here we can easily see the electrons in the second zone at the zone edges and the holes at the zone corners that are left behind in the first zone. Translation by a reciprocal lattice vector brings two pieces of the electron surface together to form a surface in the shape of a lens, and the 4 pieces at the zone corners form a rosette shaped hole pocket. Typical examples of semimetals are bismuth and graphite. For these semimetals the carrier density is on the order of one carrier/10^6 atoms.

The carrier density of a semimetal is thus not very different from that which occurs in doped semiconductors, but the behavior of the conductivity $\sigma(T)$ as a function of temperature is very different. For intrinsic semiconductors, the carriers which are excited thermally contribute significantly to conduction. Consequently, the conductivity tends to rise rapidly with increasing temperature. For a semimetal, the carrier concentration does not change significantly with temperature because the carrier density is determined by the band overlap. Since the electron scattering by lattice vibrations increases with increasing temperature, the conductivity of semimetals tends to fall as the temperature increases.

A schematic diagram of the energy bands of the semimetal bismuth is shown in Fig. 2.19(b). Electron and hole carriers exist in equal numbers but at different locations in the Brillouin zone. For Bi, electrons are at the L–point, and holes at the T–point [see Fig. 2.19(b)]. The crystal structure for Bi can be understood from the NaCl structure by considering a very small displacement of the Na FCC structure relative to the Cl FCC structure along one of the body diagonals and an elongation of that body diagonal relative to the other 3 body diagonals. The special $\{111\}$ direction corresponds to $\Gamma - T$ in the Brillouin zone while the

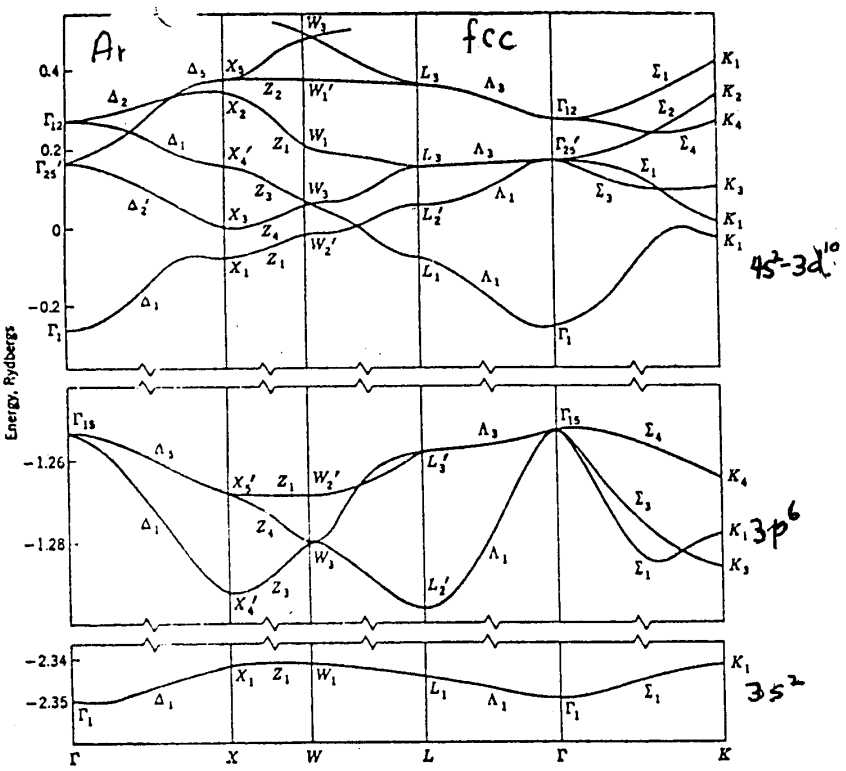


Figure 2.20: Electronic energy band structure of Argon.

other three {111} directions are labeled as $\Gamma - L$.

Instead of a band gap between valence and conduction bands (as occurs for semiconductors), semimetals are characterized by a band overlap in the millivolt range. In bismuth, a small band gap also occurs at the L–point between the conduction band and a lower filled valence band. Because the coupling between these L-point valence and conduction bands is strong, some of the effective mass components for the electrons in bismuth are anomalously small. As far as the optical properties of bismuth are concerned, bismuth behaves much like a metal with a high reflectivity at low frequencies due to strong free carrier absorption.

2.5 Insulators

The electronic structure of insulators is similar to that of semiconductors, in that both insulators and semiconductors have a band gap separating the valence and conduction bands. However, in the case of insulators, the band gap is so large that thermal energies are not sufficient to excite a significant number of carriers.

The simplest insulator is a solid formed of rare gas atoms. An example of a rare gas insulator is solid argon which crystallizes in the FCC structure with one Ar atom/primitive unit cell. With an atomic configuration $3s^2 3p^6$, argon has filled $3s$ and $3p$ bands which are easily identified in the energy band diagram in Fig. 2.20. These occupied bands have

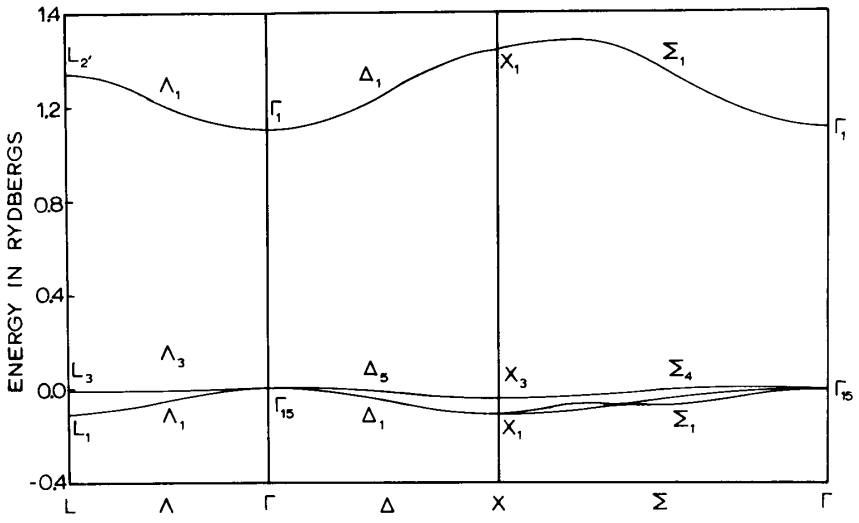


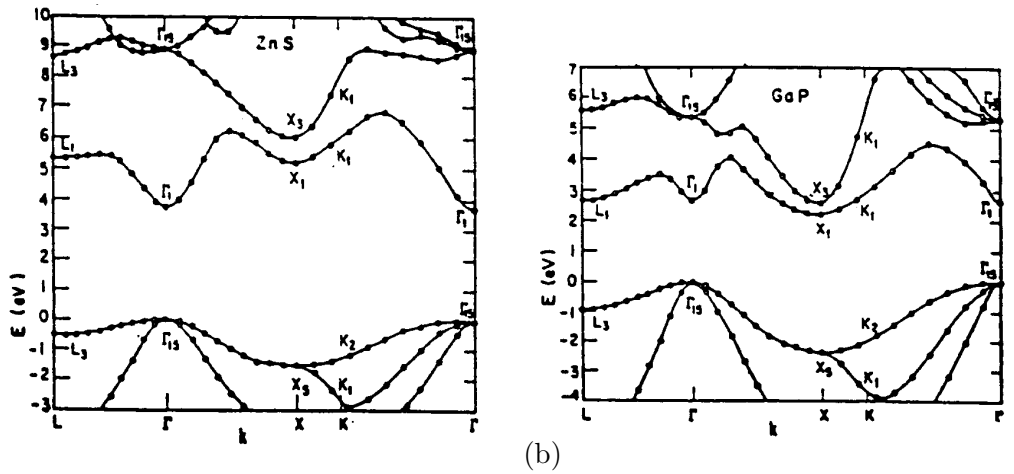
Figure 2.21: Band structure of the alkali halide insulator LiF. This ionic crystal is used extensively for UV optical components because of its large band gap.

very narrow band widths compared to their band gaps and are therefore well described by the tight binding approximation. This figure shows that the higher energy states forming the conduction bands (the hybridized $4s$ and $3d$ bands) show more dispersion than the more tightly bound valence bands. The band diagram shows argon to have a direct band gap at the Γ point of about 1 Rydberg or 13.6 eV. Although the $4s$ and $3d$ bands have similar energies, identification with the atomic levels can easily be made near $k = 0$ where the lower lying $4s$-band has considerably more band curvature than the $3d$ levels which are easily identified because of their degeneracies [the so called three-fold t_g ($\Gamma_{25'}$) and the two-fold e_g (Γ_{12}) crystal field levels for d-bands in a cubic crystal].

Another example of an insulator formed from a closed shell configuration is found in Fig. 2.21. Here the closed shell configuration results from charge transfer, as occurs in all ionic crystals. For example in the ionic crystal LiF (or in other alkali halide compounds) the valence band is identified with the filled anion orbitals (fluorine p–orbitals in this case) and at much higher energy the empty cation conduction band levels will lie (lithium s–orbitals in this case). Because of the wide band gap separation in the alkali halides between the valence and conduction bands, such materials are transparent at optical frequencies.

Insulating behavior can also occur for wide bandgap semiconductors with covalent bonding, such as diamond, ZnS and GaP (see Fig. 2.22). The $E(\vec{k})$ diagrams for these materials are very similar to the dispersion relations for typical III–V semiconducting compounds and the group IV semiconductors silicon and germanium; the main difference, however, is the large band gap separating valence and conduction bands.

Even in insulators there is a finite electrical conductivity. For these materials the band electronic transport processes become less important relative to charge hopping from one atom to another by over-coming a potential barrier. Ionic conduction can also occur in insulating ionic crystals. From a practical point of view, one of the most important applications



(a) (b)

Figure 2.22: Electronic energy band structure of (a) cubic ZnS, a direct gap semi–insulating II–VI semiconductor, and (b) cubic GaP, an indirect gap semi–insulating III–V semiconductor.

of insulators is the control of electrical breakdown phenomena.

The principal experimental methods for studying the electronic energy bands depend on the nature of the solid. For insulators, the optical properties are the most important, while for semiconductors both optical and transport studies are important. For metals, optical properties are less important and Fermi surface studies become more important.

In the case of insulators, electrical conductivity can arise through the motion of lattice ions as they move from one lattice vacancy to another, or from one interstitial site to another. Ionic conduction therefore occurs through the presence of lattice defects, and is promoted in materials with open crystal structures. In ionic crystals there are relatively few mobile electrons or holes even at high temperature so that conduction in these materials is predominantly due to the motions of ions.

Ionic conductivity (σ_{ionic}) is proportional both to the density of lattice defects (vacancies and interstitials) and to the diffusion rate, so that we can write

$$\sigma_{\text{ionic}} \sim e^{-(E+E_0)/k_B T} \tag{2.7}$$

where E_0 is the activation energy for ionic motion and E is the energy for formation of a defect (a vacancy, a vacancy pair, or an interstitial). Being an activated process, ionic conduction is enhanced at elevated temperatures. Since defects in ionic crystals can be observed visibly as the migration of color through the crystal, ionic conductivity can be distinguished from electronic conductivity by comparing the transport of charge with the transport of mass, as can, for example, be measured by the material plated out on electrodes in contact with the ionic crystal.

In this course, we will spend a good deal of time studying optical and transport properties of solids. In connection with topics on magnetism, we will also study Fermi surface measurements which are closely connected to issues relevant to transport properties. Since Fermi surface studies, for the most part, are resonance experiments and involve the use of a magnetic field, it is pedagogically more convenient to discuss these topics in Part III of

the course, devoted to Magnetism. We have presented this review of the electronic energy bands of solids because the $E(\vec{k})$ relations are closely connected with a large number of common measurements in the laboratory, and because a knowledge of the $E(\vec{k})$ relations forms the basis for many device applications.

Chapter 3

Effective Mass Theory

Reference

- Smith, Janak and Adler, *Electron Conduction in Solids*, McGraw-Hill, 1967, Chapter 6.

3.1 Wavepackets in Crystals and Group Velocity of Electrons in Solids

In a crystal lattice, the electronic motion which is induced by an applied field is conveniently described by a wavepacket composed of eigenstates of the unperturbed crystal. These eigenstates are Bloch functions

$$\psi_{nk}(\vec{r}) = e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) \tag{3.1}$$

and are associated with band n. These wavepackets are solutions of the time-dependent Schrödinger equation

$$\mathcal{H}_0 \psi_n(\vec{r}, t) = i\hbar \frac{\partial \psi_n(\vec{r}, t)}{\partial t} \tag{3.2}$$

where the time independent part of the Hamiltonian can be written as

$$\mathcal{H}_0 = \frac{p^2}{2m} + V(\vec{r}), \tag{3.3}$$

where $V(\vec{r}) = V(\vec{r} + \vec{R}_n)$ is the periodic potential. The wave packets $\psi_n(\vec{r}, t)$ can be written in terms of the Bloch states $\psi_{nk}(\vec{r})$ as

$$\psi_n(\vec{r}, t) = \sum_k A_{n,k}(t) \psi_{nk}(\vec{r}) = \int d^3k A_{n,k}(t) \psi_{nk}(\vec{r}) \tag{3.4}$$

where we have replaced the sum by an integration over the Brillouin zone, since permissible $\vec{k}$ values for a macroscopic solid are *very* closely spaced. If the Hamiltonian $\mathcal{H}_0$ is time-independent as is often the case, we can write

$$A_{n,k}(t) = A_{n,k} e^{-i\omega_n(\vec{k})t} \tag{3.5}$$

where

$$\hbar\omega_n(\vec{k}) = E_n(\vec{k}) \tag{3.6}$$

and thereby obtain

$$\psi_n(\vec{r},t) = \int d^3k A_{n,k} u_{nk}(\vec{r}) e^{i[\vec{k}\cdot\vec{r}-\omega_n(\vec{k})t]}. \tag{3.7}$$

We can localize the wavepacket in $\vec{k}$-space by requiring that the coefficients $A_{n,k}$ be large only in a confined region of $\vec{k}$-space centered at $\vec{k} = \vec{k}_0$. If we now expand the band energy in a Taylor series around $\vec{k} = \vec{k}_0$ we obtain:

$$E_n(\vec{k}) = E_n(\vec{k}_0) + (\vec{k} - \vec{k}_0) \cdot \left.\frac{\partial E_n(\vec{k})}{\partial \vec{k}}\right|_{\vec{k}=\vec{k}_0} + ..., \tag{3.8}$$

where we have written $\vec{k}$ as

$$\vec{k} = \vec{k}_0 + (\vec{k} - \vec{k}_0). \tag{3.9}$$

Since $|\vec{k} - \vec{k}_0|$ is assumed to be small compared with Brillouin zone dimensions, we are justified in retaining only the first two terms of the Taylor expansion in Eq. 3.8 given above. Substitution into Eq. 3.4 for the wave packet yields:

$$\psi_n(\vec{r},t) \simeq e^{i(\vec{k}_0\cdot\vec{r}-\omega_n(\vec{k}_0)t)} \int d^3k A_{n,k}\ u_{nk}(\vec{r}) e^{i(\vec{k}-\vec{k}_0)\cdot[\vec{r}-\frac{\partial\omega_n(\vec{k})}{\partial\vec{k}}t]} \tag{3.10}$$

where

$$\hbar\omega_n(\vec{k}_0) = E_n(\vec{k}_0) \tag{3.11}$$

and

$$\hbar\frac{\partial\omega_n(\vec{k})}{\partial\vec{k}} = \frac{\partial E_n(\vec{k})}{\partial\vec{k}} \tag{3.12}$$

and the derivative $\partial\omega_n(\vec{k})/\partial\vec{k}$ which appears in the phase factor of Eq. 3.10 is evaluated at $\vec{k} = \vec{k}_0$. Except for the periodic function $u_{nk}(\vec{r})$ the above expression is in the standard form for a wavepacket moving with *"group velocity"* $\vec{v}_g$

$$\vec{v}_g \equiv \frac{\partial\omega_n(\vec{k})}{\partial\vec{k}} \tag{3.13}$$

so that

$$\vec{v}_g = \frac{1}{\hbar}\frac{\partial E_n(\vec{k})}{\partial\vec{k}}, \tag{3.14}$$

while the phase velocity

$$\frac{\vec{v}_p = \omega_n(\vec{k})}{\vec{k} = \frac{\partial E_n(\vec{k})}{\vec{k}}}. \tag{3.15}$$

In the limit of free electrons the group velocity becomes

$$\vec{v}_g = \frac{\vec{p}}{m} = \frac{\hbar\vec{k}}{m}. \tag{3.16}$$

This result also follows from the above discussion using

$$E_n(\vec{k}) = \frac{\hbar^2 k^2}{2m} \tag{3.17}$$

$$\frac{\partial E_n(\vec{k})}{\hbar \partial \vec{k}} = \frac{\hbar \vec{k}}{m}. \tag{3.18}$$

We shall show later that the electron wavepacket moves through the crystal very much like a free electron provided that the wavepacket remains localized in k space during the time interval of interest in the particular problem under consideration. Because of the uncertainty principle, the localization of a wavepacket in reciprocal space implies a delocalization of the wavepacket in real space.

We use wavepackets to describe electronic states in a solid when the crystal is perturbed in some way (e.g., by an applied electric or magnetic field). We make frequent applications of wavepackets to transport theory (e.g., electrical conductivity). In many practical applications of transport theory, use is made of the Effective-Mass Theorem, which is the most important result of transport theory.

We note that the above discussion for the wavepacket is given in terms of the **perfect crystal**. In our discussion of the Effective-Mass Theorem we will see that these wavepackets are also of use in describing situations where the Hamiltonian which enters Schrödinger's equation contains both the unperturbed Hamiltonian of the perfect crystal $\mathcal{H}_0$ and the perturbation Hamiltonian $\mathcal{H}'$ arising from an external perturbation. Common perturbations are applied electric or magnetic fields, or a lattice defect or an impurity atom.

3.2 The Effective Mass Theorem

We shall now present the Effective Mass theorem, which is central to the consideration of the electrical and optical properties of solids. An elementary proof of the theorem will be given here for a simple but important case, namely the non–degenerate band which can be identified with the corresponding atomic state. The theorem will be discussed from a more advanced point of view which considers also the case of degenerate bands in the following courses in the physics of solids sequence.

For many practical situations we find a solid in the presence of some perturbing field (e.g., an externally applied electric field, or the perturbation created by an impurity atom or a crystal defect). The perturbation may be either time–dependent or time–independent. We will show here that under many common circumstances this perturbation can be treated in the effective mass approximation whereby the periodic potential is replaced by an effective Hamiltonian based on the $E(\vec{k})$ relations for the perfect crystal.

To derive the effective mass theorem, we start with the time–dependent Schrödinger equation

$$(\mathcal{H}_0 + \mathcal{H}')\psi_n(\vec{r}, t) = i\hbar \frac{\partial \psi_n(\vec{r}, t)}{\partial t}. \tag{3.19}$$

We then substitute the expansion for the wave packet

$$\psi_n(\vec{r}, t) = \int d^3k A_{nk}(t) e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) \tag{3.20}$$

into Schrödinger's equation and make use of the Bloch solution

$$\mathcal{H}_0 e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) = E_n(\vec{k}) e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) \tag{3.21}$$

to obtain:

$$\begin{aligned}(\mathcal{H}_0 + \mathcal{H}')\psi_n(\vec{r},t) &= \int d^3k [E_n(\vec{k}) + \mathcal{H}'] A_{nk}(t) e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) = i\hbar(\partial\psi_n(\vec{r},t)/\partial t) \\ &= i\hbar \int d^3k \dot{A}_{nk}(t) e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}).\end{aligned} \tag{3.22}$$

It follows from Bloch's theorem that $E_n(\vec{k})$ is a periodic function in the reciprocal lattice. We can therefore expand $E_n(\vec{k})$ in a Fourier series in the direct lattice

$$E_n(\vec{k}) = \sum_{\vec{R}_\ell} E_{n\ell}\, e^{i\vec{k}\cdot\vec{R}_\ell} \tag{3.23}$$

where the $\vec{R}_\ell$ are lattice vectors. Now consider the differential operator $E_n(-i\vec{\nabla})$ formed by replacing $\vec{k}$ by $-i\vec{\nabla}$

$$E_n(-i\vec{\nabla}) = \sum_{\vec{R}_\ell} E_{n\ell}\, e^{\vec{R}_\ell\cdot\vec{\nabla}}. \tag{3.24}$$

Consider the effect of $E_n(-i\vec{\nabla})$ on an arbitrary function $f(\vec{r})$. Since $e^{\vec{R}_\ell\cdot\vec{\nabla}}$ can be expanded in a Taylor series, we obtain

$$\begin{aligned}e^{\vec{R}_\ell\cdot\vec{\nabla}} f(\vec{r}) &= [1 + \vec{R}_\ell\cdot\vec{\nabla} + \tfrac{1}{2}(\vec{R}_\ell\cdot\vec{\nabla})(\vec{R}_\ell\cdot\vec{\nabla}) + ...] f(\vec{r}) \\ &= f(\vec{r}) + \vec{R}_\ell\cdot\vec{\nabla} f(\vec{r}) + \tfrac{1}{2!} R_{\ell,\alpha} R_{\ell,\beta} \frac{\partial^2}{\partial r_\alpha \partial r_\beta} f(\vec{r}) + ... \\ &= f(\vec{r} + \vec{R}_\ell).\end{aligned} \tag{3.25}$$

Thus the effect of $E_n(-i\vec{\nabla})$ on a Bloch state is

$$E_n(-i\vec{\nabla})\psi_{nk}(\vec{r}) = \sum_{\vec{R}_\ell} E_{n\ell}\psi_{nk}(\vec{r} + \vec{R}_\ell) = \sum_{\vec{R}_\ell} E_{n\ell} e^{i\vec{k}\cdot\vec{R}_\ell} e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) = E_n(\vec{k})\psi_{nk}(\vec{r}) \tag{3.26}$$

since from Bloch's theorem

$$\psi_{nk}(\vec{r} + \vec{R}_\ell) = e^{i\vec{k}\cdot\vec{R}_\ell}\Big[e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r})\Big]. \tag{3.27}$$

Substitution of

$$E_n(-i\vec{\nabla})\psi_{nk}(\vec{r}) = E_n(\vec{k})\psi_{nk}(\vec{r}) \tag{3.28}$$

from Eq. 3.26 into Schrödinger's equation (Eq. 3.22) yields:

$$\int d^3k \,\Big[E_n(-i\vec{\nabla}) + \mathcal{H}'\Big] A_{nk}(t) e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) = \Big[E_n(-i\vec{\nabla}) + \mathcal{H}'\Big] \int d^3k \; A_{nk}(t) e^{i\vec{k}\cdot\vec{r}} u_{nk}(\vec{r}) \tag{3.29}$$

so that

$$\Big[E_n(-i\vec{\nabla}) + \mathcal{H}'\Big]\psi_n(\vec{r},t) = i\hbar\frac{\partial\psi_n(\vec{r},t)}{\partial t}. \tag{3.30}$$

This result is called the *effective mass theorem*. We observe that the original crystal Hamiltonian $p^2/2m + V(\vec{r})$ does not appear in this equation. It has instead been replaced by an effective Hamiltonian which is an operator formed from the solution $E(\vec{k})$ for the perfect crystal in which we replace $\vec{k}$ by $-i\vec{\nabla}$. For example, for the free electron ($V(\vec{r}) \equiv 0$)

$$E_n(-i\vec{\nabla}) \rightarrow -\frac{\hbar^2\nabla^2}{2m}. \tag{3.31}$$

In applying the effective mass theorem, we assume that $E(\vec{k})$ is known either from the results of a theoretical calculation or from the analysis of experimental results. What is important here is that once $E(\vec{k})$ is known, the effect of various perturbations on the ideal crystal can be treated in terms of the solution to the energy levels of the perfect crystal without recourse to consideration of the full Hamiltonian. In practical cases, the solution to the effective mass equation is much easier to carry out than the solution to the original Schrödinger equation.

According to the above discussion, we have assumed that $E(\vec{k})$ is specified throughout the Brillouin zone. For many practical applications, the region of $\vec{k}$-space which is of importance is confined to a small portion of the Brillouin zone. In such cases it is only necessary to specify $E(\vec{k})$ in a local region (or regions) and to localize our wavepacket solutions to these local regions of $\vec{k}$-space. Suppose that we localize the wavepacket around $\vec{k} = \vec{k}_0$, and correspondingly expand our Bloch functions around $\vec{k}_0$,

$$\psi_{nk}(\vec{r}) = e^{i\vec{k}\cdot\vec{r}}u_{nk}(\vec{r}) \simeq e^{i\vec{k}\cdot\vec{r}}u_{nk_0}(\vec{r}) = e^{i(\vec{k}-\vec{k}_0)\cdot\vec{r}}\psi_{nk_0}(\vec{r}) \tag{3.32}$$

where we have noted that $u_{nk}(\vec{r}) \simeq u_{nk_0}(\vec{r})$ has only a weak dependence on $\vec{k}$. Then our wavepacket can be written as

$$\psi_{nk}(\vec{r},t) = \int d^3k\ A_{nk}(t)e^{i(\vec{k}-\vec{k}_0)\cdot\vec{r}}\psi_{nk_0}(\vec{r}) = F(\vec{r},t)\psi_{nk_0}(\vec{r}) \tag{3.33}$$

where $F(\vec{r},t)$ is called the *amplitude* or *envelope* function and is defined by

$$F(\vec{r},t) = \int d^3k\ A_{nk}(t)e^{i(\vec{k}-\vec{k}_0)\cdot\vec{r}}. \tag{3.34}$$

Since the time dependent Fourier coefficients $A_{nk}(t)$ are assumed here to be large only near $\vec{k} = \vec{k}_0$, then $F(\vec{r},t)$ will be a slowly varying function of $\vec{r}$, because in this case

$$e^{i(\vec{k}-\vec{k}_0)\vec{r}} \simeq 1 + i(\vec{k}-\vec{k}_0)\cdot\vec{r} + ... \tag{3.35}$$

It can be shown that the envelope function also satisfies the effective mass equation

$$\left[E_n(-i\vec{\nabla}) + \mathcal{H}'\right]F(\vec{r},t) = i\hbar\frac{\partial F(\vec{r},t)}{\partial t} \tag{3.36}$$

where we now replace $\vec{k} - \vec{k}_0$ in $E_n(\vec{k})$ by $-i\vec{\nabla}$. This form of the effective mass equation is useful for treating the problem of donor and acceptor impurity states in semiconductors, and $\vec{k}_0$ is taken as the band extremum.

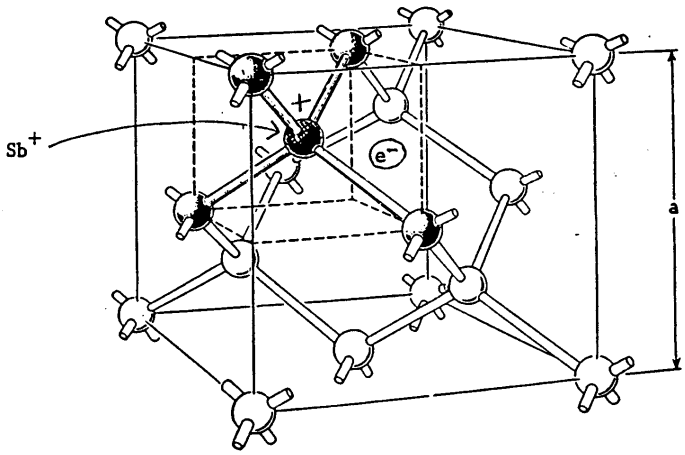


Figure 3.1: Crystal structure of diamond, showing the tetrahedral bond arrangement with an Sb^+ ion on one of the lattice sites and a free donor electron available for conduction.

3.3 Application of the Effective Mass Theorem to Donor Impurity Levels in a Semiconductor

Suppose that we add an impurity from column V in the Periodic Table to a semiconductor such as silicon or germanium, which are both members of column IV of the periodic table. This impurity atom will have one more electron than is needed to satisfy the valency requirements for the tetrahedral bonds which the germanium or silicon atoms form with their 4 valence electrons (see Fig. 3.1).

This extra electron from the impurity atom will be free to wander through the lattice, subject of course to the coulomb attraction of the ion core which will have one unit of positive charge. We will consider here the case where we add just a small number of these impurity atoms so that we may focus our attention on a single, isolated substitutional impurity atom in an otherwise perfect lattice. In the course of this discussion we will define more carefully what the limits on the impurity concentration must be so that the treatment given here is applicable.

Let us also assume that the conduction band of the host semiconductor in the vicinity of the band "minimum" at $\vec{k}_0$ has the simple analytic form

$$E_c(\vec{k}) \simeq E_c(\vec{k}_0) + \frac{\hbar^2(k-k_0)^2}{2m^*}. \tag{3.37}$$

We can consider this expression for the conduction band level $E_c(\vec{k})$ as a special case of the Taylor expansion of $E(\vec{k})$ about an energy band minimum at $\vec{k} = \vec{k}_0$. For the present discussion, $E(\vec{k})$ is assumed to be isotropic in $\vec{k}$; this typically occurs in cubic semiconductors with band extrema at $\vec{k} = 0$. The quantity m^* in this equation is the *effective mass* for the electrons. We will see that the energy levels corresponding to the donor electron will lie in the band gap below the conduction band minimum as indicated in the diagram in Fig. 3.2. To solve for the impurity levels explicitly, we may use the time-independent form of the

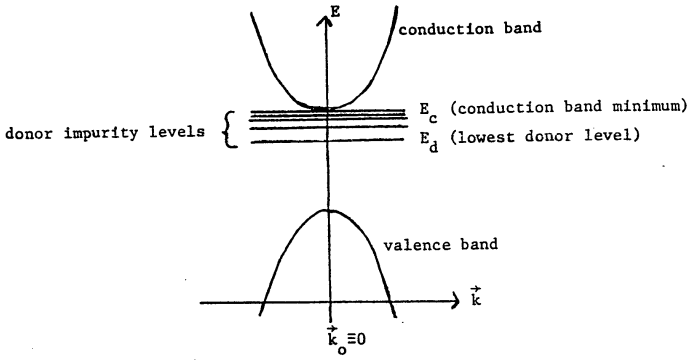


Figure 3.2: Schematic band diagram showing donor levels in a semiconductor.

effective mass theorem derived from Eq. 3.36

$$\left[E_n(-i\vec{\nabla}) + \mathcal{H}'\right] F(\vec{r}) = (E - E_c) F(\vec{r}). \tag{3.38}$$

Equation 3.38 is applicable to the impurity problem in a semiconductor provided that the amplitude function $F(\vec{r})$ is sufficiently slowly varying over a unit cell. In the course of this discussion, we will see that the donor electron in a column IV (or III-V or II-VI compound semiconductor) will wander over many lattice sites and therefore this approximation on $F(\vec{r})$ will be justified.

For a singly ionized donor impurity (such as arsenic in germanium), the perturbing potential $\mathcal{H}'$ can be represented as a Coulomb potential

$$\mathcal{H}' = -\frac{e^2}{\varepsilon r} \tag{3.39}$$

where ε is an average dielectric constant of the crystal medium which the donor electron sees as it wanders through the crystal. Experimental data on donor impurity states indicate that ε is very closely equal to the low frequency limit of the electronic dielectric constant $\varepsilon_1(\omega)|_{\omega=0}$, which we will discuss extensively in treating the optical properties of solids (Part II of this course). The above discussion involving an isotropic $E(\vec{k})$ is appropriate for semiconductors with conduction band minima at $\vec{k}_0 = 0$. The Effective Mass equation for the unperturbed crystal is

$$E_n(-i\vec{\nabla}) = -\frac{\hbar^2}{2m^*}\nabla^2 \tag{3.40}$$

in which we have replaced $\vec{k}$ by $-i\vec{\nabla}$.

The donor impurity problem in the effective mass approximation thus becomes

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2 - \frac{e^2}{\varepsilon r}\right] F(\vec{r}) = (E - E_c) F(\vec{r}) \tag{3.41}$$

where all energies are measured with respect to the bottom of the conduction band E_c. If we replace m^* by m and e^2/ε by e^2, we immediately recognize this equation as Schrödinger's

equation for a hydrogen atom under the identification of the energy eigenvalues with

$$E_n = \frac{e^2}{2n^2 a_0} = \frac{me^4}{2n^2\hbar^2} \tag{3.42}$$

where a_0 is the Bohr radius $a_0 = \hbar^2/me^2$. This identification immediately allows us to write E_ℓ for the donor energy levels as

$$E_\ell = E_c - \frac{m^* e^4}{2\varepsilon^2 \ell^2 \hbar^2} \tag{3.43}$$

where $\ell = 1, 2, 3, \ldots$ is an integer denoting the donor level quantum numbers and we identify the bottom of the conduction band E_c as the ionization energy for this effective hydrogenic problem. Physically, this means that the donor levels correspond to bound (localized) states while the band states above E_c correspond to delocalized nearly–free electron–like states. The lowest or "ground-state" donor energy level is then written as

$$E_d = E_{\ell=1} = E_c - \frac{m^* e^4}{2\varepsilon^2\hbar^2}. \tag{3.44}$$

It is convenient to identify the "effective" first Bohr radius for the donor level as

$$a_0^* = \frac{\varepsilon\hbar^2}{m^* e^2} \tag{3.45}$$

and to recognize that the wave function for the ground state donor level will be of the form

$$F(\vec{r}) = \mathcal{C} e^{-r/a_0^*} \tag{3.46}$$

where $\mathcal{C}$ is the normalization constant. Thus the solutions to Eq. 3.41 for a semiconductor are hydrogenic energy levels with the substitutions $m \to m^*$, $e^2 \to (e^2/\varepsilon)$ and the ionization energy usually taken as the zero of energy for the hydrogen atom now becomes E_c, the conduction band extremum.

For a semiconductor like germanium we have a very large dielectric constant, $\varepsilon \simeq 16$. The value for the effective mass is somewhat more difficult to specify in germanium since the constant energy surfaces for germanium are located about the L–points in the Brillouin zone (see §2.3.2) and are ellipsoids of revolution. Since the constant energy surfaces for such semiconductors are non-spherical, the effective mass tensor is anisotropic. However we will write down an average effective mass value $m^*/m \simeq 0.12$ (Kittel ISSP) so that we can estimate pertinent magnitudes for the donor levels in a typical semiconductor. With these values for ε and m^* we obtain:

$$E_c - E_d \simeq 0.007\text{eV} \tag{3.47}$$

and the effective Bohr radius

$$a_0^* \simeq 70\text{Å}. \tag{3.48}$$

These values are to be compared with the ionization energy of 13.6 eV for the hydrogen atom and with the hydrogenic Bohr orbit of $a_0 = \hbar^2/me^2 = 0.5$Å.

Thus we see that a_0^* is indeed large enough to satisfy the requirement that $F(\vec{r})$ be slowly varying over a unit cell. On the other hand, if a_0^* were to be comparable to a lattice dimension, then $F(\vec{r})$ could not be considered as a slowly varying function of $\vec{r}$

and generalizations of the above treatment would have to be made. Such generalizations involve: (1) treating $E(\vec{k})$ for a wider region of $\vec{k}$-space, and (2) relaxing the condition that impurity levels are to be associated with a single band. From the uncertainty principle, the localization in momentum space for the impurity state requires a delocalization in real space; and likewise, the converse is true, that a localized impurity in real space corresponds to a delocalized description in $\vec{k}$-space. Thus "shallow" hydrogenic donor levels can be attributed to a specific band at a specific energy extremum at $\vec{k}_0$ in the Brillouin zone. On the other hand, "deep" donor levels are not hydrogenic and have a more complicated energy level structure. Deep donor levels cannot be readily associated with a specific band or a specific $\vec{k}$ point in the Brillouin zone.

In dealing with this impurity problem, it is very tempting to discuss the donor levels in silicon and germanium. For example in silicon where the conduction band extrema are at the Δ point (see §2.3.3), the effective mass theorem requires us to replace $E(-i\vec{\nabla})$ by

$$E_n(-i\vec{\nabla}) \rightarrow -\frac{\hbar^2}{2m_\ell^*}\frac{\partial^2}{\partial x^2} - \frac{\hbar^2}{2m_t^*}\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \tag{3.49}$$

and the resulting Schrödinger equation can no longer be solved exactly. Although this is a very interesting problem from a practical point of view, it is too difficult a problem for us to solve in detail for illustrative purposes in an introductory course.

3.4 Quasi-Classical Electron Dynamics

According to the "Correspondence Principle" of Quantum Mechanics, wavepacket solutions of Schrödinger's equation follow the trajectories of classical particles and satisfy Newton's laws. One can give a Correspondence Principle argument for the form which is assumed by the velocity and acceleration of a wavepacket. According to the Correspondence Principle, the connection between the classical Hamiltonian and the quantum mechanical Hamiltonian is made by the identification of $\vec{p} \rightarrow (\hbar/i)\vec{\nabla}$. Thus

$$E_n(-i\vec{\nabla}) + \mathcal{H}'(\vec{r}) \leftrightarrow E_n(\vec{p}/\hbar) + \mathcal{H}'(\vec{r}) = \mathcal{H}_{\text{classical}}(\vec{p},\vec{r}). \tag{3.50}$$

In classical mechanics, Hamilton's equations give the velocity according to :

$$\dot{\vec{r}} = \frac{\partial \mathcal{H}}{\partial \vec{p}} = \nabla_p \mathcal{H} = \frac{\partial E(\vec{k})}{\hbar \partial \vec{k}} \tag{3.51}$$

in agreement with the group velocity for a wavepacket given by Eq. 3.13. Hamilton's equation for the acceleration is given by:

$$\dot{\vec{p}} = -\frac{\partial \mathcal{H}}{\partial \vec{r}} = -\frac{\partial \mathcal{H}'(\vec{r})}{\partial \vec{r}}. \tag{3.52}$$

For example, in the case of an applied electric field $\vec{E}$ the perturbation Hamiltonian is

$$\mathcal{H}'(\vec{r}) = -e\vec{r}\cdot\vec{E} \tag{3.53}$$

so that

$$\dot{\vec{p}} = \hbar\dot{\vec{k}} = e\vec{E}. \tag{3.54}$$

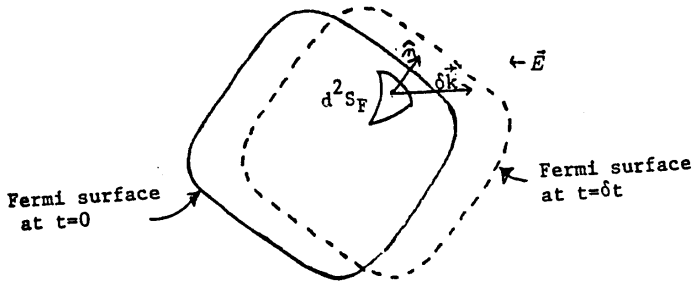


Figure 3.3: Displaced Fermi surface under the action of an electric field $\vec{E}$.

In this equation $e\vec{E}$ is the classical Coulomb force on an electric charge due to an applied field $\vec{E}$. It can be shown (to be derived rigorously in the advanced course) that in the presence of a magnetic field $\vec{B}$, the acceleration theorem follows the Lorentz force equation

$$\dot{\vec{p}} = \hbar\dot{\vec{k}} = e[\vec{E} + (1/c)\vec{v} \times \vec{B}] \tag{3.55}$$

where

$$\vec{v} = \frac{\partial E(\vec{k})}{\hbar \partial \vec{k}}. \tag{3.56}$$

In the crystal, the crystal momentum $\hbar\vec{k}$ for the wavepacket plays the role of the momentum for a classical particle.

3.5 Quasi-Classical Theory of Electrical Conductivity – Ohm's Law

We will now apply the idea of the quasi-classical electron dynamics in a solid to the problem of the electrical conductivity for a metal with an arbitrary Fermi surface and band structure. The electron is treated here as a wavepacket with momentum $\hbar\vec{k}$ moving in an external electric field $\vec{E}$ in compliance with Newton's laws. Because of the acceleration theorem, we can think of the electric field as creating a "displacement" of the electron distribution in $\vec{k}$-space. We remember that the Fermi surface encloses the region of occupied states within the Brillouin zone. The effect of the electric field is to change the wave vector $\vec{k}$ of an electron by

$$\delta\vec{k} = \frac{e}{\hbar}\vec{E}\delta t \tag{3.57}$$

(where we note that the charge on the electron e is a negative number). We picture the displacement $\delta\vec{k}$ of Eq. 3.57 by the displacement of the Fermi surface shown in Fig. 3.3. From this diagram we see that the incremental volume of $\vec{k}$-space $\delta^3 V_{\vec{k}}$ which is "swept out" in the time δt due to the presence of the field $\vec{E}$ is

$$\delta^3 V_{\vec{k}} = d^2 S_F \; \hat{n} \cdot \delta\vec{k} = d^2 S_F \; \hat{n} \cdot (\frac{e}{\hbar}\vec{E}\delta t) \tag{3.58}$$

and the electron density is found from

$$n = \frac{2}{(2\pi)^3} \int_{E \leq E_F} d^3k \tag{3.59}$$

where d^2S_F is the element of area on the Fermi surface and $\hat{n}$ is a unit vector normal to this element of area and $\delta^3 V_{\vec{k}} \rightarrow d^3k$ are both elements of volume in $\vec{k}$–space. The definition of the electrical current density is the current flowing through a unit area in real space and is given by the product of the (number of electrons per unit volume) with the (charge per electron) and with the (group velocity) so that the current density $\delta\vec{j}$ created by applying the electric field $\vec{E}$ for a time interval δt is given by

$$\delta\vec{j} = \int [2/(2\pi)^3] \ \cdot \ [\delta^3 V_{\vec{k}}] \cdot [e] \cdot [\vec{v}_g] \tag{3.60}$$

where $\vec{v}_g$ is the group velocity for electron wavepacket and $2/(2\pi)^3$ is the density of electronic states in $\vec{k}$-space (including the spin degeneracy of two) because we can put 2 electrons in each phase space state. Substitution for $\delta^3 V_{\vec{k}}$ in Eq. 3.60 by Eq. 3.58 yields the instantaneous rate of change of the current density averaged over the Fermi surface

$$\frac{\partial\vec{j}}{\partial t} = \frac{e^2}{4\pi^3\hbar} \oint\oint \vec{v}_g \ \hat{n} \cdot \vec{E} \ d^2S_F = \frac{e^2}{4\pi^3\hbar} \oint\oint \vec{v}_g \left(\frac{\vec{v}_g \cdot \vec{E}}{|v_g|} \right) (d^2S_F) \tag{3.61}$$

since the group velocity given by Eq. 3.13 is directed normal to the Fermi surface. In a real solid, the electrons will not be accelerated indefinitely, but will eventually collide with an impurity, or a lattice defect or a lattice vibration (phonon).

These collisions will serve to maintain the displacement of the Fermi surface at some steady state value, depending on τ, the average time between collisions. We can introduce this relaxation time through the expression

$$n(t) = n(0)e^{-t/\tau} \tag{3.62}$$

where $n(t)$ is the number of electrons that have not made a collision at time t, assuming that the last collision had been made at time $t = 0$. The relaxation time is the average collision time

$$\langle t \rangle = \frac{1}{\tau} \int_0^\infty t e^{-t/\tau} dt = \tau. \tag{3.63}$$

If in Eq. 3.61, we set $\langle \delta t \rangle = \tau$ and write the average current density as $\vec{j} = \langle \delta\vec{j} \rangle$, then we obtain

$$\vec{j} = \frac{e^2\tau}{4\pi^3\hbar} \int \vec{v}_g \frac{\vec{v}_g \cdot \vec{E}}{|v_g|} \ (d^2S_F). \tag{3.64}$$

We define the conductivity tensor as $\vec{j} = \overleftrightarrow{\sigma} \cdot \vec{E}$, so that Eq. 3.64 provides an explicit expression for the tensor $\overleftrightarrow{\sigma}$.

$$\overleftrightarrow{\sigma} = \frac{e^2\tau}{4\pi^3\hbar} \int \frac{\vec{v}_g \vec{v}_g}{|v_g|} \ (d^2S_F). \tag{3.65}$$

In the free electron limit $\overleftrightarrow{\sigma}$ becomes a scalar (isotropic conduction) and is given by the Drude formula which we derive below from Eq. 3.65. Using the equations for the free

electron limit

$$E=\hbar^2k^2/2m$$

$$E_F=\hbar^2k_F^2/2m \tag{3.66}$$

$$\vec{v}_g=\hbar\vec{k}_F/m.$$

We then obtain

$$\vec{v}_g\vec{v}_g \rightarrow v_x^2 = v_y^2 = v_z^2 = v^2/3 \tag{3.67}$$

$$\int d^2S_F = 4\pi k_F^2, \tag{3.68}$$

so that the number of electrons/unit volume can be written as:

$$n = \frac{1}{4\pi^3}\frac{4\pi}{3}\,k_F^3. \tag{3.69}$$

Therefore

$$\vec{j} = \frac{e^2\tau}{4\pi^3\hbar}\left(\frac{\hbar k_F}{m}\right)\frac{1}{3}\,\vec{E}(4\pi k_F^2) = \frac{ne^2\tau}{m}\,\vec{E}. \tag{3.70}$$

Thus the free electron limit gives Ohm's law in the familiar form

$$\sigma = \frac{ne^2\tau}{m} = ne\mu, \tag{3.71}$$

and showing that the electrical conductivity depends on both the carrier density n and the carrier mobility μ.

A slightly modified form of Ohm's law is also applicable to conduction in a material for which the energy dispersion relations are simple parabolic and m has been replaced by the effective mass m^*, $E(\vec{k}) = \hbar^2k^2/2m^*$. In this case σ is given by

$$\sigma = ne^2\tau/m^* \tag{3.72}$$

where the effective mass is found from the band curvature $1/m^* = \partial^2E/\hbar^2\partial k^2$. The generalization of Ohm's law can also be made to deal with solids for which the effective mass tensor is *anisotropic* and this will be discussed later in this course.

Chapter 4

Transport Phenomena

References:

- Ziman, *Principles of the Theory of Solids*, Cambridge Univ. Press, 1972, Chapters 7 and 9.
- Ashcroft and Mermin, *Solid State Physics*, Holt, Rinehart and Winston, 1976, Chapters 13.
- Smith, Janak and Adler, *Electronic Conduction in Solids*, McGraw-Hill, 1967, Chapters 7, 8, and 9.

4.1 Introduction

In this section we study some of the transport properties for metals and semiconductors. An intrinsic semiconductor at $T = 0$ has no carriers and therefore there is no transport of carriers under the influence of external fields. However at finite temperatures there are thermally generated carriers. Impurities also can serve to generate carriers and transport properties. For insulators, there is very little charge transport and in this case the defects and the ions themselves can participate in charge transport under the influence of external applied fields. Metals make use of the Fermi-Dirac distribution function but are otherwise similar to semiconductors, for which the Maxwell–Boltzmann distribution function is usually applicable.

At finite fields, the electrical conductivity will depend on the product of the carrier density and the carrier mobility. For a one carrier system, the Hall effect gives the carrier density and the magnetoresistance gives the mobility, the key parameters governing the transport properties of a semiconductor. From the standpoint of device applications, the carrier density and the carrier mobility are the parameters of greatest importance.

To the extent that electrons can be considered as particles, the electrical conductivity, the electronic contribution to the thermal conductivity and the magnetoresistance are all found by solving the Boltzmann equation. For the case of ultra-small dimensions, where the wave aspects of the electron must be considered (called mesoscopic physics), more sophisticated approaches to the transport properties must be considered. To review the

standard procedures for classical electrons, we briefly review the Boltzmann equation and its solution in the next section.

4.2 The Boltzmann Equation

The Boltzmann transport equation is a statement that in the steady state, there is no net change in the distribution function $f(\vec{r},\vec{k},t)$ which determines the probability of finding an electron at position $\vec{r}$, crystal momentum $\vec{k}$ and time t. Therefore we get a zero sum for the changes in $f(\vec{r},\vec{k},t)$ due to the 3 processes of diffusion, the effect of forces and fields, and collisions:

$$\left.\frac{\partial f(\vec{r},\vec{k},t)}{\partial t}\right|_{\text{diffusion}} + \left.\frac{\partial f(\vec{r},\vec{k},t)}{\partial t}\right|_{\text{fields}} + \left.\frac{\partial f(\vec{r},\vec{k},t)}{\partial t}\right|_{\text{collisions}} = 0. \tag{4.1}$$

It is customary to substitute the following differential form for the diffusion process

$$\left.\frac{\partial f(\vec{r},\vec{k},t)}{\partial t}\right|_{\text{diffusion}} = -\vec{v}(\vec{k}) \cdot \frac{\partial f(\vec{r},\vec{k},t)}{\partial \vec{r}} \tag{4.2}$$

which expresses the continuity equation in real space in the absence of forces, fields and collisions. For the forces and fields we write correspondingly

$$\left.\frac{\partial f(\vec{r},\vec{k},t)}{\partial t}\right|_{\text{fields}} = -\frac{\partial \vec{k}}{\partial t} \cdot \frac{\partial f(\vec{r},\vec{k},t)}{\partial \vec{k}} \tag{4.3}$$

to obtain the Boltzmann equation:

$$\frac{\partial f(\vec{r},\vec{k},t)}{\partial t} + \vec{v}(\vec{k}) \cdot \frac{\partial f(\vec{r},\vec{k},t)}{\partial \vec{r}} + \frac{\partial \vec{k}}{\partial t} \cdot \frac{\partial f(\vec{r},\vec{k},t)}{\partial \vec{k}} = \left.\frac{\partial f(\vec{r},\vec{k},t)}{\partial t}\right|_{\text{collisions}} \tag{4.4}$$

which includes derivatives for all the variables of the distribution function on the left hand side of the equation and the collision terms appear on the right hand side of Eq. 4.4. The first term in Eq. 4.4 gives the explicit time dependence of the distribution function and is needed for the solution of ac driving forces or for impulse perturbations. Boltzmann's equation is usually solved using two approximations:

1. The perturbation due to external fields and forces is assumed to be small so that the distribution function can be linearized and written as:

$$f(\vec{r},\vec{k}) = f_0(E) + f_1(\vec{r},\vec{k}) \tag{4.5}$$

   where $f_0(E)$ is the equilibrium distribution function (the Fermi function) which depends only on the energy E, while $f_1(\vec{r},\vec{k})$ is the perturbation term giving the departure from equilibrium.

2. The collision term in the Boltzmann equation is written in the **relaxation time approximation** so that the system returns to equilibrium uniformly:

$$\left.\frac{\partial f}{\partial t}\right|_{\text{collisions}} = -\frac{(f - f_0)}{\tau} = -\frac{f_1}{\tau} \tag{4.6}$$

where τ denotes the relaxation time and in general is a function of crystal momentum, i.e., $\tau = \tau(\vec{k})$. The physical interpretation of the relaxation time is the time associated with the rate of return to the equilibrium distribution when the external fields or thermal gradients are switched off. Solution to Eq. 4.6 when the fields are switched off at $t = 0$ leads to

$$\frac{\partial f}{\partial t} = -\frac{(f - f_0)}{\tau} \tag{4.7}$$

which has solutions

$$f(t) = f_0 + \Big[f(0) - f_0\Big] e^{-t/\tau} \tag{4.8}$$

where f_0 is the equilibrium distribution and $f(0)$ is the distribution function at time $t = 0$. The relaxation in Eq. 4.8 follows a Poisson distribution indicating that collisions relax the distribution function exponentially to f_0 with a time constant τ.

With these approximations, the Boltzmann equation is solved to find the distribution function which in turn determines the number density and current density. The current density $\vec{j}(\vec{r}, t)$ is given by

$$\vec{j}(\vec{r}, t) = \frac{e}{4\pi^3} \int \vec{v}(\vec{k}) f(\vec{r}, \vec{k}, t) d^3k \tag{4.9}$$

in which the crystal momentum $\hbar\vec{k}$ plays the role of the momentum $\vec{p}$ in specifying a volume in phase space. Every element of size h (Planck's constant) in phase space can accommodate one spin $\uparrow$ and one spin $\downarrow$ electron. The carrier density $n(\vec{r}, t)$ is thus simply given by integration of the distribution function over k-space

$$n(\vec{r}, t) = \frac{1}{4\pi^3} \int f(\vec{r}, \vec{k}, t) d^3k \tag{4.10}$$

where d^3k is an element of 3D wavevector space. The velocity of a carrier with crystal momentum $\hbar\vec{k}$ is related to the $E(\vec{k})$ dispersion expression by

$$\vec{v}(\vec{k}) = \frac{1}{\hbar} \frac{\partial E(\vec{k})}{\partial \vec{k}} \tag{4.11}$$

and $f_0(E)$ is the Fermi distribution function

$$f_0(E) = \frac{1}{1 + e^{(E - E_F)/k_B T}} \tag{4.12}$$

which defines the equilibrium state in which E_F is the Fermi energy and k_B is the Boltzmann constant.

4.3 Electrical Conductivity

To calculate the static electrical conductivity, we consider an applied electric field $\vec{E}$ which for convenience we will take along the x-direction. We will assume for the present that there is no magnetic field and that there are no thermal gradients present. The electrical

conductivity is expressed in terms of the conductivity tensor $\overleftrightarrow{\sigma}$ which is evaluated explicitly from the relation

$$\vec{j} = \overleftrightarrow{\sigma} \cdot \vec{E} \tag{4.13}$$

from solution of Eq. 4.9, using $\vec{v}(\vec{k})$ from Eq. 4.11 and the distribution function $f(\vec{r}, \vec{k}, t)$ from solution of the Boltzmann equation represented by Eq. 4.4. The first term in Eq. 4.4 vanishes since the dc applied field $\vec{E}$ has no time dependence.

For the second term in the Boltzmann equation Eq. 4.4, $\vec{v}(\vec{k}) \cdot \partial f(\vec{r}, \vec{k}, t)/\partial \vec{r}$, we note that

$$\frac{\partial f}{\partial \vec{r}} \simeq \frac{\partial f_0}{\partial \vec{r}} = \frac{\partial f_0}{\partial T} \frac{\partial T}{\partial \vec{r}}. \tag{4.14}$$

Since there are no thermal gradients present in the simplest calculation of the electrical conductivity given in this section, this term does not contribute to Eq. 4.4. For the third term in Eq. 4.4, which we write as

$$\dot{\vec{k}} \cdot \frac{\partial f(\vec{r}, \vec{k}, t)}{\partial \vec{k}} = \sum_{\alpha} \dot{k}_{\alpha} \frac{\partial f(\vec{r}, \vec{k}, t)}{\partial k_{\alpha}} \tag{4.15}$$

where the right hand side shows the summation over the vector components, we do get a contribution, since the equations of motion ($F = ma$) give

$$\hbar \dot{\vec{k}} = e\vec{E} \tag{4.16}$$

and

$$\frac{\partial f(\vec{r}, \vec{k}, t)}{\partial \vec{k}} = \frac{\partial (f_0 + f_1)}{\partial \vec{k}} = \frac{\partial f_0}{\partial E} \frac{\partial E}{\partial \vec{k}} + \frac{\partial f_1}{\partial \vec{k}}. \tag{4.17}$$

In considering the linearized Boltzmann equation, we retain only the leading terms in the perturbing electric field, so that $(\partial f_1/\partial \vec{k})$ can be neglected and only the term $(\partial f_0/\partial E)\hbar \vec{v}(\vec{k})$ need be retained. We thus obtain the linearized Boltzmann equation for the case on an applied static electric field and no thermal gradients:

$$\dot{\vec{k}} \cdot \frac{\partial f(\vec{r}, \vec{k}, t)}{\partial \vec{k}} = \frac{\phi}{\tau} \frac{\partial f_0}{\partial E} = -\frac{f_1}{\tau} \tag{4.18}$$

where it is convenient to write:

$$f_1 = -\phi \left(\frac{\partial f_0}{\partial E} \right) \tag{4.19}$$

in order to show the $(\partial f_0/\partial E)$ dependence explicitly. Substitution of Eqs. 4.16 and 4.17 into Eq. 4.18 yields

$$\left[\frac{e\vec{E}}{\hbar} \left(\frac{\partial f_0}{\partial E} \right) \right] \cdot [\hbar \vec{v}(\vec{k})] = \frac{\phi(\vec{k})}{\tau} \left(\frac{\partial f_0}{\partial E} \right) \tag{4.20}$$

so that

$$\phi(\vec{k}) = e\tau \vec{E} \cdot \vec{v}(\vec{k}). \tag{4.21}$$

Thus we can relate $\phi(\vec{k})$ to $f_1(\vec{k})$ by

$$f_1(\vec{k}) = -\phi(\vec{k}) \frac{\partial f_0(E)}{\partial E} = -e\tau \vec{E} \cdot \vec{v}(\vec{k}) \frac{\partial f_0(E)}{\partial E}. \tag{4.22}$$

The current density is then found from the distribution function $f(\vec{k})$ by calculation of the average value of $\langle ne\vec{v}\rangle$ over all k-space

$$\vec{j} = \frac{1}{4\pi^3}\int e\vec{v}(\vec{k})f(\vec{k})d^3k = \frac{1}{4\pi^3}\int e\vec{v}(\vec{k})f_1(\vec{k})d^3k \tag{4.23}$$

since

$$\int e\vec{v}(\vec{k})f_0(\vec{k})d^3k = 0. \tag{4.24}$$

Equation 4.24 states that no net current flows in the absence of an applied electric field, another statement of the equilibrium condition. Substitution for $f_1(\vec{k})$ given by Eq. 4.22 into Eq. 4.23 for $\vec{j}$ yields

$$\vec{j} = -\frac{e^2\vec{E}}{4\pi^3}\cdot\int \tau\vec{v}\vec{v}\frac{\partial f_0}{\partial E}d^3k \tag{4.25}$$

where in general $\tau = \tau(\vec{k})$ and $\vec{v}$ is given by Eq. 4.11. A comparison of Eqs. 4.25 and 4.13 thus yields the desired result for the conductivity tensor $\overleftrightarrow{\sigma}$

$$\overleftrightarrow{\sigma} = -\frac{e^2}{4\pi^3}\int \tau\vec{v}\vec{v}\,\frac{\partial f_0}{\partial E}d^3k \tag{4.26}$$

where $\overleftrightarrow{\sigma}$ is a symmetric second rank tensor ($\sigma_{ij} = \sigma_{ji}$). The evaluation of the integral in Eq. 4.26 over all k-space depends on the $E(\vec{k})$ relations through the $\vec{v}\vec{v}$ terms and the temperature dependence comes through the $\partial f_0/\partial E$ term.

4.4 Electrical Conductivity of Metals

To exploit the energy dependence of $(\partial f_0/\partial E)$ in applying Eq. 4.26 to metals, it is more convenient to evaluate $\overleftrightarrow{\sigma}$ if we replace $\int d^3k$ with an integral over the constant energy surfaces

$$\int d^3k = \int d^2S dk_\perp \equiv \int d^2S dE/|\partial E/\partial\vec{k}|. \tag{4.27}$$

Thus Eq. 4.26 is written

$$\overleftrightarrow{\sigma} = -\frac{e^2}{4\pi^3}\int \frac{\tau\vec{v}\vec{v}}{|\partial E/\partial\vec{k}|}\,\frac{\partial f_0}{\partial E}d^2S\,dE. \tag{4.28}$$

From the Fermi-Dirac distribution function $f_0(E)$ shown in Fig. 4.1, we see that the derivative $(-\partial f_0/\partial E)$ can approximately be replaced by a δ-function so that Eq. 4.28 can be written as

$$\overleftrightarrow{\sigma} = \frac{e^2}{4\pi^3\hbar}\int_{\text{Fermi surface}} \tau\vec{v}\vec{v}\,\frac{d^2S}{v}. \tag{4.29}$$

For a cubic crystal, $[v_x v_x] = v^2/3$ and thus $\overleftrightarrow{\sigma}$ has only diagonal components σ that are all equal to each other:

$$\sigma = \frac{e^2}{4\pi^3\hbar}\int_{\text{Fermi surface}} \tau v\,\frac{d^2S}{3} = \frac{ne^2\tau}{m^*} \tag{4.30}$$

since

$$n = (1/4\pi^3)(4\pi/3)k_F^3 \tag{4.31}$$

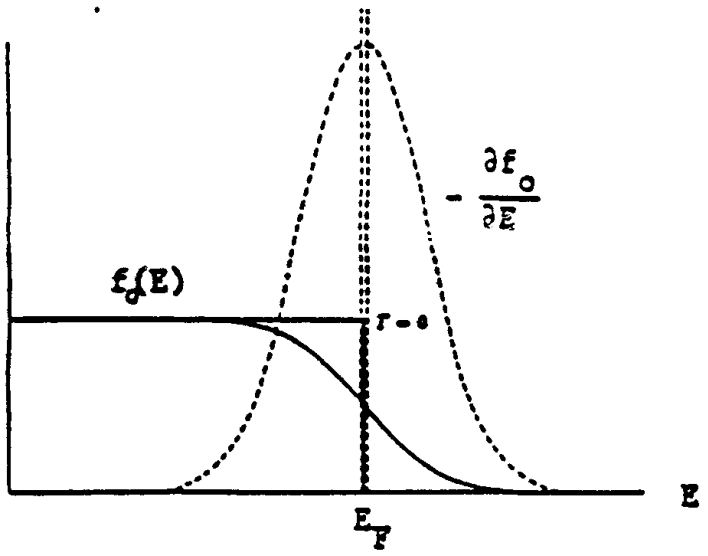


Figure 4.1: Schematic plot of $f_0(E)$ and $-\partial f_0(E)/\partial E$ for a metal showing the δ-function like behavior near the Fermi level E_F for the derivative.

and

$$v_F = \hbar k_F/m. \tag{4.32}$$

The result

$$\sigma = ne^2\tau/m^* \tag{4.33}$$

is called the Drude formula for the dc electrical conductivity.

4.5 Electrical Conductivity of Semiconductors

We show in this section that the simple Drude model $\sigma = ne^2\tau/m^*$ can also be recovered for a semiconductor from the general relation given by Eq. 4.26, using a simple parabolic band model and a constant relaxation time. When a more complete theory is used, departures from the simple Drude model will result.

In deriving the Drude model for a semiconductor we make three approximations:

- Approximation #1
  In the case of electron states in intrinsic semiconductors having no donor or acceptor impurities, we have the condition $(E - E_F) \gg k_BT$ since E_F is in the band gap and E is the energy of an electron in the conduction band, as shown in Fig. 4.2.

  Thus the first approximation is equivalent to writing

$$f_0(E) = \frac{1}{1 + \exp[(E - E_F)/k_BT]} \simeq \exp[-(E - E_F)/k_BT] \tag{4.34}$$

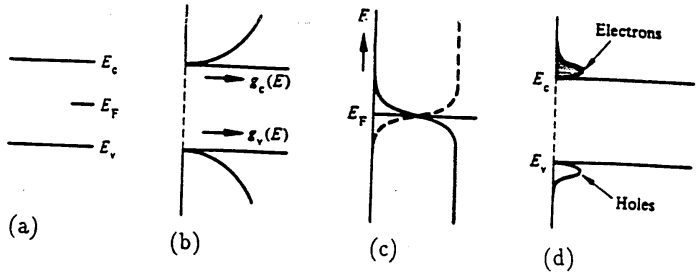


Figure 4.2: Electron and hole states in the conduction and valence bands of an intrinsic semiconductor. (a) Location of E_F in an intrinsic semiconductor. (b) The corresponding density of electron and hole states. (c) The Fermi functions for electrons and holes. (d) The occupation of electron and hole states in an intrinsic semiconductor.

which is equivalent to using the Maxwell–Boltzmann distribution in place of the full Fermi-Dirac distribution. Since E is usually measured with respect to the bottom of the conduction band, E_F is a negative energy and it is therefore convenient to write $f_0(E)$ as

$$f_0(E) \simeq e^{-|E_F|/k_BT} e^{-E/k_BT} \tag{4.35}$$

so that the derivative of the Fermi function becomes

$$\frac{\partial f_0(E)}{\partial E} = -\frac{e^{-|E_F|/k_BT}}{k_BT} e^{-E/k_BT}. \tag{4.36}$$

- Approximation #2
  For simplicity we assume a constant relaxation time τ independent of $\vec{k}$ and E. This approximation is made for simplicity and is not valid for specific cases. Some common scattering mechanisms yield an energy dependent relaxation time such as acoustic deformation potential scattering or ionized impurity scattering where $r = -1/2$ and $r = +3/2$ respectively in the relation $\tau = \tau_0(E/k_BT)^r$.

- Approximation #3
  To illustrate the explicit evaluation of the integral in Eq. 4.26, we consider the simplest case, assuming an isotropic, parabolic band $E = \hbar^2k^2/2m^*$ for the evaluation of $\vec{v} = \partial E/\hbar\partial\vec{k}$ about the conduction band extremum.

Using this third approximation we can write

$$\vec{v}\vec{v}=\tfrac{1}{3}v^2 \overleftrightarrow{1}$$

$$k^2=2m^*E/\hbar^2$$

$$2kdk=2m^*dE/\hbar^2 \tag{4.37}$$

$$v^2=2E/m^*$$

$$v=\hbar k/m^*$$

where $\overleftrightarrow{1}$ is the unit second rank tensor. We next convert Eq. 4.26 to an integration over energy and write

$$d^3k = 4\pi k^2 dk = 4\pi\sqrt{2}(m^*/\hbar^2)^{3/2}\sqrt{E}dE \tag{4.38}$$

so that Eq. 4.26 becomes

$$\sigma = \frac{e^2\tau}{4\pi^3}\left(\frac{8\sqrt{2}\pi\sqrt{m^*}}{3\hbar^3 k_B T}\right)e^{-|E_F/k_BT}\int_0^\infty E^{3/2}dEe^{-E/k_BT} \tag{4.39}$$

in which the integral over energy E is extended to ∞ because there is negligible contribution for large E and because the definite integral

$$\int_0^\infty x^p dx e^{-x} = \Gamma(p+1) \tag{4.40}$$

can be evaluated exactly, $\Gamma(p)$ being the Γ function which has the property

$$\Gamma(p+1)=p\Gamma(p)$$

$$\Gamma(1/2)=\sqrt{\pi}. \tag{4.41}$$

Substitution into Eq. 4.39 thus yields

$$\sigma = \frac{2e^2\tau}{m^*}\left(\frac{m^*k_BT}{2\pi\hbar^2}\right)^{3/2}e^{-|E_F|/k_BT} \tag{4.42}$$

which gives the temperature dependence of σ. Now the carrier density calculated using the same approximations becomes

$$\begin{aligned} n&=(4\pi^3)^{-1}\int f_0(E)\, d^3k \\ &=(4\pi^3)^{-1}\, e^{-|E_F|/k_BT}\int e^{-E/k_BT}4\pi\, k^2dk \\ &=(\sqrt{2}/\pi^2)\left(m^*/\hbar^2\right)^{3/2}e^{-|E_F|/k_BT}\textstyle\int_0^\infty \sqrt{E}dEe^{-E/k_BT} \end{aligned} \tag{4.43}$$

where

$$\int_0^\infty \sqrt{E}dEe^{-E/k_BT} = \frac{\sqrt{\pi}}{2}(k_BT)^{3/2} \tag{4.44}$$

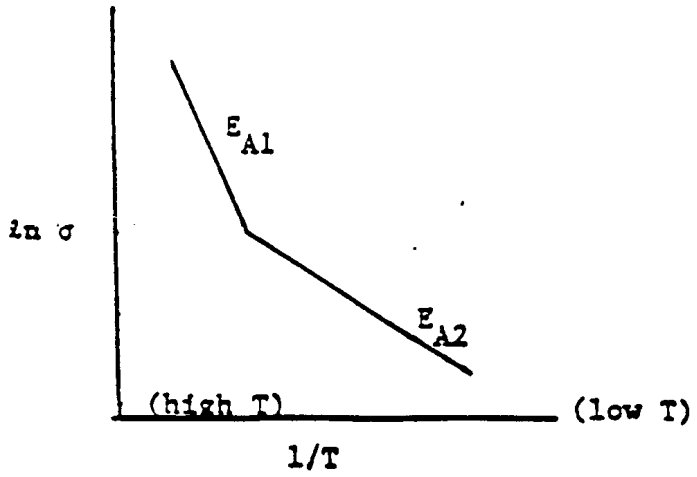


Figure 4.3: Schematic diagram of an Arrhenius plot of $\ln \sigma$ vs $1/T$ showing two carrier types with different activation energies.

which gives the final result for the temperature dependence of the carrier density

$$n = 2\left(\frac{m^* k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-|E_F|/k_B T} \tag{4.45}$$

so that by substitution into Eq. 4.42, the Drude formula is recovered

$$\sigma = \frac{ne^2\tau}{m^*} \tag{4.46}$$

for a semiconductor with constant τ and isotropic, parabolic dispersion relations.

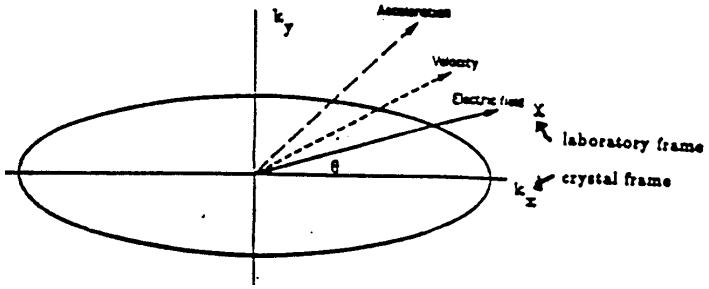
To find σ for a semiconductor with more than one spherical carrier pocket, the conductivities per carrier pocket are added

$$\sigma = \sum_i \sigma_i \tag{4.47}$$

where i is the carrier pocket index. We use these simple formulae to make rough estimates for the carrier density and conductivity of semiconductors. For more quantitative analysis, the details of the $E(\vec{k})$ relation must be considered, as well as an energy dependent τ and use of the complete Fermi function.

The electrical conductivity and carrier density of a semiconductor with one carrier type exhibits an exponential temperature dependence so that the slope of $\ln \sigma$ vs $1/T$ yields an activation energy (see Fig. 4.3). The plot of $\ln \sigma$ vs $1/T$ is called an "Arrhenius plot". If a plot of $\ln \sigma$ vs $1/T$ exhibits one temperature range with activation energy E_{A1} and a second temperature range with activation energy E_{A2}, then two carrier behavior is suggested. Also in such cases, the activation energies can be extracted from an Arrhenius plot as shown in the schematic of Fig. 4.3.

Figure 4.4: Schematic diagram of an ellipsoidal constant energy surface.



4.5.1 Ellipsoidal Carrier Pockets

The conductivity results given above for a spherical Fermi surface can easily be generalized to an ellipsoidal Fermi surface which is commonly found in degenerate semiconductors. Semiconductors are degenerate at $T = 0$ when the Fermi level is in the valence or conduction band rather than in the energy band gap.

For an ellipsoidal Fermi surface, we write

$$E(\vec{k}) = \frac{\hbar^2 k_x^2}{2m_{xx}} + \frac{\hbar^2 k_y^2}{2m_{yy}} + \frac{\hbar^2 k_z^2}{2m_{zz}} \tag{4.48}$$

where the effective mass components m_{xx}, m_{yy} and m_{zz} are appropriate to the band curvatures in the x, y, z directions, respectively. Substitution of

$$k'_\alpha = k_\alpha \sqrt{m_0/m_\alpha} \tag{4.49}$$

for $\alpha = x, y, z$ brings Eq. 4.48 into spherical form

$$E(\vec{k'}) = \frac{\hbar^2 k'^2}{2m_0} \tag{4.50}$$

where $k'^2 = k'^2_x + k'^2_y + k'^2_z$. For the volume element d^3k in Eq. 4.26 we have

$$d^3k = \sqrt{m_{xx}m_{yy}m_{zz}/m_0^3}\, d^3k' \tag{4.51}$$

and the carrier density associated with a single carrier pocket becomes

$$n_i = 2\sqrt{m_{xx}m_{yy}m_{zz}} \left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-|E_F|/k_B T}. \tag{4.52}$$

For an ellipsoidal constant energy surface (see Fig. 4.4), the directions of the electric field, electron velocity and electron acceleration will in general be different. Let (x, y, z) be the coordinate system for the major axes of the constant energy ellipsoid and (X, Y, Z) be

the laboratory coordinate system. Then in the laboratory system the current density $\vec{j}$ and electric field $\vec{E}$ are related by

$$\begin{pmatrix} j_X \\ j_Y \\ j_Z \end{pmatrix} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_{YY} & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_{ZZ} \end{pmatrix} \begin{pmatrix} E_X \\ E_Y \\ E_Z \end{pmatrix} \tag{4.53}$$

As an example, suppose that the electric field is applied in the XY plane along the X axis at an angle θ with respect to the x axis of the constant energy ellipsoid (see Fig. 4.4). The conductivity tensor is easily written in the xyz crystal coordinate system where the xyz axes are along the principal axes of the ellipsoid:

$$\begin{pmatrix} j_x \\ j_y \\ j_z \end{pmatrix} = ne^2\tau \begin{pmatrix} 1/m_{xx} & 0 & 0 \\ 0 & 1/m_{yy} & 0 \\ 0 & 0 & 1/m_{zz} \end{pmatrix} \begin{pmatrix} E\cos\theta \\ E\sin\theta \\ 0 \end{pmatrix} \tag{4.54}$$

A coordinate transformation from the crystal axes to the laboratory frame allows us to relate $\overleftrightarrow{\sigma}_{\text{crystal}}$ which we have written easily by Eq. 4.54 to $\overleftrightarrow{\sigma}_{\text{Lab}}$ which we measure by Eq. 4.53. In general

$$\overleftrightarrow{\sigma}_{\text{Lab}} = R\, \overleftrightarrow{\sigma}_{\text{crystal}}\, R^{-1} \tag{4.55}$$

where

$$R = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.56}$$

and

$$R^{-1} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.57}$$

so that the conductivity tensor $\overleftrightarrow{\sigma}_{\text{Lab}}$ in the lab frame becomes:

$$\overleftrightarrow{\sigma}_{\text{Lab}} = ne^2\tau \begin{pmatrix} \cos^2\theta/m_{xx} + \sin^2\theta/m_{yy} & \cos\theta\sin\theta(1/m_{yy} - 1/m_{xx}) & 0 \\ \cos\theta\sin\theta(1/m_{yy} - 1/m_{xx}) & \sin^2\theta/m_{xx} + \cos^2\theta/m_{yy} & 0 \\ 0 & 0 & 1/m_{zz} \end{pmatrix} \tag{4.58}$$

Semiconductors with ellipsoidal Fermi surfaces usually have several such surfaces located in crystallographically equivalent locations. In the case of cubic symmetry, the sum of the

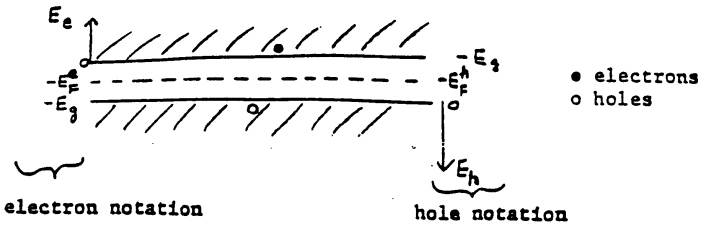


Figure 4.5: Schematic diagram of the band gap in a semiconductor showing the symmetry of electrons and holes.

conductivity components results in an isotropic conductivity even though the contribution from each ellipsoid is anisotropic. Thus measurement of the electrical conductivity provides no information on the anisotropy of the Fermi surfaces of cubic materials. However, measurement of the magnetoresistance does provide such information, since the application of a magnetic field gives special importance to the magnetic field direction, thereby lowering the effective crystal symmetry.

4.6 Electrons and Holes in Intrinsic Semiconductors

In the absence of doping, carriers are generated by thermal or optical excitations. Thus at $T = 0$, all valence band states are occupied and all conduction band states are empty. Thus, for each electron that is excited into the conduction band, a hole is left behind. For intrinsic semiconductors, conduction is by both holes and electrons. The Fermi level is thus determined by the condition that the number of electrons is equal to the number of holes. Writing $g_v(E_h)$ and $g_c(E_e)$ as the density of hole and electron states, respectively, we obtain

$$n_h = \int_0^\infty g_v(E_h)\hat{f}_0(E_h + E_F^h)dE_h = \int_0^\infty g_c(E_e)\hat{f}_0(E_e + E_F^e)dE_e = n_e \tag{4.59}$$

where the notation we have used is shown in Fig. 4.5. Here the energy gap E_g is written as

$$E_F^e + E_F^h = E_g. \tag{4.60}$$

The condition $n_e = n_h$ for intrinsic semiconductors is used to determine the position of the Fermi levels for electrons and holes within the band gap. If the band curvatures of the valence and conduction bands are the same, then their effective masses are the same magnitude and E_F lies at midgap. We derive in this section the general result for the placement of E_F when $m_e^* \neq m_h^*$.

On the basis of this interpretation, the holes obey Fermi statistics as do the electrons, only we must measure their energies downward, while electron energies are measured upwards, as indicated in Fig. 4.5. This approach clearly builds on the symmetry relation between electrons and holes. It is convenient to measure electron energies E_e with respect to the bottom of the conduction band E_c so that $E_e = E - E_c$ and hole energies E_h with respect to the top of the valence band E_v so that $E_h = -(E - E_v)$ and the Fermi level for

the electrons is $-E_F^e$ and for holes is $-E_F^h$. Referring to Eq. 4.59, $\hat{f}_0(E_e + E_F^e)$ denotes the Fermi function where the Fermi energy is written explicitly

$$\hat{f}_0(E_e + E_F^e) = \frac{1}{1 + \exp[(E + E_F^e)/k_B T]} \tag{4.61}$$

and is consistent with the definitions given above.

In an intrinsic semiconductor, the magnitudes of the energies E_F^e and E_F^h are both much greater than thermal energies, i.e., $|E_F^e| \gg k_B T$ and $|E_F^h| \gg k_B T$, so that the distribution functions can be approximated by the Boltzmann form

$$\begin{aligned} \hat{f}_0(E_e + E_F^e) &\simeq e^{-(E_e + E_F^e)/k_B T} \\ \hat{f}_0(E_h + E_F^h) &\simeq e^{-(E_h + E_F^h)/k_B T}. \end{aligned} \tag{4.62}$$

If m_e and m_h are respectively the electron and hole effective masses and if we write the dispersion relations around the valence and conduction band extrema as

$$\begin{aligned} E_e &= \hbar^2 k^2/(2m_e) \\ E_h &= \hbar^2 k^2/(2m_h) \end{aligned} \tag{4.63}$$

then the density of states for electrons at the bottom of the conduction band and for holes at the top of the valence band can be written in their respective nearly free electron forms (see Eq. 4.68)

$$\begin{aligned} g_c(E_e) &= \tfrac{1}{2\pi^2}\left(2m_e/\hbar^2\right)^{3/2} E_e^{1/2} \\ g_v(E_h) &= \tfrac{1}{2\pi^2}\left(2m_h/\hbar^2\right)^{3/2} E_h^{1/2}. \end{aligned} \tag{4.64}$$

These expressions follow from

$$n = \frac{1}{4\pi^3}\frac{4\pi}{3}k^3 \tag{4.65}$$

and substitution of k via the simple parabolic relation

$$E = \frac{\hbar^2 k^2}{2m^*} \tag{4.66}$$

so that

$$n = \frac{1}{3\pi^2}\left(\frac{2m^* E}{\hbar^2}\right)^{3/2} \tag{4.67}$$

and

$$g(E) = \frac{dn}{dE} = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} E^{1/2}. \tag{4.68}$$

Substitution of this density of states expression into Eq. 4.45 results in a carrier density

$$n_e = 2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-E_F^e/k_B T}. \tag{4.69}$$

Likewise for holes we obtain

$$n_h = 2\left(\frac{m_h k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-E_F^h/k_B T}. \tag{4.70}$$

Thus the famous product rule is obtained

$$n_e n_h = 4\left(\frac{k_B T}{2\pi\hbar^2}\right)^3 (m_e m_h)^{3/2} e^{-E_g/k_B T} \tag{4.71}$$

where $E_g = E_F^e + E_F^h$. But for an intrinsic semiconductor $n_e = n_h$. Thus by taking the square root of the above expression, we obtain both n_e and n_h

$$n_e = n_h = 2\left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2} (m_e m_h)^{3/4} e^{-E_g/2k_B T}. \tag{4.72}$$

Comparison with the expressions given in Eqs. 4.69 and 4.70 for n_e and n_h allows us to solve for the Fermi levels E_F^e and E_F^h

$$n_e = 2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-E_F^e/k_B T} = 2\left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2} (m_e m_h)^{3/4} e^{-E_g/2k_B T} \tag{4.73}$$

so that

$$\exp(-E_F^e/k_B T) = (m_h/m_e)^{3/4} \exp(-E_g/2k_B T) \tag{4.74}$$

and

$$E_F^e = \frac{E_g}{2} - \frac{3}{4} k_B T \ln(m_h/m_e). \tag{4.75}$$

If $m_e = m_h$, we obtain the simple result that $E_F^e = E_g/2$ which says that the Fermi level lies in the middle of the energy gap. However, if the masses are not equal, E_F will lie closer to the band edge with higher curvature, thereby enhancing the Boltzmann factor term in the thermal excitation process, to compensate for the lower density of states for the higher curvature band.

If however $m_e \ll m_h$, the Fermi level approaches the conduction band edge and the full Fermi functions have to be considered. In this case

$$n_e = \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2} \int_{E_c}^{\infty} \frac{(E - E_c)^{1/2} dE}{\exp[(E - E_F^e)/k_B T] + 1} \equiv N_e F_{1/2}\left(\frac{E_F^e - E_c}{k_B T}\right) \tag{4.76}$$

where E_c is the bottom of the conduction band and the "effective electron density" is in accordance with Eq. 4.69 given by

$$N_e = 2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} \tag{4.77}$$

and the Fermi integral is written as

$$F_j(\eta) = \frac{1}{j!} \int_0^{\infty} \frac{x^j dx}{\exp(x - \eta) + 1}. \tag{4.78}$$

We can take $F_j(\eta)$ from the tables in Blakemore, "Semiconductor Physics" (Appendix B). For the semiconductor limit ($\eta < -4$), then $F_j(\eta) \to \exp(\eta)$. Clearly, when $F_j(\eta)$ is required to describe the carrier density, then $F_j(\eta)$ is also needed to describe the conductivity. These refinements are important for a detailed solution of the transport properties of semiconductors over the entire temperature range of interest.

Table 4.1: Occupation of impurity states in the grand canonical ensemble.

j	N_j	spin	E_j
1	0	-	0
2	1	$\uparrow$	$-E_d$
3	1	$\downarrow$	$-E_d$
4	2	$\uparrow\downarrow$	$-E_d + E_{\text{Coulomb}}$

4.7 Donor and Acceptor Doping of Semiconductors

In general a semiconductor has electron and hole carriers due to the presence of impurities as well as from thermal excitation processes. For many applications, impurities are intentionally introduced to generate carriers: donor impurities to generate electrons in n-type material and acceptor impurities to generate holes in p-type material. Assuming for the moment that each donor contributes one electron to the conduction band, then the donors can contribute an excess carrier concentration up to N_d, where N_d is the donor impurity concentration. Similarly, if every acceptor contributes one hole to the valence band, then the excess hole concentration will be N_a, where N_a is the acceptor impurity concentration. In general, the semiconductor is **partly compensated**, with both donor and acceptor impurities present. Furthermore, at finite temperatures, the donor and acceptor levels will be partially occupied so that somewhat less than the maximum charge will be released as **mobile** charge into the conduction and valence bands. The density of electrons bound to a donor site n_d is found from the grand canonical ensemble as

$$\frac{n_d}{N_d} = \frac{\sum_j N_j e^{-(E_j - \mu N_j)/k_B T}}{\sum_j e^{-(E_j - \mu N_j)/k_B T}} \tag{4.79}$$

where E_j and N_j are respectively the energy and number of electrons that can be placed in state j, and μ is the chemical potential (Fermi energy). Referring to Table 4.1, the system can be found in one of three states: one where no electrons are present (hence no contribution is made to the energy), and two states (one with spin $\uparrow$, the other with spin $\downarrow$) corresponding to the donor energy E_d, where E_d is a positive energy. Placing two electrons in the same energy state would result in a very high energy because of the Coulomb repulsion between the two electrons; therefore this possibility is neglected in practical calculations. Writing either $N_j = 0, 1$ for the 3 states of importance, we obtain for the relative ion concentration of occupied donor sites

$$\frac{n_d}{N_d} = \frac{2e^{-(\varepsilon_d - \mu)/k_B T}}{1 + 2e^{-(\varepsilon_d - \mu)/k_B T}} = \frac{1}{1 + \frac{1}{2}e^{(\varepsilon_d - \mu)/k_B T}} = \frac{1}{1 + \frac{1}{2}e^{-(E_d - E_F^e)/k_B T}} \tag{4.80}$$

in which E_d and E_F^e are positive numbers and the zero of energy is taken at the bottom of the conduction band. The energy ε_d denotes the energy for the donor level.

Consequently, the concentration of electrons thermally ionized into the conduction band will be

$$N_d - n_d = \frac{N_d}{1 + 2e^{(E_d - E_F^e)/k_B T}} = n_e - n_h \tag{4.81}$$

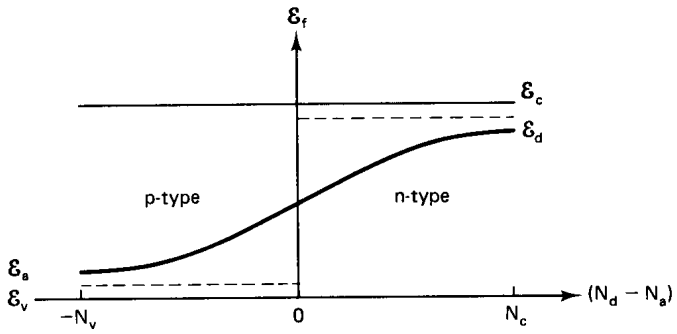


Figure 4.6: Variation of the Fermi energy with donor and acceptor concentrations. For a heavily doped n-type semiconductor E_F is close to the donor level E_d. Also shown on the figure are the bottom of the conduction band E_c and the donor energy E_d. This plot is made assuming almost all the donor and acceptor states are ionized.

where n_e and n_h are the mobile electron and hole concentrations. At low temperatures, where $E_d \sim k_B T$, almost all of the carriers in the conduction band will be generated by the ionized donors, so that $n_h \ll n_e$ and $(N_d - n_d) \simeq n_e$. The Fermi level will then adjust itself so that from Eqs. 4.45 and 4.81 the following equation determines E_F^e:

$$n_e = 2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-E_F^e/k_B T} \simeq \frac{N_d}{1 + 2e^{(E_d - E_F^e)/k_B T}}. \tag{4.82}$$

Solution of Eq. 4.82 shows that the presence of the ionized donor carriers moves the Fermi level up above the middle of the band gap and close to the bottom of the conduction band. For the donor impurity problem, the Fermi level will be close to the position of the donor level E_d, as shown in Fig. 4.6. The position of the Fermi level also varies with temperature. Figure 4.6 assumes that almost all the donor electrons (or acceptor holes) are ionized and are in the conduction band, which is typical of temperatures where the n-type (or p-type) semiconductor would be used. Figure 4.7 shows the dependence of the Fermi level on temperature. Here T_1 denotes the temperature at which the thermal excitation of intrinsic electrons and holes become important, and T_1 is normally a high temperature. In contrast, T_2 is normally a very low temperature and denotes the temperature below which donor-generated electrons begin to freeze out in impurity level bound states and no longer contribute to conduction. This carrier freeze-out is illustrated in Fig. 4.8. In the temperature range $T_2 < T < T_1$, the Fermi level in Fig. 4.7 falls as T increases according to

$$E_F = E_c - k_B T \ln(N_c/N_d) \tag{4.83}$$

where $N_c = 2m_e k_B T/(2\pi\hbar^2)$. In Fig. 4.8 we see the temperature dependence of the carrier concentration in the intrinsic range ($T > T_1$), the saturation range ($T_2 < T < T_1$), and finally the low temperature range ($T < T_2$) where carriers freeze out into bound states in the impurity band at E_d. The plot is as a function of $(1/T)$ and the corresponding temperature values are shown on the upper scale of the figure.

For the case of acceptor impurities, an ionized acceptor level releases a hole into the valence band, or alternatively, an electron from the valence band gets excited into an acceptor level, leaving a hole behind. At very low temperature, the acceptor levels are filled with

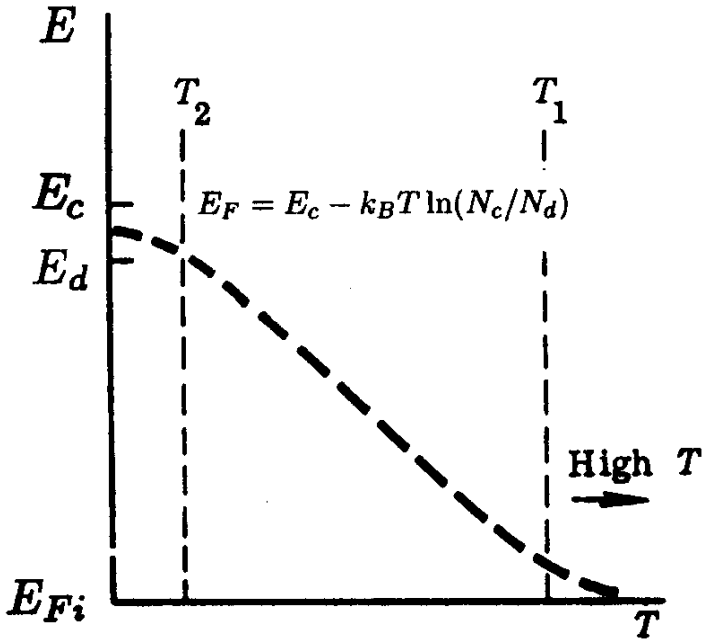


Figure 4.7: Temperature dependence of the Fermi energy for an n-type doped semiconductor. See the text for definitions of T_1 and T_2. Here E_{Fi} is the position of the Fermi level in the high temperature limit where the thermal excitation of carriers far exceeds the electron density contributed by the donor impurities.

holes. Because of hole-hole Coulomb repulsion, we can place no more than one hole in each acceptor level. A singly occupied hole can have either spin up or spin down. Thus for the acceptor levels, a formula analogous to Eq. 4.80 for donors is obtained for the occupation of an acceptor level

$$\frac{n_a}{N_a} = \frac{1}{1 + \frac{1}{2}e^{-(E_a - E_F^h)/k_B T}} \tag{4.84}$$

so that the essential symmetry between holes and electrons is maintained.

A situation which commonly arises for the acceptor levels relates to the degeneracy of the valence bands for group IV and III-V compound semiconductors. We will illustrate the degenerate valence band in the case where spin-orbit interaction is considered and the two degenerate levels are only weakly coupled, so that we can approximate the impurity acceptor levels by hydrogenic acceptor levels for the heavy hole $\varepsilon_{a,h}$ and light hole $\varepsilon_{a,l}$ bands. In this case the split-off band does not contribute significantly. The density of holes bound to both types of acceptor sites is given by

$$\frac{n_a}{N_a} = \frac{\sum_j N_j e^{-(E_j - \mu N_j)/k_B T}}{\sum_j e^{-(E_j - \mu N_j)/k_B T}}. \tag{4.85}$$

Using the same arguments as above, we obtain:

$$\frac{n_a}{N_a} = \frac{2e^{-(\varepsilon_{a,l} - \mu)/k_B T} + 2e^{-(\varepsilon_{a,h} - \mu)/k_B T}}{1 + 2e^{-(\varepsilon_{a,l} - \mu)/k_B T} + 2e^{-(\varepsilon_{a,h} - \mu)/k_B T}} \tag{4.86}$$

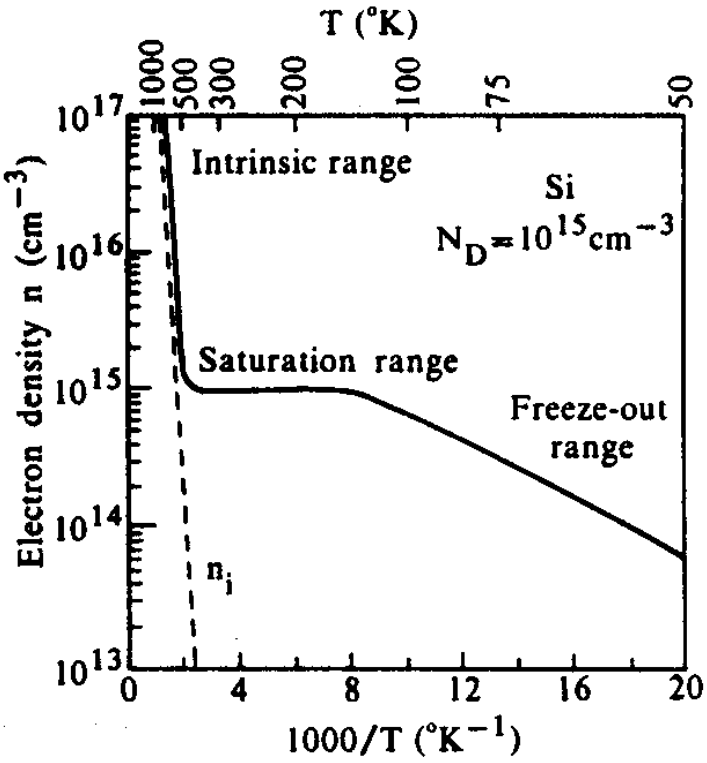


Figure 4.8: Temperature dependence of the electron density for Si doped with 10^{15} cm^{-3} donors.

so that

$$\frac{n_a}{N_a} = \frac{1 + e^{-(\varepsilon_{a,h}-\varepsilon_{a,l})/k_BT}}{1 + \frac{1}{2}e^{(\varepsilon_{a,l}-\mu)/k_BT} + e^{-(\varepsilon_{a,h}-\varepsilon_{a,l})/k_BT}}. \tag{4.87}$$

If the thermal energy is large in comparison to the difference between the acceptor levels for the heavy and light hole bands, then

$$[(\varepsilon_{a,h} - \varepsilon_{a,l})/k_BT] \ll 1 \tag{4.88}$$

and

$$\exp[-(\varepsilon_{a,h} - \varepsilon_{a,l})/k_BT] \simeq 1 \tag{4.89}$$

so that

$$\frac{n_a}{N_a} \simeq \frac{1}{1 + \frac{1}{4}e^{-(\varepsilon_{a,l}-\mu)/k_BT}} = \frac{1}{1 + \frac{1}{4}e^{(E_a-E_F^h)/k_BT}} \tag{4.90}$$

where E_a and E_F^h are the positive values corresponding to $\varepsilon_{a,l}$ and μ, respectively. From Eqs. 4.81 and 4.90, the temperature dependence of E_F can be calculated for the case of doped semiconductors. Figure 4.6 shows the doping dependence of E_F for p-doped semiconductors as well as n-doped semiconductors.

4.8 Characterization of Semiconductors

In describing the electrical conductivity of semiconductors, it is customary to write the conductivity as

$$\sigma = n_e|e|\mu_e + n_h|e|\mu_h \tag{4.91}$$

in which n_e and n_h are the carrier densities for the carriers, and μ_e and μ_h are their **mobilities**. We have shown in Eq. 4.46 that for cubic materials the static conductivity can under certain approximations be written as

$$\sigma = \frac{ne^2\tau}{m^*} \tag{4.92}$$

for each carrier type, so that the mobilities and effective masses are related by

$$\mu_e = \frac{|e|\langle\tau_e\rangle}{m_e} \tag{4.93}$$

and

$$\mu_h = \frac{|e|\langle\tau_h\rangle}{m_h} \tag{4.94}$$

which show that materials with small effective masses have high mobilities. By writing the electrical conductivity as a product of the carrier density with the mobility, it is easy to contrast the temperature dependence of σ for metals and semiconductors. For metals, the carrier density n is essentially independent of T, while μ is temperature dependent. In contrast, n for semiconductors is highly temperature dependent in the intrinsic regime (see Fig. 4.8) and μ is relatively less temperature dependent. Figure 4.9 shows the carrier concentration for intrinsic Si and Ge in the neighborhood of room temperature, demonstrating the rapid increase of the carrier concentration. These values of n indicate the doping levels

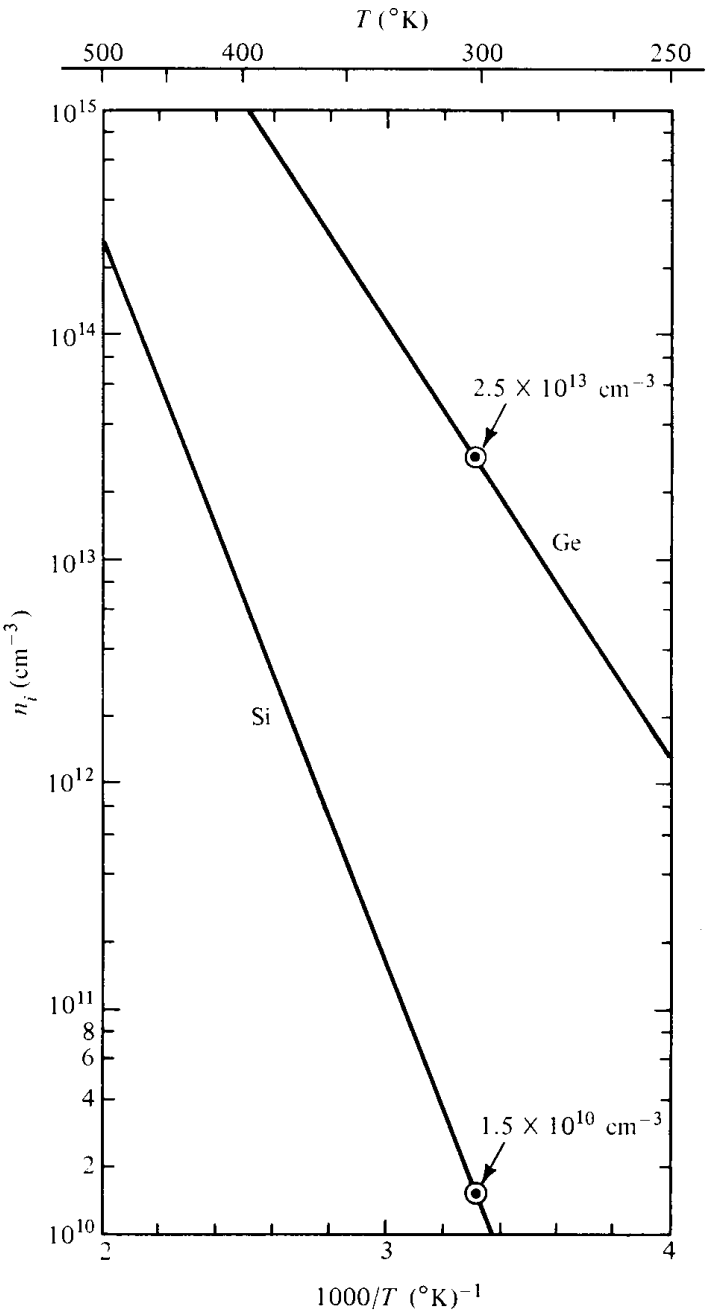


Figure 4.9: Temperature dependence of the electron concentration for intrinsic Si and Ge in the range $250 < T < 500$ K.

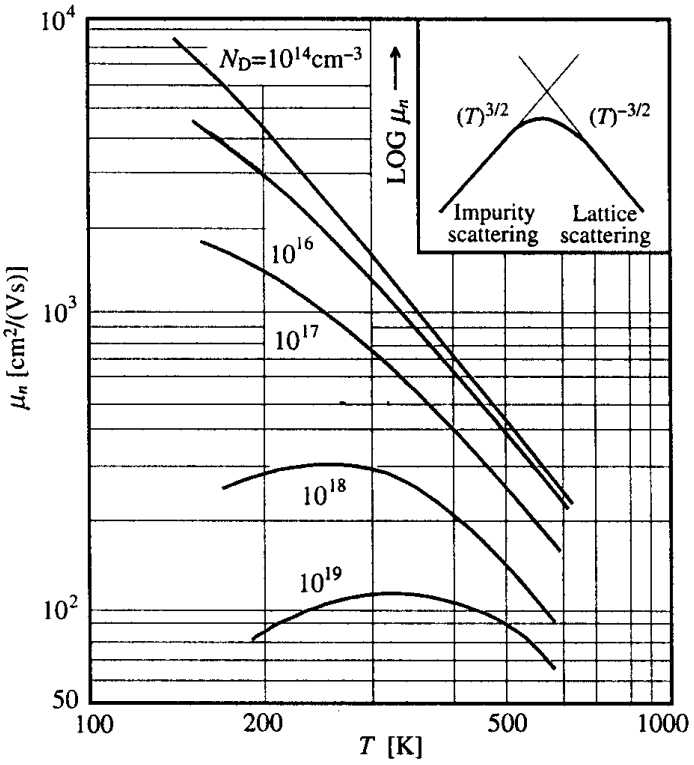


Figure 4.10: Temperature dependence of the mobility for n-type Si for a series of samples with different impurity concentrations. Note that the mobility is not as strong a function of temperature as is the carrier density shown in Fig. 4.9. At low temperature impurity scattering by the donor ions becomes important as shown in the inset.

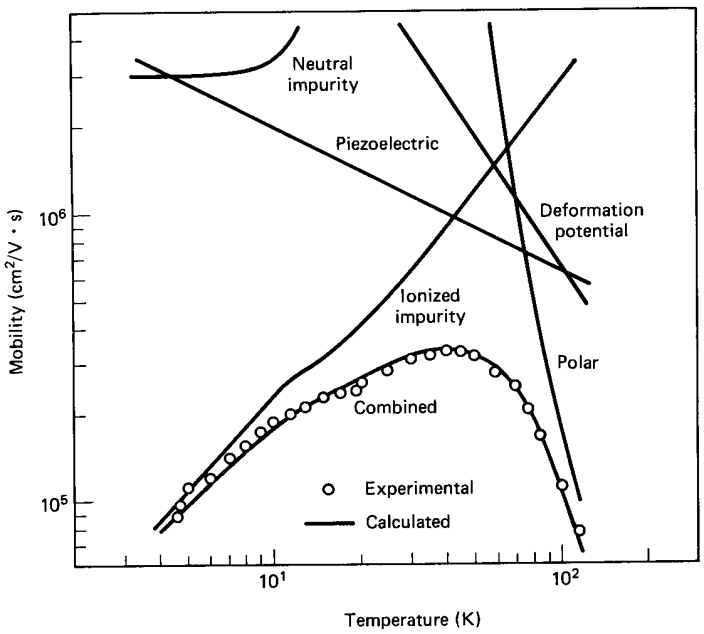


Figure 4.11: Temperature dependence of the mobility for n-type GeAs showing the separate and combined scattering processes.

necessary to exceed the intrinsic carrier level at a given temperature. Figure 4.10 shows the mobility for Ge samples with various impurity levels. The dependence expected for the intrinsic sample is also indicated. the observed temperature dependence can be explained by the different temperature dependences of the impurity scattering and phonon scattering mechanisms (see Fig. 4.11). This is further discussed in Chapter 6.

A table of typical mobilities for semiconductors is given in Table 4.2. By way of comparison, μ for copper at room temperature is 35cm^2/volt-sec. When using conductivity formulae in esu units, remember that the mobility is expressed in cm^2/statvolt-sec and that all the numbers in Table 4.2 have to be multiplied by 300 to match the units given in the notes.

In the characterization of a semiconductor for device applications, we are expected to provide information on the carrier density and mobility, preferably as a function of temperature. Such plots are shown in Figs. 4.9 and 4.10. When presenting characterization data in condensed form, the carrier density and mobility of semiconductors are traditionally given at 300 K and 77 K. Other information of values in semiconductor physics are values of the effective masses (Table 4.3). and of the energy gaps (Table 4.4)

Table 4.2: Mobilities for some typical semiconductors at room temperature in units of cm^2/V-sec.

Crystal	Electrons	Holes	Crystal	Electrons	Holes
Diamond	1800	1200	GaAs	8000	300
Si	1350	480	GaSb	5000	1000
Ge	3600	1800	PbS	550	600
InSb	77,000	750	PbSe	1020	930
InAs	30,000	460	PbTe	2500	1000
InP	4600	100	AgCl	50	–
AlAs	280	–	KBr (100K)	100	–
AlSb	900	400	SiC	100	10–20

Table 4.3: Semiconductor effective masses of electrons and holes in direct gap materials.

Crystal	Electron	Holes			Spin-orbit
	m_e/m_0	m_{hh}/m_0	m_{lh}/m_0	m_{soh}/m_0	Δ(eV)
InSb	0.015	0.39	0.021	(0.11)	0.82
InAs	0.026	0.41	0.025	0.08	0.43
InP	0.073	0.4	(0.078)	(0.15)	0.11
GaSb	0.047	0.3	0.06	(0.14)	0.80
GaAs	0.066	0.5	0.082	0.17	0.34
Cu_2O	0.99	–	0.58	0.69	0.13

Table 4.4: Semiconductor energy gaps between the valence and conduction bands.

		E_g, eV				E_g, eV	
Crystal	Gap[a]	0 K	300 K	Crystal	Gap[a]	0 K	300 K
Diamond	*i*	5.4	–	HgTe[b]	*d*	–0.30	–
Si	*i*	1.17	1.11	PbS	*d*	0.286	0.34-0.37
Ge	*i*	0.744	0.66	PbSe	*i*	0.165	0.27
αSn	*d*	0.00	0.00	PbTe	*i*	0.190	0.29
InSb	*d*	0.23	0.17	CdS	*d*	2.582	2.42
InAs	*d*	0.43	0.36	CdSe	*d*	1.840	1.74
InP	*d*	1.42	1.27	CdTe	*d*	1.607	1.44
GaP	*i*	2.32	2.25	ZnO	–	3.436	3.2
GaAs	*d*	1.52	1.43	ZnS	–	3.91	3.6
GaSb	*d*	0.81	0.68	SnTe	*d*	0.3	0.18
AlSb	*i*	1.65	1.6	AgCl	–	–	3.2
SiC(hex)	*i*	3.0	–	AgI	–	–	2.8
Te	*d*	0.33	–	Cu_2O	*d*	2.172	–
ZnSb	–	0.56	0.56	TiO_2	–	3.03	–

[a]The indirect gap is labelled by *i*, and the direct gap is labelled by *d*.
[b]HgTe is a semimetal and the bands overlap, hence a negative gap.

Chapter 5

Thermal Transport

References:

- Ziman, *Principles of the Theory of Solids*, Cambridge Univ. Press, 1972, Chapters 7.
- Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill, 1965, pp. 393-397.
- Wolfe, Holonyak and Stillman, *Physical Properties of Semiconductors*, Prentice Hall, 1989, chapter 5.

5.1 Thermal Transport

The electrons in solids not only conduct electricity but also conduct heat, as they transfer energy from a hot junction to a cold junction. Just as the electrical conductivity characterizes the response of a material to an applied voltage, the thermal conductivity likewise characterizes the material with regard to heat flow. In fact the electrical conductivity and thermal conductivity are coupled, since thermal conduction also transports charge and electrical conduction transports energy. This coupling gives rise to thermo-electricity. In this chapter, we discuss first the thermal conductivity for metals, semiconductors and insulators and then consider the coupling between electrical and thermal transport which gives rise to thermoelectric phenomena. In Chapter 6, we discuss scattering mechanisms for electrons and phonons.

5.2 Thermal Conductivity

5.2.1 General Considerations

Thermal transport, like electrical transport follows from the Boltzmann equation. We will first derive a general expression for the electronic contribution to the thermal conductivity using Boltzmann's equation. We will then apply this general expression to find the thermal conductivity for metals and then for semiconductors. The total thermal conductivity $\overleftrightarrow{\kappa}$ of any material is, of course, the superposition of the electronic part $\overleftrightarrow{\kappa}_e$ with the lattice part $\overleftrightarrow{\kappa}_L$:

$$\overleftrightarrow{\kappa} = \overleftrightarrow{\kappa}_e + \overleftrightarrow{\kappa}_L \ . \tag{5.1}$$

We now consider the calculation of the electronic contribution to the thermal conductivity. The application of a temperature gradient to a solid gives rise to a flow of heat. We define $\vec{U}$ as the thermal current which is driven by the heat energy $E - E_F$, which is the excess energy of an electron above the equilibrium energy E_F. Neglecting time dependent effects, we define $\vec{U}$ as

$$\vec{U} = \frac{1}{4\pi^3} \int \vec{v}\,(E - E_F)\, f(\vec{r}, \vec{k}) d^3k \tag{5.2}$$

where the distribution function $f(\vec{r}, \vec{k})$ is related to the Fermi function f_0 by $f = f_0 + f_1$. Under equilibrium conditions there is no thermal current density

$$\int \vec{v}(E - E_F) f_0 \; d^3k = 0 \tag{5.3}$$

so that the thermal current is driven by the thermal gradient which causes a departure from the equilibrium distribution:

$$\vec{U} = \frac{1}{4\pi^3} \int \vec{v}(E - E_F) f_1 \; d^3k \tag{5.4}$$

where the electronic contribution to the thermal conductivity tensor $\overleftrightarrow{\kappa}_e$ is defined by the relation

$$\vec{U} = - \overleftrightarrow{\kappa}_e \cdot \frac{\partial T}{\partial \vec{r}}. \tag{5.5}$$

Assuming no explicit time dependence for the distribution function, the function f_1 representing the departure of the distribution from equilibrium is found from solution of Boltzmann's equation

$$\vec{v} \cdot \frac{\partial f}{\partial \vec{r}} + \dot{\vec{k}} \cdot \frac{\partial f}{\partial \vec{k}} = -\frac{f_1}{\tau}. \tag{5.6}$$

In the absence of an electric field, $\dot{\vec{k}} = 0$ and the drift velocity $\vec{v}$ is found from the equation

$$\vec{v} \cdot \frac{\partial f}{\partial \vec{r}} = -\frac{f_1}{\tau}. \tag{5.7}$$

Using the linear approximation for the term $\partial f / \partial \vec{r}$ in the Boltzmann equation, we obtain

$$\begin{aligned} \tfrac{\partial f}{\partial \vec{r}} \cong \tfrac{\partial f_0}{\partial \vec{r}} &= \tfrac{\partial}{\partial \vec{r}} [\tfrac{1}{1+e^{(E-E_F)/k_BT}}] \\ &= \{ -\tfrac{e^{(E-E_F)/k_BT}}{[1+e^{(E-E_F)/k_BT}]^2} \} \{ -\tfrac{1}{k_BT} \tfrac{\partial E_F}{\partial \vec{r}} - \tfrac{(E-E_F)}{k_BT^2} \tfrac{\partial T}{\partial \vec{r}} \} \\ &= \{ k_BT \tfrac{\partial f_0}{\partial E} \} \{ -\tfrac{1}{k_BT} \tfrac{\partial T}{\partial \vec{r}} \} \{ \tfrac{\partial E_F}{\partial T} + \tfrac{(E-E_F)}{T} \} = -\tfrac{\partial f_0}{\partial E} \{ \tfrac{\partial E_F}{\partial T} + \tfrac{(E-E_F)}{T} \} \tfrac{\partial T}{\partial \vec{r}}. \end{aligned} \tag{5.8}$$

We will now give some typical values for these two terms for semiconductors and metals. For semiconductors, we evaluate the expression in Eq. 5.8 by referring to Eq. 4.75

$$E_F = \frac{1}{2} E_g - \frac{3}{4} k_B T \; \ell n(m_h / m_e) \tag{5.9}$$

from which

$$\frac{\partial E_F}{\partial T} \sim \frac{3}{4} k_B \; \ell n(m_h / m_e) \tag{5.10}$$

showing that the temperature dependence of E_F arises from the inequality of the valence and conduction band effective masses. If $m_h = m_e$, which would be the case of strongly coupled "mirror" bands, then $\partial E_F/\partial T$ would vanish. For a significant mass difference such as m_h/m_e=2, we obtain $\partial E_f/\partial T \sim 0.5 k_B$ from Eq. 5.10. For a band gap of 0.5 eV and the Fermi level in the middle of the gap, we obtain for the other term in Eq. 5.8

$$[(E - E_F)/T] \approx [0.5/(1/40)]k_B = 20k_B \tag{5.11}$$

where $k_B T \approx 1/40$ eV at room temperature. Thus for a semiconductor, the term $(E-E_F)/T$ is much larger than the term $(\partial E_F/\partial T)$.

For a metal with a spherical Fermi surface, the following relation

$$E_F = E_F^0 - \frac{\pi^2}{12}\frac{(k_B T)^2}{E_F^0} \tag{5.12}$$

is derived in standard textbooks on statistical mechanics, so that at room temperature and assuming that for a typical metal $E_F^0 = 5$ eV, we obtain from Eq. 5.12

$$\left|\frac{\partial E_F}{\partial T}\right| = \frac{\pi^2}{6}\frac{(k_B T)}{E_F^0}k_B \approx \frac{10}{6}\frac{(\frac{1}{40})}{5}k_B \approx 8 \times 10^{-3} k_B. \tag{5.13}$$

Thus, for both semiconductors and metals, the term $(E - E_F)/T$ tends to dominate over $(\partial E_F/\partial T)$, though there can be situations where the term $(\partial E_F/\partial T)$ cannot be neglected. In this presentation, we will temporarily neglect the term ∇E_F in Eq. 5.8 in calculation of the electronic contribution to the thermal conductivity, but we will include this term formally in our derivation of thermoelectric effects in §5.3.

Typically, the electron energies of importance in any transport problem are those within $k_B T$ of the Fermi energy so that for many applications for metals, we can make the rough approximation,

$$\frac{E - E_F}{T} \approx k_B \tag{5.14}$$

though the results given in this section are derived without the above approximation for $(E - E_F)/T$. Rather, all integrations are carried out in terms of the variable $(E - E_F)/T$.

We return now to the solution of the Boltzmann equation in the relaxation time approximation

$$\frac{\partial f}{\partial \vec{r}} = \left(-\frac{\partial f_0}{\partial E}\right)\left(\frac{E - E_F}{T}\right)\left(\frac{\partial T}{\partial \vec{r}}\right). \tag{5.15}$$

Solution of the Boltzmann equation yields

$$f_1 = -\tau \vec{v} \cdot \left(\frac{\partial f}{\partial \vec{r}}\right) = \tau \vec{v}\left(\frac{\partial f_0}{\partial E}\right)\left(\frac{E - E_F}{T}\right)\frac{\partial T}{\partial \vec{r}}. \tag{5.16}$$

Substitution of f_1 in the equation for the thermal current

$$\vec{U} = \frac{1}{4\pi^3}\int \vec{v}(E - F_F) f_1 d^3k \tag{5.17}$$

then results in

$$\vec{U} = \frac{1}{4\pi^3 T}\left(\frac{\partial T}{\partial \vec{r}}\right) \cdot \int \tau \vec{v}\vec{v}(E - E_F)^2 \left(\frac{\partial f_0}{\partial E}\right) d^3k. \tag{5.18}$$

Using the definition of the thermal conductivity tensor $\overleftrightarrow{\kappa}_e$ given by Eq. 5.5 we write the electronic contribution to the thermal conductivity $\overleftrightarrow{\kappa}_e$ as

$$\overleftrightarrow{\kappa}_e = \frac{-1}{4\pi^3 T} \int \tau \vec{v}\vec{v}(E - E_F)^2 \left(\frac{\partial f_0}{\partial E}\right) d^3k \tag{5.19}$$

where $d^3k = d^2S\ dk_\perp = d^2S\ dE/|\partial E/\partial \vec{k}| = d^2S dE/(\hbar v)$ is used to exploit our knowledge of the dependence of the distribution function on the energy as discussed below.

5.2.2 Thermal Conductivity for Metals

In the case of a metal, the integral for $\overleftrightarrow{\kappa}_e$ given by Eq. 5.19 can be evaluated easily by converting the integral over phase space $\int d^3k$ to an integral over $\int dE\ d^2S_F$ in order to exploit the δ-function property of $-(\partial f_0/\partial E)$. We then make use of the following result that you will show for homework and can be found in any standard statistical mechanics text (see for example, Reif)

$$\int G(E) \left(-\frac{\partial f_0}{\partial E}\right) dE = G(E_F) + \frac{\pi^2}{6}(k_B T)^2 \left[\frac{\partial^2 G}{\partial E^2}\right]_{E_F} + \ldots \tag{5.20}$$

It is necessary to consider the expansion given in Eq. 5.20 in solving Eq. 5.19 since $G(E_F)$ vanishes at $E = E_F$ for the integral defined in Eq. 5.19 for $\overleftrightarrow{\kappa}_e$. To solve the integral equation of Eq. 5.19 we make the identification of

$$G(E) = g(E)(E - E_F)^2 \tag{5.21}$$

where

$$g(E) = \frac{1}{4\pi^3} \int \tau \vec{v}\vec{v} d^2S/\hbar v \tag{5.22}$$

so that $G(E_F) = 0$ and $(\partial G/\partial E)|_{E_F} = 0$ while

$$\left[\frac{\partial^2 G}{\partial E^2}\right]_{E_F} = G''(E_F) = 2g(E_F). \tag{5.23}$$

These relations will be used again in connection with the calculation of the thermopower in §5.3.1. For the case of the thermal conductivity for a metal we then obtain

$$\overleftrightarrow{\kappa}_e = \frac{\pi^2}{3}(k_B T)^2 g(E_F) = \frac{(k_B T)^2}{12\pi\hbar} \int \tau \vec{v}\vec{v} \frac{d^2S_F}{v} \tag{5.24}$$

where the integration is over the Fermi surface. We immediately recognize that the integral appearing in Eq. 5.24 is the same as that for the electrical conductivity (see Eqs. 4.26 and 4.29)

$$\overleftrightarrow{\sigma} = \frac{e^2}{4\pi^3\hbar} \int \tau \vec{v}\vec{v} \frac{d^2S_F}{v} \tag{5.25}$$

so that the electronic contribution to the thermal conductivity and the electrical conductivity tensors are proportional to each other

$$\overleftrightarrow{\kappa}_e = \overleftrightarrow{\sigma}\ T \left(\frac{\pi^2 k_B^2}{3e^2}\right) \tag{5.26}$$

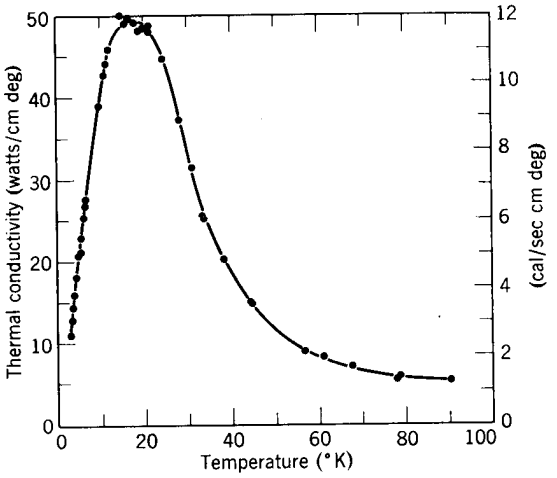


Figure 5.1: The temperature dependence of the thermal conductivity of copper. Note that both κ and T are plotted on linear scales. At low temperatures where the phonon density is low, the thermal transport is by electrons predominantly, while at high temperatures, thermal transport by phonons becomes more important.

and Eq. 5.26 is known as the Wiedemann–Franz Law. The physical basis for this relation is that in electrical conduction each electron carries a charge e and experiences an electrical force $e\vec{E}$ so that the electrical current per unit field is e^2. In thermal conduction, each electron carries a unit of thermal energy k_BT and experiences a thermal force $k_B\partial T/\partial\vec{r}$ so that the heat current per unit thermal gradient is proportional to k_B^2T. Therefore the ratio of $|\kappa_e|/|\sigma|$ must be on the order of (k_B^2T/e^2). The Wiedemann–Franz law suggests that the ratio $\kappa_e/(\sigma T)$ should be a constant (called the Lorenz constant)

$$\left|\frac{\kappa_e}{\sigma T}\right| = \frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^2 = 2.45\times10^{-8} \quad \text{watt ohm/deg}^2. \tag{5.27}$$

The ratio $(\kappa_e/\sigma T)$ is approximately constant for all metals at high temperatures $T > \Theta_D$ and at very low temperatures $T \ll \Theta_D$, where Θ_D is the Debye temperature. The derivation of the Wiedemann–Franz Law depends on the relaxation time approximation, which is valid at high temperatures $T > \Theta_D$ where the electron scattering is dominated by the quasi-elastic phonon scattering process and is valid also at very low temperatures $T \ll \Theta_D$ where phonon scattering is unimportant and the dominant scattering mechanism is impurity and lattice defect scattering, both of which tend to be elastic scattering processes. These scattering processes are discussed in Chapter 6 where we discuss in more detail the temperature dependence for κ.

The temperature dependence of the thermal conductivity of a metal is given in Fig. 5.1. From Eq. 5.27 we can write the following relation for the electronic contribution to the thermal conductivity κ_e when the Wiedemann–Franz law is satisfied

$$\kappa_e = \left(\frac{ne^2\tau}{m^*}\right)T\frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^2. \tag{5.28}$$

At very low temperatures where scattering by impurities, defects, crystal boundaries is dominant, σ is independent of T and therefore from the Wiedemann–Franz law, $\kappa_e \sim T$. At somewhat higher temperatures, but still in the regime $t \ll \Theta_D$, electron-phonon scattering starts to dominate. In this regime, the electrical conductivity exhibits a T^{-5} dependence. However only small q phonons participate in this regime. Thus it is only the phonon density which increases as T^3 that is relevant to the phonon-electron scattering, thereby yielding a resistivity with a T^3 dependence and a conductivity with a T^{-3} dependence. Using Eq. 5.28, we thus find that in the low T range where only low q phonons participate in thermal transport κ_e should show a T^{-2} dependence, in agreement with Fig. 5.1. At high T where all the phonons contribute to thermal transport we have $\sigma \sim 1/T$ so that κ_e becomes independent of T. Since $\Theta_D \sim 300$ K for Cu, this temperature range far exceeds the upper limit of Fig. 5.1.

5.2.3 Thermal Conductivity for Semiconductors

For the case of non-degenerate semiconductors, the integral for $\overleftrightarrow{\kappa}_e$ in Eq. 5.19 is evaluated by replacing $(E - E_F) \rightarrow E$, since in a semiconductor the electrons that can conduct heat must be in the conduction band, and the lowest energy an electron can have in the thermal conduction process is at the conduction band minimum. Then the thermal conductivity for a non-degenerate semiconductor can be written as

$$\overleftrightarrow{\kappa}_e = \frac{1}{4\pi^3 T} \int \tau \vec{v}\vec{v}\, E^2 \Big(-\frac{\partial f_0}{\partial E} \Big) d^3k. \tag{5.29}$$

For intrinsic semiconductors, the distribution function can be approximated by the Maxwell–Boltzmann distribution so that

$$(\partial f_0/\partial E) \rightarrow -(1/k_B T) e^{(-|E_F^e|/k_B T)} e^{(-E/k_B T)}. \tag{5.30}$$

For a parabolic band we have $E = \hbar^2 k^2/2m^*$, so that the volume element in reciprocal space can be written

$$\int d^3k = \int 4\pi k^2 dk = \int_0^\infty 2\pi (2m^*/\hbar^2)^{3/2} E^{1/2} dE, \tag{5.31}$$

and $\vec{v} = (1/\hbar)(\partial E/\partial \vec{k}) = \hbar \vec{k}/m$. Assuming a constant relaxation time, we then substitute all these terms into Eq. 5.29 for $\overleftrightarrow{\kappa}_e$ and integrate to obtain

$$\begin{aligned} \kappa_{exx} &= (1/4\pi^3 T) \int \tau v_x^2 E^2 \Big[(k_B T)^{-1} e^{-|E_F^e|(/k_B T)}\, e^{-E/(k_B T)} \Big] 2\pi (2m^*/\hbar^2)^{3/2} E^{1/2} dE \\ &= \Big[k_B (k_B T) \tau / (3\pi^2 m^*) \Big] (2m^* k_B T/\hbar^2)^{3/2}\, e^{-|E_F^e|/(k_B T)} \int_0^\infty x^{7/2} e^{-x} dx \end{aligned} \tag{5.32}$$

where $\int_0^\infty x^{7/2} e^{-x} dx = 105\sqrt{\pi}/8$, from which it follows that κ_{exx} has a temperature dependence of the form

$$T^{5/2} e^{-|E_F^e|/(k_B T)} \tag{5.33}$$

in which the exponential term is dominant for temperatures of physical interest, where $k_B T \ll |E_F^e|$. We note from Eq. 4.42 that for a semiconductor the temperature dependence

of the electrical conductivity is given by

$$\sigma_{xx} = \frac{2e^2\tau}{m^*}\left(\frac{m^* k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-|E_F^e|/(k_B T)}. \tag{5.34}$$

Assuming cubic symmetry, we can write the conductivity tensor as

$$\overleftrightarrow{\sigma} = \begin{pmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{xx} & 0 \\ 0 & 0 & \sigma_{xx} \end{pmatrix} \tag{5.35}$$

so that the electronic contribution to the thermal conductivity of a semiconductor can be written as

$$\kappa_{exx} = \left(\frac{35}{2}\right)\left(\frac{k_B^2}{e^2}\right)\sigma_{xx} T \tag{5.36}$$

where

$$\sigma_{xx} = ne^2\tau/m_{xx} = ne\mu_{xx} \tag{5.37}$$

and we note that the coefficient (35/2) for this calculation for semiconductors corresponds to $(\pi^2/3)$ for metals (see Eq. 5.27). Except for numerical constants, the formal results relating the electronic contribution to the thermal conductivity κ_{exx} and σ_{xx} are similar for metals and semiconductors, with the electronic thermal conductivity and electrical conductivity being proportional.

A major difference between semiconductors and metals is the magnitude of the electrical conductivity and hence of the electronic contribution to the thermal conductivity. Since σ_{xx} is much smaller for semiconductors than for metals, κ_e for semiconductors is relatively unimportant and the thermal conductivity is dominated by the lattice contribution κ_L.

5.2.4 Thermal Conductivity for Insulators

In the case of insulators, heat is only carried by phonons (lattice vibrations). The thermal conductivity in insulators therefore depends on phonon scattering mechanisms (see Chapter 6). The lattice thermal conductivity is calculated from kinetic theory and is given by

$$\kappa_L = \frac{C_p \overline{v}_q \Lambda_{\text{ph}}}{3} \tag{5.38}$$

where C_p is the heat capacity, $\overline{v}_q$ is the average phonon velocity and Λ_{ph} is the phonon mean free path. As discussed above, the total thermal conductivity of a solid is given as the sum of the lattice contribution κ_L and the electronic contribution κ_e. For metals the electronic contribution dominates, while for insulators and semiconductors the phonon contribution dominates. Let us now consider the temperature dependence of κ_{exx} (see Fig. 5.2). At very low T in the defect scattering range, the heat capacity has a dependence $C_p \propto T^3$ while $\overline{v}_q$ and Λ_{ph} are almost independent of T. As T increases and we enter the phonon-phonon scattering regime due to normal scattering processes and involving only low q phonons, C_p is still increasing with T but more slowly that T^3 while $\overline{v}_q$ remains independent of T and Λ_{ph}. As T increases further, the thermal conductivity increases more and more gradually and eventually starts to decrease because of phonon-phonon events where the density of phonons available for scattering depends on $[\exp(\hbar\omega/k_B T) - 1]$. This causes a peak in

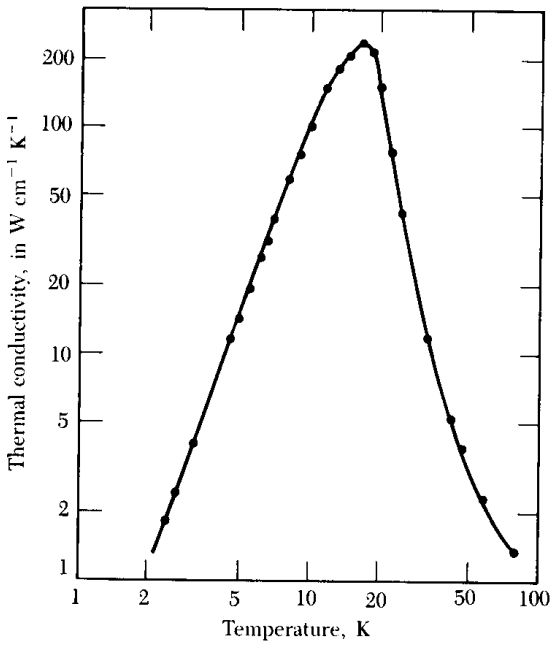


Figure 5.2: Temperature dependence of the thermal conductivity of a highly purified insulating crystal of NaF. Note that both κ and T are plotted on a log scale, and that the peak in κ occurs at low temperature (~ 17 K). The temperature dependence of κ is further discussed in §6.4.

$\kappa_L(T)$. The decrease in $\kappa_L(T)$ becomes more pronounced as C_p becomes independent of T and $\Lambda_{\rm ph}$ continues to be proportional to $[\exp(\hbar\bar{\omega}/k_BT) - 1]$ where $\bar{\omega}$ is a typical phonon frequency (see §6.4.1). As T increases further we eventually enter the $T \gg \Theta_D$ regime where Θ_D is the Debye temperature. In this regime, the temperature dependence of $\Lambda_{\rm ph}$ simply becomes $\Lambda_{\rm ph} \sim (1/T)$. Referring to Fig. 5.2 for $\kappa(T)$ for NaF we see that the peak in κ occurs at about 18 K where the complete Bose factor $[\exp(\hbar\bar{\omega}/k_BT) - 1]$ must be used to describe the T dependence of κ_L. For much of the temperature range in Fig. 5.2, only low q phonons participate in the thermal conduction process. At higher temperatures where larger q phonons contribute to thermal conduction, umklapp processes become important in the phonon scattering process, as discussed in §6.3.1.

5.3 Thermoelectric Phenomena

In many metals and semiconductors there exists a coupling between the electrical current and the thermal current. This coupling can be appreciated by observing that when electrons carry thermal current, they are also transporting charge and therefore generating electric fields. This coupling between the charge transport and heat transport gives rise to thermoelectric phenomena. In our discussion of thermoelectric phenomena we start with a general derivation of the coupled equations for the electrical current density $\vec{j}$ and the

thermal current density $\vec{U}$:

$$\vec{j} = \frac{e}{4\pi^3}\int \vec{v} f_1 d^3k \tag{5.39}$$

$$\vec{U} = \frac{1}{4\pi^3}\int \vec{v}(E - E_F) f_1 d^3k \tag{5.40}$$

and the perturbation to the distribution function f_1 is found from solution of Boltzmann's equation in the relaxation time approximation:

$$\vec{v}\cdot\frac{\partial f}{\partial \vec{r}} + \dot{\vec{k}}\cdot\frac{\partial f}{\partial \vec{k}} = -\frac{(f - f_0)}{\tau}, \tag{5.41}$$

which is written here for the case of time independent forces and fields. Substituting for $(\partial f/\partial \vec{r})$ from Eqs. 5.8 and 5.15, for $\partial f/\partial \vec{k}$ from Eq. 4.17, and for $\partial f/\partial \vec{k} = (\partial f_0/\partial E)(\partial E/\partial \vec{k})$ yields

$$\vec{v}\cdot(\partial f_0/\partial E)\Big[(e\vec{E} - \vec{\nabla}E_F) - \frac{(E - E_F)}{T}(\vec{\nabla}T)\Big] = -\frac{f_1}{\tau}, \tag{5.42}$$

so that the solution to the Boltzmann equation in the presence of an electric field and a temperature gradient is

$$f_1 = \vec{v}\tau\cdot(\partial f_0/\partial E)\Big\{[(E - E_F)/T]\vec{\nabla}T - e\vec{E} + \vec{\nabla}E_F\Big\}. \tag{5.43}$$

The electrical and thermal currents in the presence of both an applied electric field and a temperature gradient can thus be obtained by substituting f_1 into $\vec{j}$ and $\vec{U}$ in Eqs. 5.39 and 5.40 to yield expressions of the form:

$$\vec{j} = e^2 \overleftrightarrow{\kappa}_0 \cdot(\vec{E} - \frac{1}{e}\vec{\nabla}E_F) - (e/T)\overleftrightarrow{\kappa}_1 \cdot\vec{\nabla}T \tag{5.44}$$

and

$$\vec{U} = e\overleftrightarrow{\kappa}_1 \cdot(\vec{E} - \frac{1}{e}\vec{\nabla}E_F) - (1/T)\overleftrightarrow{\kappa}_2 \cdot\vec{\nabla}T \tag{5.45}$$

where $\overleftrightarrow{\kappa}_0$ is related to the conductivity tensor $\overleftrightarrow{\sigma}$ by

$$\overleftrightarrow{\kappa}_0 = \frac{1}{4\pi^3}\int \tau\vec{v}\vec{v}(-\partial f_0/\partial E)d^3k = \frac{\overleftrightarrow{\sigma}}{e^2}, \tag{5.46}$$

and

$$\overleftrightarrow{\kappa}_1 = \frac{1}{4\pi^3}\int \tau\vec{v}\vec{v}(E - E_F)\Big(-\frac{\partial f_0}{\partial E}\Big)d^3k, \tag{5.47}$$

and $\overleftrightarrow{\kappa}_2$ is related to the thermal conductivity tensor $\overleftrightarrow{\kappa}_e$ by

$$\overleftrightarrow{\kappa}_2 = \frac{1}{4\pi^3}\int \tau\vec{v}\vec{v}(E - E_F)^2(-\partial f_0/\partial E)d^3k = T\overleftrightarrow{\kappa}_e\ . \tag{5.48}$$

Note that the integrands for $\overleftrightarrow{\kappa}_1$ and $\overleftrightarrow{\kappa}_2$ are both related to that for $\overleftrightarrow{\kappa}_0$ by introducing factors of $(E - E_F)$ and $(E - E_F)^2$, respectively. Note also that the same integral $\overleftrightarrow{\kappa}_1$ occurs in the expression for the electric current $\vec{j}$ induced by a thermal gradient $\vec{\nabla}T$ and in the expression for the thermal current $\vec{U}$ induced by an electric field $\vec{E}$. The motion of charged

carriers across a temperature gradient results in a flow of electric current expressed by the term $-(e/T)(\overleftrightarrow{\kappa}_1)\cdot\vec{\nabla}T$. This term is the origin of thermoelectric effects.

The discussion up to this point has been general. If specific boundary conditions are imposed, we obtain a variety of thermoelectric effects such as the Seebeck effect, the Peltier effect and the Thomson effect. We now define the conditions under which each of these thermoelectric effects occur.

We define the thermopower $\overleftrightarrow{\mathcal{S}}$ (Seebeck coefficient) and the Thomson coefficient $\mathcal{T}_b$ under conditions of zero current flow. Then referring to Eq. 5.44, we obtain under open circuit conditions

$$\vec{j}=0=e^2\overleftrightarrow{\kappa}_0\cdot(\vec{E}-\frac{1}{e}\vec{\nabla}E_F)-(e/T)\overleftrightarrow{\kappa}_1\cdot\vec{\nabla}T \tag{5.49}$$

so that the Seebeck coefficient $\overleftrightarrow{\mathcal{S}}$ is defined by

$$\vec{E}-\frac{1}{e}\vec{\nabla}E_F=(1/eT)\overleftrightarrow{\kappa}_0^{-1}\cdot\overleftrightarrow{\kappa}_1\cdot\vec{\nabla}T\equiv\overleftrightarrow{\mathcal{S}}\cdot\vec{\nabla}T, \tag{5.50}$$

and $\mathcal{S}$ is sometimes called the thermopower. Using the relation $\vec{\nabla}E_F=\frac{\partial E_F}{\partial T}\vec{\nabla}T$ we obtain the definition for the Thomson coefficient $\mathcal{T}_b$

$$\vec{E}=\left(\frac{1}{e}\frac{\partial E_F}{\partial T}+\overleftrightarrow{\mathcal{S}}\right)\vec{\nabla}T\equiv\overleftrightarrow{\mathcal{T}}_b\cdot\vec{\nabla}T \tag{5.51}$$

where

$$\overleftrightarrow{\mathcal{T}}_b=T\frac{\partial}{\partial T}\overleftrightarrow{\mathcal{S}}\,. \tag{5.52}$$

For many thermoelectric systems of interest, $\overleftrightarrow{\mathcal{S}}$ has a linear temperature dependence, and in this case it follows from Eq. 5.52 that $\overleftrightarrow{\mathcal{T}}_b$ and $\overleftrightarrow{\mathcal{S}}$ are almost equivalent for practical purposes. Therefore the Seebeck and Thomson coefficients are used almost interchangeably in the literature.

From Eq. 5.50 we have

$$\overleftrightarrow{\mathcal{S}}=(1/eT)\overleftrightarrow{\kappa}_0^{-1}\cdot\overleftrightarrow{\kappa}_1 \tag{5.53}$$

which is simplified by assuming an isotropic medium, yielding the scalar quantities

$$\mathcal{S}=(1/eT)(\kappa_1/\kappa_0). \tag{5.54}$$

However in an anisotropic medium the tensor components of $\overleftrightarrow{\mathcal{S}}$ are found from

$$\mathcal{S}_{ij}=(1/eT)(\kappa_0^{-1})_{i\alpha}(\kappa_1)_{\alpha j}, \tag{5.55}$$

where the Einstein summation convention is assumed. Figure 5.3 shows a schematic diagram for measuring the thermopower or Seebeck effect in an n-type semiconductor. At the hot junction the Fermi level is higher than at the cold junction. Electrons will move from the hot junction to the cold junction in an attempt to equalize the Fermi level, thereby creating an electric field which can be measured in terms of the open circuit voltage V shown in Fig. 5.3.

Another important thermoelectric coefficient is the Peltier coefficient $\overleftrightarrow{\Pi}$, defined as the proportionality between $\vec{U}$ and $\vec{j}$

$$\vec{U}\equiv\overleftrightarrow{\Pi}\cdot\vec{j} \tag{5.56}$$

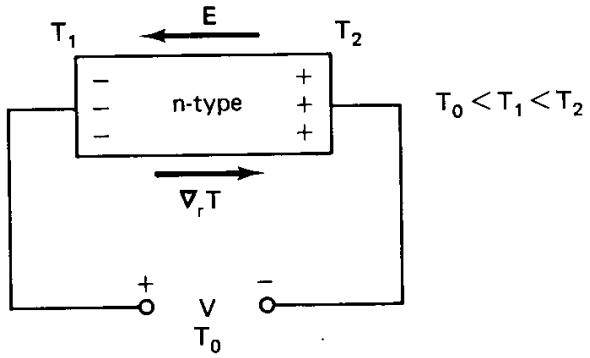


Figure 5.3: Determination of the Seebeck effect for an n-type semiconductor. In the presence of a temperature gradient electrons, will move from the hot junction to the cold junction, thereby creating an electric field and a voltage V across the semiconductor.

in the absence of a thermal gradient. For $\vec{\nabla}T = 0$, Eqs. 5.44 and 5.45 become

$$\vec{j} = e^2 \overleftrightarrow{\kappa}_0 \cdot \left(\vec{E} - \frac{1}{e} \vec{\nabla} E_F \right) \tag{5.57}$$

$$\vec{U} = e \overleftrightarrow{\kappa}_1 \cdot \left(\vec{E} - \frac{1}{e} \vec{\nabla} E_F \right) \tag{5.58}$$

so that

$$\begin{aligned} \vec{U} = & \ (1/e) \overleftrightarrow{\kappa}_1 \cdot (\overleftrightarrow{\kappa}_0)^{-1} \cdot \vec{j} \\ = & \ \overleftrightarrow{\Pi} \cdot \vec{j} \end{aligned} \tag{5.59}$$

where

$$\overleftrightarrow{\Pi} = (1/e) \overleftrightarrow{\kappa}_1 \cdot (\overleftrightarrow{\kappa}_0)^{-1}. \tag{5.60}$$

Comparing Eqs. 5.53 and 5.60 we see that $\overleftrightarrow{\Pi}$ and $\overleftrightarrow{\mathcal{S}}$ are related by

$$\overleftrightarrow{\Pi} = T \overleftrightarrow{\mathcal{S}}, \tag{5.61}$$

where T is the temperature. For isotropic materials the Peltier coefficient thus becomes a scalar, and is proportional to the thermopower $\mathcal{S}$:

$$\Pi = \frac{1}{e}(\kappa_1/\kappa_0) = T\mathcal{S}, \tag{5.62}$$

while for anisotropic materials the tensor components of $\overleftrightarrow{\Pi}$ can be found in analogy with Eq. 5.55. We note that both $\overleftrightarrow{\mathcal{S}}$ and $\overleftrightarrow{\Pi}$ exhibit a linear dependence on e and therefore depend explicitly on the **sign of the carrier**, and measurements of $\overleftrightarrow{\mathcal{S}}$ or $\overleftrightarrow{\Pi}$ can be used to determine whether transport is dominated by electrons or holes.

We have already considered the evaluation of $\overleftrightarrow{\kappa}_0$ in treating the electrical conductivity and $\overleftrightarrow{\kappa}_2$ in treating the thermal conductivity. To treat thermoelectric phenomena we need now to evaluate $\overleftrightarrow{\kappa}_1$

$$\overleftrightarrow{\kappa}_1 = \frac{1}{4\pi^3} \int \tau \vec{v}\vec{v}(E - E_F)(-\partial f_0/\partial E) d^3k. \tag{5.63}$$

In §5.3.1 we evaluate $\overleftrightarrow{\kappa}_1$ for the case of a metal and in §5.3.2 we evaluate $\overleftrightarrow{\kappa}_1$ for the case of the electrons in an intrinsic semiconductor.

5.3.1 Thermoelectric Phenomena in Metals

All thermoelectric effects in metals depend on the tensor $\overleftrightarrow{\kappa}_1$ which we evaluate below for the case of a metal. We can then obtain the thermopower

$$\overleftrightarrow{\mathcal{S}} = \frac{1}{eT}(\overleftrightarrow{\kappa}_1 \cdot \overleftrightarrow{\kappa}_0^{-1}) \tag{5.64}$$

or the Peltier coefficient

$$\overleftrightarrow{\Pi} = \frac{1}{e}(\overleftrightarrow{\kappa}_1 \cdot \overleftrightarrow{\kappa}_0^{-1}) \tag{5.65}$$

or the Thomson coefficient

$$\overleftrightarrow{\mathcal{T}}_b = T\frac{\partial}{\partial T}\overleftrightarrow{\mathcal{S}}\,. \tag{5.66}$$

To evaluate $\overleftrightarrow{\kappa}_1$ for metals we wish to exploit the δ-function behavior of $(-\partial f_0/\partial E)$. This is accomplished by converting the integration d^3k to an integration over dE and a constant energy surface, $d^3k = d^2S\; dE/\hbar v$. From Fermi statistics we have the general relation (see Eq. 5.20)

$$\int G(E)\Big(-\frac{\partial f_0}{\partial E}\Big)dE = G(E_F) + \frac{\pi^2}{6}(k_BT)^2\left[\frac{\partial^2 G}{\partial E^2}\right]_{E_F} + \cdots. \tag{5.67}$$

For the integral in Eq. 5.63 which defines $\overleftrightarrow{\kappa}_1$, we can write

$$G(E) = g(E)(E - E_F) \tag{5.68}$$

where

$$g(E) = \frac{1}{4\pi^3}\int \tau\vec{v}\vec{v}\; d^2S/v \tag{5.69}$$

and the integration in Eq. 5.69 is carried out over a constant energy surface at energy E. Differentiation of $G(E)$ then yields

$$\begin{aligned} G'(E) &= g'(E)(E - E_F) + g(E) \\ G''(E) &= g''(E)(E - E_F) + 2g'(E). \end{aligned} \tag{5.70}$$

Evaluation at $E = E_F$ yields

$$\begin{aligned} G(E_F) &= 0 \\ G''(E_F) &= 2g'(E_F). \end{aligned} \tag{5.71}$$

We therefore obtain

$$\overleftrightarrow{\kappa}_1 = \frac{\pi^2}{3}(k_BT)^2 g'(E_F). \tag{5.72}$$

We interpret $g'(E_F)$ in Eq. 5.72 to mean that the same integral $\overleftrightarrow{\kappa}_0$ which determines the conductivity tensor is evaluated on a constant energy surface E, and $g'(E_F)$ is the energy derivative of that integral evaluated at the Fermi energy E_F. The temperature dependence of $g'(E_F)$ is related to the temperature dependence of τ, since v is essentially temperature independent. For example, we will see in Chapter 6 that acoustic phonon scattering in the high temperature limit $T \gg \Theta_D$ yields a temperature dependence $\tau \sim T^{-1}$ so that $\overleftrightarrow{\kappa}_1$ in this important case for metals will be proportional to T.

For a spherical constant energy surface $E = \hbar^2 k^2/2m^*$ and assuming a relaxation time τ that is independent of energy, we can readily evaluate Eq. 5.72 to obtain

$$g(E) = \frac{\tau}{3\pi^2 m^*}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} E^{3/2} \tag{5.73}$$

$$g'(E_F) = \frac{\tau}{2\pi^2 m^*}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} E_F^{1/2} \tag{5.74}$$

and

$$\kappa_1 = \frac{\tau}{6m^*}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} E_F^{1/2}(k_B T)^2. \tag{5.75}$$

Using the same approximations, we can write for κ_0:

$$\kappa_0 = \frac{\tau}{3\pi^2 m^*}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} E_F^{3/2} \tag{5.76}$$

so that from Eq. 5.64 we have for the Seebeck coefficient

$$\mathcal{S} = \frac{\pi^2 k_B}{2e}\,\frac{k_B T}{E_F}. \tag{5.77}$$

From Eq. 5.77 we see that $\mathcal{S}$ exhibits a linear dependence on T and a sensitivity to the sign of the carriers. We note from Eq. 5.64 that a low carrier density implies a large $\mathcal{S}$ value. Thus degenerate (heavily doped with $n \sim 10^{18}$–$10^{19}/\mathrm{cm}^3$) semiconductors tend to have higher thermopowers than metals.

5.3.2 Thermopower for Semiconductors

In this section we evaluate $\overleftrightarrow{\kappa}_1$ for electrons in an intrinsic or lightly doped semiconductor for illustrative purposes. Intrinsic semiconductors are not important for practical thermoelectric devices since the contributions of electrons and holes to $\overleftrightarrow{\kappa}_1$ are of opposite signs and tend to cancel. Thus it is only heavily doped semiconductors with a single carrier type that are important for thermoelectric applications.

The evaluation of the general expression for the integral $\overleftrightarrow{\kappa}_1$

$$\overleftrightarrow{\kappa}_1 = \frac{1}{4\pi^3}\int \tau \vec{v}\vec{v}(E - E_F)\left(-\frac{\partial f_0}{\partial E}\right)d^3k \tag{5.78}$$

is different for semiconductors and metals. Referring to Fig. 5.4 for an intrinsic semiconductor we need to make the substitution $(E - E_F) \to E$ in Eq. 5.78, since only conduction electrons can carry heat. The equilibrium distribution function for an intrinsic semiconductor can be written as

$$f_0 = e^{-E/(k_B T)}\; e^{-|E_F^e|/(k_B T)} \tag{5.79}$$

so that

$$\frac{\partial f_0}{\partial E} = -\frac{1}{k_B T}\; e^{-E/(k_B T)}\; e^{-|E_F^e|/(k_B T)}. \tag{5.80}$$

To evaluate d^3k we need to assume a model for $E(\vec{k})$. For simplicity, assume a simple parabolic band

$$E = \hbar^2 k^2/2m^* \tag{5.81}$$

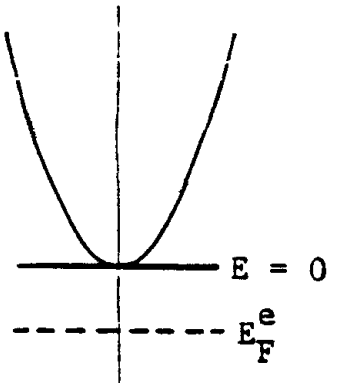


Figure 5.4: Schematic E vs k diagram, showing that $E = 0$ is the lowest electronic energy for heat conduction.

$$d^3k = 4\pi k^2 dk \tag{5.82}$$

so that

$$d^3k = 2\pi\left(\frac{2m^*}{\hbar^2}\right)^{3/2} E^{1/2} dE \tag{5.83}$$

and also

$$\vec{v} = \frac{1}{\hbar}(\partial E/\partial \vec{k}) = \hbar\vec{k}/m^*. \tag{5.84}$$

Substitution into the equation for $\overleftrightarrow{\kappa}_1$ for a semiconductor with the simple $E = \hbar^2k^2/2m^*$ dispersion relation then yields upon integration

$$\kappa_{1xx} = \frac{5\tau k_B T}{m^*}(m^* k_B T/2\pi\hbar^2)^{3/2} e^{-|E_F^e|/(k_B T)}. \tag{5.85}$$

This expression is valid for a semiconductor for which the Fermi level is far from the band edge $(E - E_F) \gg k_B T$. The thermopower is then found by substitution

$$\mathcal{S} = \frac{1}{eT}\,(\kappa_{1xx}/\kappa_{0xx}) \tag{5.86}$$

where the expression for $\kappa_{0xx} = \sigma_{xx}/e^2$ is given by Eq. 5.46. We thus obtain the result

$$\mathcal{S} = \frac{5}{2}\,\frac{k_B}{e} \tag{5.87}$$

which is a constant independent of temperature, independent of the band structure, but sensitive to the sign of the carriers. The calculation in this section was for the contribution of electrons. In an actual intrinsic semiconductor, the contribution of both electrons and holes to κ_1 must be found. Likewise the calculation for κ_{0xx} would also include contributions from both electrons and holes. Since the contribution to $(1/e)\kappa_{1xx}$ for holes and electrons are of opposite sign, we can from Eq. 5.87 expect that $\mathcal{S}$ for holes will cancel $\mathcal{S}$ for electrons for an intrinsic semiconductor.

Materials with a high thermopower or Seebeck coefficient are heavily doped degenerate semiconductors for which the Fermi level is close to the band edge and the complete Fermi function must be used. Since $\mathcal{S}$ depends on the sign of the charge carriers, thermoelectric materials are doped either heavily doped n-type or heavily doped p-type semiconductors to prevent cancellation of the contribution from electrons and holes, as occurs in intrinsic semiconductors which because of thermal excitations of carriers have equal concentrations of electrons and holes.

5.3.3 Effect of Thermoelectricity on the Thermal Conductivity

From the coupled equations given by Eqs. 5.44 and 5.45 it is seen that the proportionality between the thermal current $\vec{U}$ and the temperature gradient $\vec{\nabla}T$ in the absence of electrical current ($\vec{j}=0$) contains terms related to $\overleftrightarrow{\kappa}_1$. We now solve Eqs. 5.44 and 5.45 to find the contribution of the thermoelectric terms to the electronic thermal conductivity. When $\vec{j}=0$, Eq. 5.44 becomes

$$(\vec{E}-\frac{1}{e}\vec{\nabla}E_F)=\frac{1}{eT}\,\overleftrightarrow{\kappa^{-1}}_0\cdot\overleftrightarrow{\kappa}_1\cdot\vec{\nabla}T \tag{5.88}$$

so that

$$\vec{U}=-(1/T)\left[\overleftrightarrow{\kappa}_2-\overleftrightarrow{\kappa}_1\cdot\overleftrightarrow{\kappa^{-1}}_0\cdot\overleftrightarrow{\kappa}_1\right]\cdot\vec{\nabla}T \tag{5.89}$$

where $\overleftrightarrow{\kappa}_0$, $\overleftrightarrow{\kappa}_1$, and $\overleftrightarrow{\kappa}_2$ are given by Eqs. 5.46, 5.47, and 5.48, respectively, or

$$\overleftrightarrow{\kappa}_0=\frac{1}{4\pi^3\hbar}\int\tau\vec{v}\vec{v}\frac{d^2S_F}{v}, \tag{5.90}$$

$$\overleftrightarrow{\kappa}_1=\frac{\pi^2}{3}(k_BT)^2\left(\frac{\partial\,\overleftrightarrow{\kappa}_0}{\partial E}\right)_{E_F}, \tag{5.91}$$

and

$$\overleftrightarrow{\kappa}_2=\frac{(k_BT)^2}{12\pi\hbar}\int\tau\vec{v}\vec{v}\frac{d^2S_F}{v}. \tag{5.92}$$

We now evaluate the contribution to the thermal conductivity from the thermoelectric coupling effects for the case of a metal having a simple dispersion relation

$$E=\hbar^2k^2/2m^*. \tag{5.93}$$

In this case where τ is considered to be independent of E, Eqs. 5.91 and 5.90, respectively, provide expressions for κ_1 and κ_0 from which

$$\frac{1}{T}\kappa_1\kappa_0^{-1}\kappa_1=\frac{\pi^4n\tau}{4m^*}k_B^2T\left(\frac{k_BT}{E_F}\right)^2 \tag{5.94}$$

so that from Eqs. 5.28 and 5.94 the total electronic thermal conductivity for the metal becomes

$$\kappa_e=\frac{\pi^2n\tau}{3m^*}k_B^2T\left[1-\frac{3\pi^2}{4}\left(\frac{k_BT}{E_F}\right)^2\right]. \tag{5.95}$$

For typical metals $(T/T_F)\sim(1/30)$ at room temperature so that the thermoelectric correction term is less than 1%.

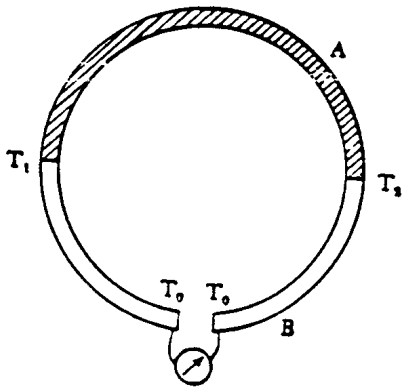


Figure 5.5: Thermopower between two different metals showing the principle of operation of a thermocouple under open circuit conditions (i.e., $j = 0$).

5.4 Thermoelectric Measurements

5.4.1 Seebeck Effect (Thermopower)

The thermopower $\mathcal{S}$ as defined in Eq. 5.50 and is the characteristic coefficient in the Seebeck effect, where a metal subjected to a thermal gradient $\vec{\nabla}T$ exhibits an electric field $\vec{E} = \mathcal{S}\vec{\nabla}T$, or an open-circuit voltage V under conditions of no current flow.

In the application of the Seebeck effect to thermocouple operation, we usually measure the difference in thermopower $\mathcal{S}_A - \mathcal{S}_B$ between two different metals A and B by measuring the open circuit voltage V_{AB} as shown in Fig. 5.5. This voltage can be calculated from

$$\begin{aligned} V_{AB} &= -\oint \vec{E} \cdot d\vec{r} = -\oint \mathcal{S} \frac{\partial T}{\partial \vec{r}}\, d\vec{r} \\ &= \int_{T_0}^{T_1} \mathcal{S}_B dT + \int_{T_1}^{T_2} \mathcal{S}_A dT + \int_{T_2}^{T_0} \mathcal{S}_B dT \\ &= \int_{T_1}^{T_2} (\mathcal{S}_A - \mathcal{S}_B) dT. \end{aligned} \tag{5.96}$$

With $T_1 \neq T_2$, an open-circuit potential difference V_{AB} can be measured and Eq. 5.96 shows that V_{AB} is independent of the temperature T_0. Thus if T_1 is known and V_{AB} is measured, then temperature T_2 can be found from the calibration table of the thermocouple. From the simple expression of Eq. 5.77

$$\mathcal{S} = \frac{\pi^2 k_B}{2e} \frac{k_B T}{E_F} \tag{5.97}$$

a linear dependence of $\mathcal{S}$ on T is predicted for simple metals. For actual thermocouples used for temperature measurements, the $\mathcal{S}(T)$ dependence is approximately linear, but is given by an accurate calibration table to account for small deviations from this linear relation. Thermocouples are calibrated at several fixed temperatures and the calibration table comes from a fit of these thermal data to a polynomial function that is approximately linear in T.

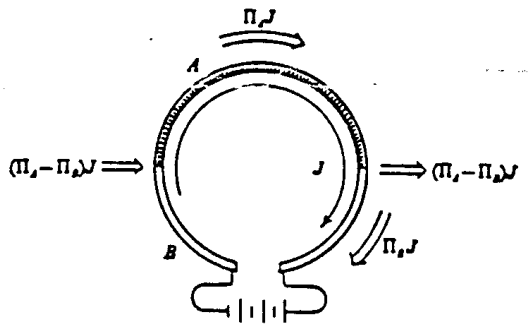


Figure 5.6: A heat engine based on the Peltier Effect with heat introduced at one junction and extracted at another under the conditions of no temperature gradient ($\nabla T = 0$).

5.4.2 Peltier Effect

The Peltier effect is the observation of a thermal current $\vec{U} = \overleftrightarrow{\Pi} \cdot \vec{j}$ in the presence of an electric current $\vec{j}$ with no thermal gradient ($\vec{\nabla} T = 0$) so that

$$\overleftrightarrow{\Pi} = T \overleftrightarrow{\mathcal{S}} \, . \tag{5.98}$$

The Peltier effect measures the heat generated (or absorbed) at the junction of two dissimilar metals held at constant temperature, when an electric current passes through the junction. Sending electric current around a circuit of two dissimilar metals cools one junction and heats another and is the basis of the operation of thermoelectric coolers. This thermoelectric effect is represented schematically in Fig. 5.6. Because of the similarities between the Peltier coefficient and the Seebeck coefficient, materials exhibiting a large Seebeck effect also show a large Peltier effect. Since both $\overleftrightarrow{\mathcal{S}}$ and $\overleftrightarrow{\Pi}$ are proportional to $(1/e)$, the sign of $\overleftrightarrow{\mathcal{S}}$ and $\overleftrightarrow{\Pi}$ is negative for electrons and positive for holes in the case of degenerate semiconductors. Reversing the direction of $\vec{j}$, will interchange the junctions where heat is generated (absorbed).

5.4.3 Thomson Effect

Assume that we have an electric circuit consisting of a single metal conductor. The power generated in a sample, such as an n-type semiconductor as shown in Fig. 5.7, is

$$P = \vec{j} \cdot \vec{E} \tag{5.99}$$

where the electric field can be obtained from Eqs. 5.49 and 5.51 as

$$\vec{E} = (\overleftrightarrow{\sigma^{-1}}) \cdot \vec{j} - \overleftrightarrow{\mathcal{T}}_b \cdot \vec{\nabla} T \tag{5.100}$$

where $\overleftrightarrow{\mathcal{T}}_b$ is the Thomson coefficient defined in Eq. 5.51 and is related to the Seebeck coefficient $\overleftrightarrow{\mathcal{S}}$ as discussed in §5.3 and §5.3.1. Substitution of Eq. 5.100 into Eq. 5.99 yields the total power dissipation

$$P = \vec{j} \cdot (\overleftrightarrow{\sigma^{-1}}) \cdot \vec{j} - \vec{j} \cdot \overleftrightarrow{\mathcal{T}}_b \cdot \vec{\nabla} T. \tag{5.101}$$

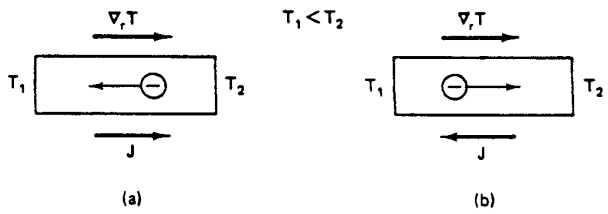


Figure 5.7: The Thomson term in an n-type semiconductor produces (a) heating when $\vec{j}$ and $\vec{\nabla}T$ are in the same direction and (b) cooling when $\vec{j}$ and $\vec{\nabla}T$ are in opposite directions.

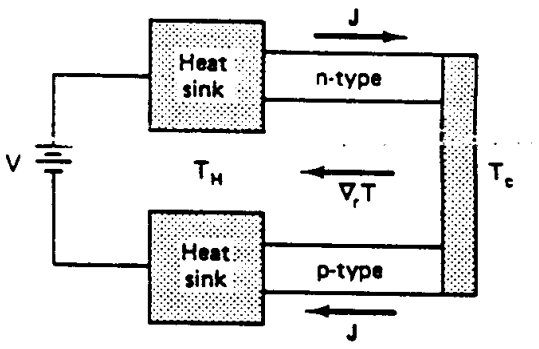


Figure 5.8: Schematic diagram of a thermoelectric cooler. The heat sinks and cold junctions are metals that form ohmic contacts to the active thermoelectric n-type and p-type semiconductors.

The first term in Eq. 5.101 is the conventional joule heating term while the second term is the contribution from the Thomson effect. For an n-type semiconductor $\overleftrightarrow{\mathcal{T}}_b$ is negative. Thus when $\vec{j}$ and $\vec{\nabla}T$ are parallel, heating will result, as in Fig. 5.7(a). However if $\vec{j}$ and $\vec{\nabla}T$ are antiparallel, as in Fig. 5.7(b), cooling will occur. Thus reversal of the direction of $\vec{j}$ without changing the direction of $\vec{\nabla}T$ will reverse the sign of the Thomson contribution. Likewise, a reversal in the direction of $\vec{\nabla}T$ keeping the direction of $\vec{j}$ unchanged will also reverse the sign of the Thomson contribution.

Thus, if either (but not both) the direction of the electric current or the direction of the thermal gradient is reversed, an absorption of heat from the surroundings will take place. The Thomson effect is utilized in thermoelectric refrigerators which are useful as practical low temperature laboratory coolers. Referring to Fig. 5.8, we see a schematic diagram explaining the operation of a thermoelectric cooler. We see that for a degenerate n-type semiconductor where $\overleftrightarrow{\mathcal{T}}_b$, $\overleftrightarrow{\mathcal{S}}$, and $\overleftrightarrow{\Pi}$ are all negative, when $\vec{j}$ and $\vec{\nabla}T$ are antiparallel then cooling occurs and heat is extracted from the cold junction and transferred to the heat sink at temperature T_H. For the p-type leg, all the thermoelectric coefficients are positive, so Eq. 5.101 shows that cooling occurs when $\vec{j}$ and $\vec{\nabla}T$ are parallel. Thus both the n-type and p-type legs in a thermoelectric element contribute to cooling in a thermoelectric cooler.

5.4.4 The Kelvin Relations

The three thermoelectric effects are related and these relations were first derived by Lord Kelvin after he became a Lord and changed his name from Thomson to Kelvin. The Kelvin relations are based on arguments of irreversible thermodynamics and relate Π, $\mathcal{S}$, and $\mathcal{T}_b$.

If we define the thermopower $\mathcal{S}_{AB} = \mathcal{S}_B - \mathcal{S}_A$ and the Peltier coefficient similarly $\Pi_{AB} = \Pi_A - \Pi_B$ for material A joined to material B, then we obtain the first Kelvin relation:

$$\mathcal{S}_{AB} = \frac{\Pi_{AB}}{T}. \tag{5.102}$$

The Thomson coefficient $\mathcal{T}_b$ is defined by

$$\mathcal{T}_b = T\frac{\partial \mathcal{S}}{\partial T} \tag{5.103}$$

which allows determination of the Seebeck coefficient at temperature T_0 by integration of the above equation

$$\mathcal{S}(T_0) = \int_0^{T_0} \left[\frac{\mathcal{T}_b(T)}{T}\right] dT. \tag{5.104}$$

Furthermore, from the above definitions, we deduce the second Kelvin relation

$$\mathcal{T}_{b,A} - \mathcal{T}_{b,B} = T\frac{\partial \mathcal{S}_{AB}}{\partial T} = T\frac{\partial \mathcal{S}_A}{\partial T} - T\frac{\partial \mathcal{S}_B}{\partial T} \tag{5.105}$$

from which we obtain an expression relating all three thermoelectric coefficients

$$\mathcal{T}_{b,A} = T\frac{\partial \mathcal{S}_A}{\partial T} = T\frac{\partial (\Pi_A/T)}{\partial T} = \frac{\partial \Pi_A}{\partial T} - \frac{\Pi_A}{T} = \frac{\partial \Pi_A}{\partial T} - \mathcal{S}_A. \tag{5.106}$$

5.4.5 Thermoelectric Figure of Merit

A good thermoelectric material for cooling applications must have a high thermoelectric figure of merit, Z, which is defined by

$$Z = \frac{\mathcal{S}^2\sigma}{\kappa} \tag{5.107}$$

where $\mathcal{S}$ is the thermoelectric power (Seebeck coefficient), σ is the electrical conductivity, and κ is the thermal conductivity. In order to achieve a high Z, one requires a high thermoelectric power $\mathcal{S}$, a high electrical conductivity σ to maintain high carrier mobility, and a low thermal conductivity κ to retain the applied thermal gradient. In general, it is difficult in practical systems to increase Z for the following reasons: increasing $\mathcal{S}$ for simple materials also leads to a simultaneous decrease in σ, and an increase in σ leads to a comparable increase in the electronic contribution to κ because of the Wiedemann–Franz law. So with known conventional solids, a limit is rapidly obtained where a modification to any one of the three parameters $\mathcal{S}$, σ, or κ adversely affects the other transport coefficients, so that the resulting Z does not vary significantly. Currently, the materials with the highest Z are Bi_2Te_3 alloys such as $Bi_{0.5}Sb_{1.5}Te_3$ with $ZT \sim 1$ at 300K.

Only small increases in Z have been achieved in the last two to three decades. Research on thermoelectric materials has therefore been at a low level since about 1960. Since 1994,

new interest has been revived in thermoelectricity with the discovery of new materials: skutterudites – $CeFe_{4-x}Co_xSb_{12}$ or $LaFe_{4-x}Co_xSb_{12}$ for $0 < x < 4$, which offer promise for higher Z values in bulk materials, and low dimensional systems (quantum wells, quantum wires) which offer promise for enhanced Z relative to bulk Z values in the same material. Thus thermoelectricity has again become an active research field.

5.5 Phonon Drag Effect

For a simple metal such as an alkali metal one would expect the thermopower $\mathcal{S}$ to be given by the simple expression in Eq. 5.77, and to be negative since the carriers are electrons. This is true at room temperature for all of the alkali metals except Li. Furthermore, $\mathcal{S}$ is positive for the noble metals Au, Ag and Cu. This can be understood by recalling the complex Fermi surfaces for these metals (see Fig. 2.6), where we note that copper in fact exhibits hole orbits in the extended zone. In general, with multiple carrier types as occur in semiconductors, the interpretation of thermopower data can become complicated.

Another complication which must also be considered, especially at low temperatures, is the *phonon drag effect*. In the presence of a thermal gradient, the phonons will diffuse and "drag" the electrons along with them because of the electron-phonon interaction. For a simple explanation of phonon drag, consider a gas of phonons with an average energy density $E_{\rm ph}/V$. Using kinetic theory, we find that the phonon gas exerts a pressure

$$P = \frac{1}{3}\left(\frac{E_{\rm ph}}{V}\right) \tag{5.108}$$

on the electron gas. In the presence of a thermal gradient, the electrons are subject to a force density

$$F_x/V = -dP/dx = -\frac{1}{3V}\left(\frac{dE_{\rm ph}}{dT}\right)\frac{dT}{dx}. \tag{5.109}$$

To prevent the flow of current, this force must be balanced by the electric force. Thus, for an electron density n, we obtain

$$-neE_x + F_x/V = 0 \tag{5.110}$$

giving a phonon-drag contribution to the thermopower. Using the definition of the Seebeck coefficient for an open circuit system, we can write

$$\mathcal{S}_{\rm ph} = \frac{E_x}{(dT/dx)} \approx -\left(\frac{1}{3enV}\right)\frac{dE_{\rm ph}}{dT} = \frac{C_{\rm ph}}{3en} \tag{5.111}$$

where $C_{\rm ph}$ is the phonon heat capacity per unit volume. Although this is only a rough approximate derivation, it predicts the correct temperature dependence, in that the phonon-drag contribution is important at temperatures where the phonon specific heat is large. The total thermopower is a sum of the diffusion contribution (considered in §5.4.1) and the phonon drag term $\mathcal{S}_{\rm ph}$.

The phonon drag effect depends on the electron-phonon coupling; at higher temperatures where the phonon-phonon coupling (Umklapp processes) becomes more important than the electron-phonon coupling, phonon drag effects become less important (see §6.4.4).

Chapter 6

Electron and Phonon Scattering

References:

- Kittel, *Introduction to Solid State Physics*, 6^{th} Ed., Wiley, 1986, Appendix C.
- Ashcroft and Mermin, *Solid State Physics*, Holt, Rinehart and Winston, 1976, Chapters 16 and 26.
- Hang, *Theoretical Solid State Physics*, Volume 2, Pergamon 1972, Chapter 4.

6.1 Electron Scattering

The transport properties of solids depend on the availability of carriers and on their scattering rates. In the previous chapters, we focused on the carriers and their generation. In this chapter, we focus on the scattering mechanisms.

Electron scattering brings an electronic system which has been subjected to external perturbations back to equilibrium. Collisions also alter the momentum of the carriers as the electrons are brought back into equilibrium. Electron collisions can occur through a variety of mechanisms such as electron-phonon, electron-impurity, electron-defect, and electron-electron scattering processes.

In principle, the collision rates can be calculated from scattering theory. To do this, we introduce a transition probability $S(\vec{k}, \vec{k'})$ for scattering from a state $\vec{k}$ to a state $\vec{k'}$. Since electrons obey the Pauli principle, scattering will occur from an occupied to an unoccupied state. The process of scattering from $\vec{k}$ to $\vec{k'}$ decreases the distribution function $f(\vec{r}, \vec{k}, t)$ and depends on the probability that $\vec{k}$ is occupied and that $\vec{k'}$ is unoccupied. The process of scattering from $\vec{k'}$ to $\vec{k}$ increases $f(\vec{r}, \vec{k}, t)$ and depends on the probability that state $\vec{k'}$ is occupied and state $\vec{k}$ is unoccupied. We will use the following notation for describing the scattering process:

- f_k is the probability that an electron occupies a state $\vec{k}$
- $[1 - f_k]$ is the probability that state $\vec{k}$ is unoccupied
- $S(\vec{k}, \vec{k'})$ is the probability per unit time that an electron in state $\vec{k}$ will be scattered to state $\vec{k'}$

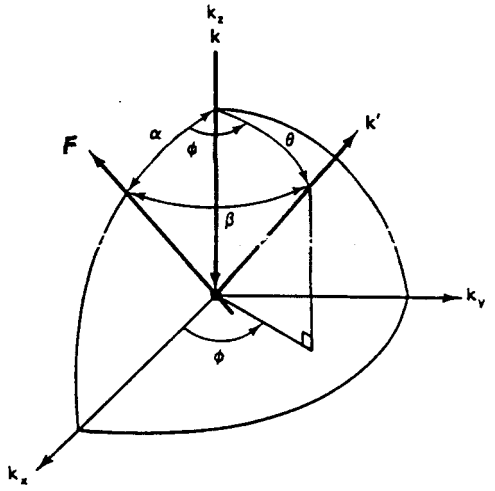


Figure 6.1: Spherical coordinate system in reciprocal space for an electron with wave vector $\vec{k}$ (along the k_z axis) scattering into a state with wavevector $\vec{k}'$ in an arbitrary force field $\vec{F}$. The scattering center is at the origin. For simplicity the event is rotated so that $\vec{F}$ has no k_y component.

- $S(\vec{k}', \vec{k})$ is the probability per unit time that an electron in state $\vec{k}'$ will be scattered back into state $\vec{k}$.

Using these definitions, the rate of change in the distribution function (see Eq. 4.4) due to collisions can be written as:

$$\left.\frac{\partial f(\vec{r}, \vec{k}, t)}{\partial t}\right|_{\text{collisions}} = \int d^3k' [f_{k'}(1 - f_k)S(\vec{k}', \vec{k}) - f_k(1 - f_{k'})S(\vec{k}, \vec{k}')] \tag{6.1}$$

where d^3k' is a volume element in $\vec{k}'$ space. The integration in Eq. 6.1 is over k space and the spherical coordinate system is shown in Fig. 6.1, together with the arbitrary force $\vec{F}$ responsible for the scattering event that introduces a perturbation

$$f_{\vec{k}} = f_{0\vec{k}} + \frac{\partial f_{0\vec{k}}}{\partial E}\frac{\hbar}{m^*}\vec{k} \cdot \vec{F} + \ldots \tag{6.2}$$

Using Fermi's Golden Rule for the transition probability per unit time between states $\vec{k}$ and $\vec{k}'$ we can write

$$S(\vec{k}, \vec{k}') \simeq \frac{2\pi}{\hbar}|\mathcal{H}_{\vec{k}\vec{k}'}|^2\{\delta[E(\vec{k})] - \delta[E(\vec{k}')]\} \tag{6.3}$$

where the matrix element of the Hamiltonian coupling states $\vec{k}$ and $\vec{k}'$ is

$$\mathcal{H}_{\vec{k}\vec{k}'} = \frac{1}{N}\int_V \psi^*_{\vec{k}}(\vec{r})\nabla V \psi_{\vec{k}'}(\vec{r})d^3r \tag{6.4}$$

in which N is the number of unit cells in the sample and ∇V is the perturbation Hamiltonian responsible for the scattering event associated with the force $\vec{F}$.

At equilibrium $f_k = f_0(E)$ and the principle of detailed balance applies

$$S(\vec{k'},\vec{k})f_0(E')[1-f_0(E)] = S(\vec{k},\vec{k'})f_0(E)[1-f_0(E')] \tag{6.5}$$

so that the distribution function does not experience a net change via collisions when in the equilibrium state

$$(\partial f(\vec{r},\vec{k},t)/\partial t)|_{\text{collisions}} = 0. \tag{6.6}$$

We define collisions as elastic when $E(\vec{k'}) = E(\vec{k})$ and in this case $f_0(E') = f_0(E)$ so that $S(\vec{k'},\vec{k}) = S(\vec{k},\vec{k'})$. Collisions for which $E(\vec{k'}) \neq E(\vec{k})$ are termed inelastic. The term quasi-elastic is used to characterize collisions where the percentage change in energy is small. For our purposes here, we shall consider $S(\vec{k},\vec{k'})$ as a known function which can be calculated quantum mechanically by a detailed consideration of the scattering mechanisms which are important for a given practical case; this statement is true in principle, but in practice $S(\vec{k},\vec{k'})$ is usually specified in an approximate way.

The return to equilibrium depends on the frequency of collisions and the effectiveness of a scattering event in randomizing the motion of the electrons. Thus, small angle scattering is not as effective in restoring a system to equilibrium as large angle scattering. For this reason we distinguish between τ_D, the time for the system to be restored to equilibrium, and τ_c, the time between collisions. These times are related by

$$\tau_D = \frac{\tau_c}{1-\cos\theta} \tag{6.7}$$

where θ is the mean change of angle of the electron on collision (see Fig. 6.1). The time τ_D is the quantity which enters into Boltzmann's equation while $1/\tau_c$ determines the actual scattering rate.

The mean free time between collisions, τ_c, is related to several other quantities of interest: the mean free path ℓ_f, the scattering cross section σ_d, and the concentration of scattering centers N_c by

$$\tau_c = \frac{1}{N_c\sigma_d v} \tag{6.8}$$

where v is the drift velocity given by

$$v = \frac{\ell_f}{\tau_c} = \frac{1}{N_c\sigma_d\tau_c} \tag{6.9}$$

and is in the direction of the electron transport. The drift velocity is of course very much smaller in magnitude than the instantaneous velocity of the electron at the Fermi level, which is typically of magnitude $v_F \sim 10^8$ cm/sec. Electron scattering centers include phonons, impurities, dislocations, the crystal surface, etc.

The most important electron scattering mechanism for both metals and semiconductors is electron-phonon scattering (scattering of electrons by the thermal motion of the lattice), though the scattering process for metals differs in detail from that in semiconductors. In the case of metals, much of the Brillouin zone is occupied by electrons, while in the case of semiconductors, most of the Brillouin zone is unoccupied and represents states into which electrons can be scattered. In the case of metals, electrons are scattered from one point on the Fermi surface to another point, and a large change in momentum occurs, corresponding to a large change in $\vec{k}$. In the case of semiconductors, changes in wave vector from $\vec{k}$ to $-\vec{k}$

normally correspond to a very small change in wave vector, and thus changes from $\vec{k}$ to $-\vec{k}$ can be accomplished much more easily in the case of semiconductors. By the same token, small angle scattering (which is not so efficient for returning the system to equilibrium) is especially important for semiconductors where the change in wavevector is small. Since the scattering processes in semiconductors and metals are quite different, they will be discussed separately below.

Scattering probabilities for more than one scattering process are additive and therefore so are the reciprocal scattering time or scattering rates:

$$(\tau^{-1})_{\text{total}} = \sum_i \tau_i^{-1} \tag{6.10}$$

since $1/\tau$ is proportional to the scattering probability. Metals have large Fermi wavevectors k_F, and therefore large momentum transfers Δk can occur as a result of electronic collisions. In contrast, for semiconductors, k_F is small and so also is Δk on collision.

6.2 Scattering Processes in Semiconductors

6.2.1 Electron-Phonon Scattering

Electron-phonon scattering is the dominant scattering mechanism in crystalline semiconductors except at very low temperatures. Conservation of energy in the scattering process, which creates or absorbs a phonon of energy $\hbar\omega(\vec{q})$, is written as:

$$E_i - E_f = \pm\hbar\omega(\vec{q}) = \frac{\hbar^2}{2m^*}(k_i^2 - k_f^2), \tag{6.11}$$

where E_i is the initial energy, E_f is the final energy, k_i the initial wavevector, and k_f the final wavevector. Here, the "+" sign corresponds to the creation of phonons (the phonon emission process), while the "–" sign corresponds to the absorption of phonons. Conservation of momentum in the scattering by a phonon of wavevector $\vec{q}$ yields

$$\vec{k_i} - \vec{k_f} = \pm\vec{q}. \tag{6.12}$$

For semiconductors, the electrons involved in the scattering event generally remain in the vicinity of a single band extremum and involve only a small change in $\vec{k}$ and hence only low phonon $\vec{q}$ vectors participate. The probability that an electron makes a transition from an initial state i to a final state f is proportional to:

(a) the availability of final states for electrons,

(b) the probability of absorbing or emitting a phonon,

(c) the strength of the electron-phonon coupling or interaction.

The first factor, the availability of final states, is proportional to the density of final electron states $\rho(E_f)$ times the probability that the final state is unoccupied. (This occupation probability for a semiconductor is assumed to be unity since the conduction band is essentially empty.) For a parabolic band $\rho(E_f)$ is (from Eq. 4.64):

$$\rho(E_f) = \frac{(2m^*)^{3/2}E_f^{1/2}}{2\pi^2\hbar^3} = (2m^*)^{3/2}\frac{[E_i \pm \hbar\omega(\vec{q})]^{1/2}}{2\pi^2\hbar^3}, \tag{6.13}$$

where Eq. 6.11 has been employed and the "+" sign corresponds to absorption of a phonon and the "–" sign corresponds to phonon emission.

The probability of absorbing or emitting a phonon is proportional to the electron-phonon coupling $G(\vec{q})$ and to the phonon density $n(\vec{q})$ for absorption, and $[1 + n(\vec{q})]$ for emission, where $n(\vec{q})$ is given by the Bose-Einstein factor

$$n(\vec{q}) = \frac{1}{e^{\frac{\hbar\omega(\vec{q})}{k_B T}} - 1}. \tag{6.14}$$

Combining the terms in Eqs. 6.13 and 6.14 gives a scattering probability (or $1/\tau_c$) proportional to a sum over final states

$$\frac{1}{\tau_c} \sim \frac{(2m^*)^{3/2}}{2\pi^2\hbar^3} \sum_{\vec{q}} G(\vec{q}) \left[\frac{[E_i + \hbar\omega(\vec{q})]^{1/2}}{e^{\frac{\hbar\omega(\vec{q})}{k_B T}} - 1} + \frac{[E_i - \hbar\omega(\vec{q})]^{1/2}}{1 - e^{\frac{-\hbar\omega(\vec{q})}{k_B T}}} \right] \tag{6.15}$$

where the first term in the big bracket of Eq. 6.15 corresponds to phonon absorption and the second term to phonon emission. If $E_i < \hbar\omega(\vec{q})$, only the phonon absorption process is energetically allowed.

The electron-phonon coupling coefficient $G(\vec{q})$ in Eq. 6.15 depends on the electron-phonon coupling mechanism. There are three important coupling mechanisms in semiconductors which we briefly describe below: electromagnetic coupling, piezoelectric coupling, and deformation-potential coupling.

Electromagnetic coupling is important only for semiconductors where the charge distribution has different signs on neighboring ion sites. In this case, the oscillatory electric field can give rise to oscillating dipole moments associated with the motion of neighboring ion sites in the LO mode (see Fig. 6.2). The electromagnetic coupling mechanism is important in coupling electrons to longitudinal optical phonon modes in III-V and II-VI compound semiconductors, but does not contribute in the case of silicon. To describe the LO mode we can use the Einstein approximation, since $\omega(\vec{q})$ is only weakly dependent on $\vec{q}$ for the optical modes. In this case $\hbar\omega_0 \gg k_B T$ and $\hbar\omega_0 \gg E$, so that from Eq. 6.15 the collision rate is proportional to

$$\frac{1}{\tau_c} \sim \frac{m^{*3/2}(\hbar\omega_0)^{1/2}}{e^{\hbar\omega_0/k_B T} - 1}. \tag{6.16}$$

Thus, the collision rate depends on the temperature T, the LO phonon frequency ω_0 and the electron effective mass m^*. The corresponding mobility for the optical phonon scattering is

$$\mu = \frac{e\langle\tau\rangle}{m^*} \sim \frac{e(e^{\hbar\omega_0/k_B T} - 1)}{m^{*5/2}(\hbar\omega_0)^{1/2}} \tag{6.17}$$

Thus for optical phonon scattering, the mobility μ is independent of the electron energy E and decreases with increasing temperature.

As in the case of electromagnetic coupling, piezoelectric coupling is important in semiconductors which are ionic or partly ionic. If these crystals lack inversion symmetry, then *acoustic* mode vibrations generate regions of compression and rarefaction in a crystal which lead to electric fields (see Fig. 6.3). The piezoelectric scattering mechanism is thus associated with the coupling between electrons and phonons arising from these electromagnetic

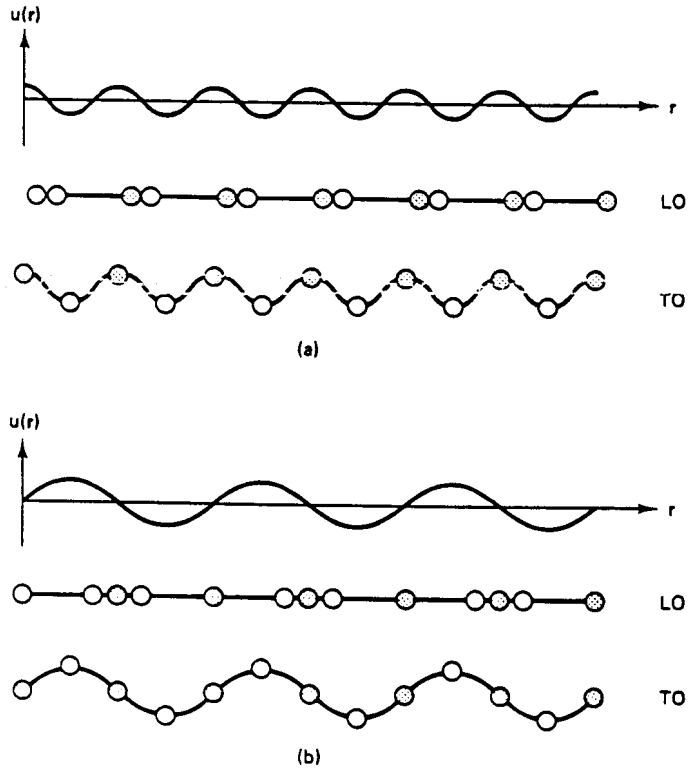


Figure 6.2: Displacements $\vec{u}(\vec{r})$ of a diatomic chain for LO and TO phonons at (a) the center and (b) the edge of the Brillouin zone. The lighter mass atoms are indicated by open circles. For zone edge optical phonons, only the lighter atoms are displaced.

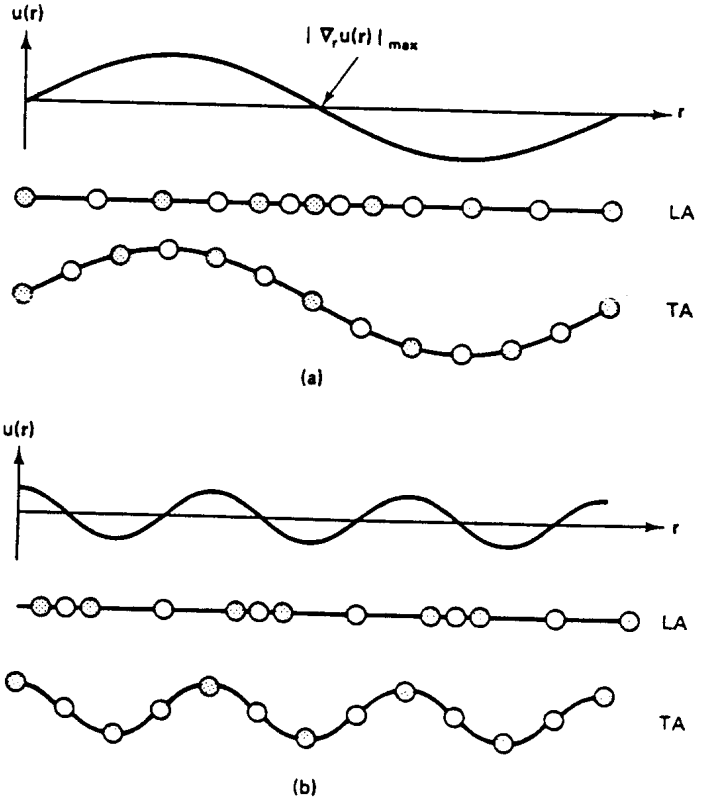


Figure 6.3: Displacements $\vec{u}(\vec{r})$ of a diatomic chain for LA and TA phonons at (a) the center and (b) the edge of the Brillouin zone. The lighter mass atoms are indicated by open circles. For zone edge acoustic phonons, only the heavier atoms are displaced.

fields. The zincblende structure of the III–V compounds (e.g., GaAs) lacks inversion symmetry. In this case the perturbation potential

$$\Delta V(\vec{r},t) = \frac{-ie\varepsilon_{pz}}{\epsilon_0 q}\vec{\nabla}\cdot\vec{u}(\vec{r},t) \tag{6.18}$$

where ε_{pz} is the piezoelectric coefficient and $\vec{u}(\vec{r},t) = u\exp(i\vec{q}\cdot\vec{r}-\omega t)$ is the displacement during a normal mode oscillation. Note that the phase of $\Delta V(\vec{r},t)$ in piezoelectric coupling is shifted by $\pi/2$ relative to the case of electromagnetic coupling.

The deformation-potential coupling mechanism is associated with energy shifts of the energy band extrema caused by the compression and rarefaction of crystals during acoustic mode vibrations. The deformation potential scattering mechanism is important in crystals like silicon which have inversion symmetry (and hence no piezoelectric scattering coupling) and have the same species on each site (and hence no electromagnetic coupling). The longitudinal acoustic modes are important for phonon coupling in n-type Si and Ge where the conduction band minima occur away from $\vec{k}=0$.

For deformation potential coupling, it is the LA acoustical phonons that are most important though contributions by LO optical phonons still make a contribution. For the acoustic phonons, we have the condition $\hbar\omega \ll k_BT$ and $\hbar\omega \ll E$, while for the optical phonons it is usually the case that $\hbar\omega \gg k_BT$ at room temperature. For the range of acoustic phonon modes of interest, $G(\vec{q}) \sim q$, where q is the phonon wave vector and $\omega \sim q$ for acoustic phonons. Furthermore for the LA phonon branch the phonon absorption process will depend on $n(q)$ in accordance with the Bose factor

$$\frac{1}{e^{\hbar\omega/k_BT}-1} \simeq \frac{1}{[1+\frac{\hbar\omega}{k_BT}+\cdots]-1} \sim \frac{k_BT}{\hbar\omega} \sim \frac{k_BT}{q}, \tag{6.19}$$

while for phonon emission

$$\frac{1}{1-e^{-\hbar\omega/k_BT}} \simeq \frac{1}{1-[1-\frac{\hbar\omega}{k_BT}+\cdots]} \sim \frac{k_BT}{\hbar\omega} \sim \frac{k_BT}{q}. \tag{6.20}$$

Therefore, in considering both phonon absorption and phonon emission, the factors

$$G(\vec{q})[e^{\hbar\omega/k_BT}-1]^{-1}$$

and

$$G(\vec{q})[1-e^{-\hbar\omega/k_BT}]^{-1}$$

are independent of q for the LA branch. Consequently for the acoustic phonon scattering process, the carrier mobility μ decreases with increasing T according to (see Eq. 6.15)

$$\mu = \frac{e\langle\tau\rangle}{m^*} \sim m^{*-5/2}E^{-1/2}(k_BT)^{-1}. \tag{6.21}$$

For the optical LO contribution, we have $G(\vec{q})$ independent of $\vec{q}$ but an $E^{1/2}$ factor is introduced by Eq. 6.15 for both phonon absorption and emission, leading to the same basic dependence as given by Eq. 6.21. Thus, we find that the temperature and energy dependence of the mobility μ is different for the various electron-phonon coupling mechanisms. These differences in the E and T dependences can thus be used to identify which scattering mechanism is dominant in specific semiconducting samples. Furthermore, when explicit account is taken of the energy dependence of τ, departures from the strict Drude model $\sigma = ne^2\tau/m^*$ can be expected.

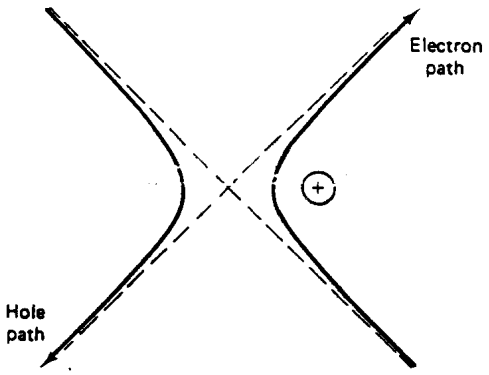


Figure 6.4: Trajectories of electrons and holes in ionized impurity scattering. The scattering center is at the origin.

6.2.2 Ionized Impurity Scattering

As the temperature is reduced, phonon scattering becomes less important so that in this regime, ionized impurity scattering and other defect scattering mechanisms can become dominant. Ionized impurity scattering can also be important in heavily doped semiconductors over a wider temperature range because of the larger defect density. This scattering mechanism involves the deflection of an electron with velocity v by the Coulomb field of an ion with charge Ze, as modified by the dielectric constant ε of the medium and by the screening of the impurity ion by free electrons (see Fig. 6.4). Most electrons are scattered through small angles as they are scattered by ionized impurities. The perturbation potential is given by

$$\Delta V(\vec{r}) = \frac{\pm Ze^2}{4\pi\epsilon_0 r} \tag{6.22}$$

and the $\pm$ signs denote the different scattering trajectories for electrons and holes (see Fig. 6.4). In Eq. 6.22 the screening of the electron by the semiconductor environment is handled by the static dielectric constant of the semiconductor ϵ_0. Because of the long-range nature of the Coulomb interaction, screening by other free carriers and by other ionized impurities could be important. Such screening effects are further discussed in §6.2.4.

The scattering rate $1/\tau_I$ due to ionized impurity scattering is given to a good approximation by the Conwell–Weisskopf formula

$$\frac{1}{\tau_I} \sim \frac{Z^2 N_I}{m^{*1/2} E^{3/2}}\ \ell n \left\{ 1 + \left[\frac{4\pi\varepsilon E}{Ze^2 N_I^{1/3}} \right]^2 \right\} \tag{6.23}$$

in which N_I is the ionized charged impurity density. The Conwell–Weisskopf formula works quite well for heavily doped semiconductors. We note here that $\tau_I \sim E^{3/2}$, so that it is the low energy electrons that are most effected by ionized impurity scattering (see Fig. 4.11).

Neutral impurities also introduce a scattering potential, but it is much weaker than that for the ionized impurity. Free carriers can polarize a neutral impurity and interact with the

resulting dipole moment, or can undergo an exchange interaction. In the case of neutral impurity scattering, the perturbation potential is given by

$$\Delta V(\vec{r}) \simeq \frac{\hbar^2}{m^*}\left(\frac{r_B}{r^5}\right)^{1/2} \tag{6.24}$$

where r_B is the ground state Bohr radius of the electron in a doped semiconductor and r is the distance of the electron to the impurity scattering center.

6.2.3 Other Scattering Mechanisms

Other scattering mechanisms in semiconductors include:

(a) neutral impurity centers — these make contributions at very low temperatures, and are mentioned in §6.2.2.

(b) dislocations — these defects give rise to anisotropic scattering at low temperatures.

(c) boundary scattering by crystal surfaces — this scattering becomes increasingly important the smaller the crystal size.

(d) intervalley scattering from one equivalent conduction band minimum to another. This scattering process requires a phonon with large q and consequently results in a relatively large energy transfer.

(e) electron-electron scattering – similar to charged impurity scattering. This mechanism can be important in distributing energy and momentum among the electrons in the solid and thus can act in conjunction with other scattering mechanisms in establishing equilibrium.

(f) electron-hole scattering — depends on having both electrons and holes present. Because the electron and hole motions induced by an applied electric field are in opposite directions, electron-hole scattering tends to reverse the direction of the incident electrons and holes.

6.2.4 Screening Effects in Semiconductors

In the vicinity of a charged impurity or an acoustic phonon, charge carriers are accumulated or depleted by the scattering potential, giving rise to a charge density

$$\rho(\vec{r}) = e[n(\vec{r}) - p(\vec{r}) + N_a^-(\vec{r}) - N_d^+(\vec{r})] = en^*(\vec{r}) \tag{6.25}$$

where $n(\vec{r})$, $p(\vec{r})$, $N_a^-(\vec{r})$, $N_d^+(\vec{r})$, and $n^*(\vec{r})$ are, respectively, the electron, hole, ionized acceptor, ionized donor, and effective total carrier concentrations as a function of distance r to the scatterer. We can then write expressions for these quantities in terms of their excess charge above the uniform potential in the absence of impurities:

$$\begin{aligned} n(\vec{r}) &= n + \delta n(\vec{r}) \\ N_d^+(\vec{r}) &= N_d^+ + \delta N_d^+(\vec{r}), \end{aligned} \tag{6.26}$$

and similarly for the holes and acceptors. The space charge $\rho(\vec{r})$ is related to the perturbing potential by Poisson's equation

$$\nabla^2\phi(\vec{r}) = -\frac{\rho(\vec{r})}{\epsilon_0}. \tag{6.27}$$

Approximate relations for the excess concentrations are

$$\begin{aligned} \delta n(\vec{r})/n &\simeq -e\phi(\vec{r})/(k_B T) \\ \delta N_d^+(\vec{r})/N_d^+ &\simeq e\phi(\vec{r})/(k_B T) \end{aligned} \tag{6.28}$$

and similar relations for the holes. Substitution of Eq. 6.25 into Eqs. 6.26 and 6.28 yield

$$\nabla^2\phi(\vec{r}) = -\frac{n^* e^2}{\epsilon_0 k_B T}\phi(\vec{r}). \tag{6.29}$$

We define an effective Debye screening length λ such that

$$\lambda^2 = \frac{\epsilon_0 k_B T}{n^* e^2}. \tag{6.30}$$

For a spherically symmetric potential Eq. 6.29 becomes

$$\frac{d^2}{dr^2}\Big(r\phi(r)\Big) = \frac{r\phi(r)}{\lambda^2} \tag{6.31}$$

which yields a solution

$$\phi(r) = \frac{Ze^2}{4\pi\epsilon_0 r}e^{-r/\lambda}. \tag{6.32}$$

Thus, the screening effect produces an exponential decay of the scattering potential $\phi(r)$ with a characteristic length λ that depends through Eq. 6.30 on the effective electron concentration. When the concentration gets large, λ decreases and screening becomes more effective.

When applying screening effects to the ionized impurity scattering problem, we Fourier expand the scattering potential to take advantage of the overall periodicity of the lattice

$$\Delta V(\vec{r}) = \sum_G A_G \exp(i\vec{G}\cdot\vec{r}) \tag{6.33}$$

where the Fourier coefficients are given by

$$A_G = \frac{1}{V}\int_V \nabla V(\vec{r}) \exp(-i\vec{G}\cdot\vec{r}) d^3r \tag{6.34}$$

and the matrix element of the perturbation Hamiltonian in Eq. 6.4 becomes

$$\mathcal{H}_{\vec{k},\vec{k}'} = \frac{1}{N}\sum_G \int_V e^{-i\vec{k}\cdot\vec{r}} u_k^*(r) A_G e^{-i\vec{G}\cdot\vec{r}} e^{i\vec{k}'\cdot\vec{r}} u_{k'}(r) d^3r. \tag{6.35}$$

We note that the integral in Eq. 6.35 vanishes unless $\vec{k} - \vec{k}' = \vec{G}$ so that

$$\mathcal{H}_{\vec{k},\vec{k}'} = \frac{A_G}{N}\int_V u_k^*(r) u_{k'}(r) d^3r \tag{6.36}$$

within the first Brillouin zone so that for parabolic bands $u_k(\vec{r}) = u_{k'}(\vec{r})$ and

$$\mathcal{H}_{\vec{k},\vec{k}'} = A_{\vec{k}-\vec{k}'}. \tag{6.37}$$

Now substituting for the screening potential in Eq. 6.34 we obtain

$$A_G = \frac{Ze^2}{4\pi\epsilon_0 V}\int_V \exp(-i\vec{G}\cdot\vec{r})d^3r \tag{6.38}$$

where $d^3r = r^2\sin\theta d\theta d\phi dr$ so that the angular integration gives 4π and the spatial integration gives

$$A_G = \frac{Ze^2}{\epsilon_0 V|\vec{G}|^2} \tag{6.39}$$

and

$$\mathcal{H}_{\vec{k},\vec{k}'} = \frac{Ze^2}{\epsilon_0 V|\vec{k}-\vec{k}'|^2}. \tag{6.40}$$

When screening is included in considering the ionized impurity scattering mechanism the integration becomes

$$A_G = \frac{Ze^2}{4\pi\epsilon_0 V}\int_V e^{-r/\lambda}e^{-i\vec{G}\cdot\vec{r}}d^3r = \frac{Ze^2}{\epsilon_0 V[|\vec{G}|^2 + |1/\lambda|^2]} \tag{6.41}$$

and

$$\mathcal{H}_{\vec{k},\vec{k}'} = \frac{Ze^2}{\epsilon_0 V[|\vec{k}-\vec{k}'|^2 + |1/\lambda|^2]} \tag{6.42}$$

so that screening clearly reduces the scattering due to ionized impurity scattering. The discussion given here also extends to the case of scattering in metals, which is treated below.

Combining the various scattering mechanisms discussed above for semiconductors, the following picture emerges (see Fig. 6.5), where the effect of these processes on the carrier mobility is considered. Here it is seen that screening effects are important for carrier mobilities at low temperature.

6.3 Electron Scattering in Metals

Basically the same class of scattering mechanisms are present in metals as in semiconductors, but because of the large number of occupied states in the conduction bands of metals, the temperature dependences of the various scattering mechanisms are quite different.

6.3.1 Electron-Phonon Scattering

For a review of phonons in the harmonic oscillator approximation see Appendix B.

In metals as in semiconductors, the dominant scattering mechanism is usually electron-phonon scattering. In the case of metals, electron scattering is mainly associated with an electromagnetic interaction of ions with nearby electrons, the longer range interactions being **screened** by the numerous mobile electrons. For metals, we must therefore consider explicitly the probability that a state $\vec{k}$ is occupied $f_0(\vec{k})$ or unoccupied $[1 - f_0(\vec{k})]$. The

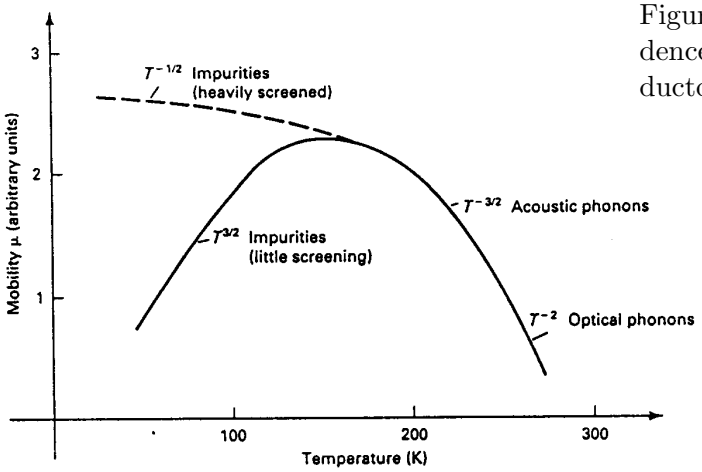


Figure 6.5: Typical temperature dependence of the carrier mobility in semiconductors.

scattering rate is found by explicit consideration of the scattering rate into a state $\vec{k}$ and the scattering out of that state. Using the same arguments as in §6.2.1, the collision term in Boltzmann's equation is given by

$$\left.\frac{\partial f}{\partial t}\right|_{\text{collisions}} \sim \frac{1}{\tau} \simeq$$

$$\sum_q G(\vec{q}) \left\{ \overbrace{[1-f_0(\vec{k})]\Big[\underbrace{f_0(\vec{k}-\vec{q})n(\vec{q})}_{\text{phonon absorption}} + \underbrace{f_0(\vec{k}+\vec{q})[1+n(\vec{q})]}_{\text{phonon emission}}\Big]}^{\text{scattering into } \vec{\text{k}}} \right. \tag{6.43}$$

$$\left. - \overbrace{[f_0(\vec{k})]\Big[\underbrace{[1-f_0(\vec{k}+\vec{q})]n(\vec{q})}_{\text{phonon absorption}} + \underbrace{[1-f_0(\vec{k}-\vec{q})][1+n(\vec{q})]}_{\text{phonon emission}}\Big]}^{\text{scattering out of } \vec{\text{k}}} \right\}$$

Here the first term in Eq. 6.43 is associated with scattering electrons into an element of phase space at $\vec{k}$ with a probability given by $[1-f_0(\vec{k})]$ that state $\vec{k}$ is unoccupied and has contributions from both phonon absorption processes and phonon emission processes. The second term arises from electrons scattered out of state $\vec{k}$ and here, too, there are contributions from both phonon absorption processes and phonon emission processes. The equilibrium distribution function $f_0(\vec{k})$ for the electron is the Fermi distribution function while the function $n(\vec{q})$ for the phonons is the Bose distribution function (Eq. 6.14). Phonon absorption depends on the phonon density $n(\vec{q})$, while phonon emission depends on the factor $\{1+n(\vec{q})\}$. These factors arise from the properties of the creation and annihilation operators for phonons (to be reviewed in recitation). The density of final states for metals is the density of states at the Fermi level which is consequently approximately independent

Symbol	Metal	Θ_D(K)
$\oplus$	Au	175
$\circ$	Na	202
$\triangle$	Cu	333
$\square$	Al	395
$\bullet$	Ni	472

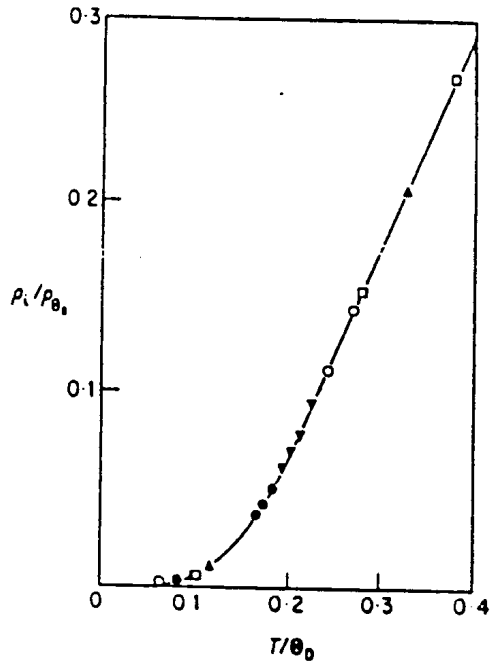


Figure 6.6: Universal curve of the temperature dependence of the ideal resistivity of various metals normalized to the value at the Debye temperature as a function of the dimensionless temperature T/Θ_D.

of energy and temperature. The condition that, in metals, electron scattering takes place to states near the Fermi level implies that the largest phonon wave vector in an electron collision is $2k_F$ where k_F is the electron wave vector at the Fermi surface.

Of particular interest is the temperature dependence of the phonon scattering mechanism in the limit of low and high temperatures. Experimentally, the temperature dependence of the resistivity of metals can be plotted on a universal curve (see Fig. 6.6) in terms of ρ_T/ρ_{Θ_D} vs. T/Θ_D where Θ_D is the Debye temperature. This plot includes data for several metals, and values for the Debye temperature of these metals are given with the figure.

In accordance with the plot in Fig. 6.6, $T \ll \Theta_D$ defines the low temperature limit and $T \gg \Theta_D$ the high temperature limit. Except for the very low temperature defect scattering limit, the electron-phonon scattering mechanism dominates, and the temperature dependence of the scattering rate depends on the product of the density of phonon states

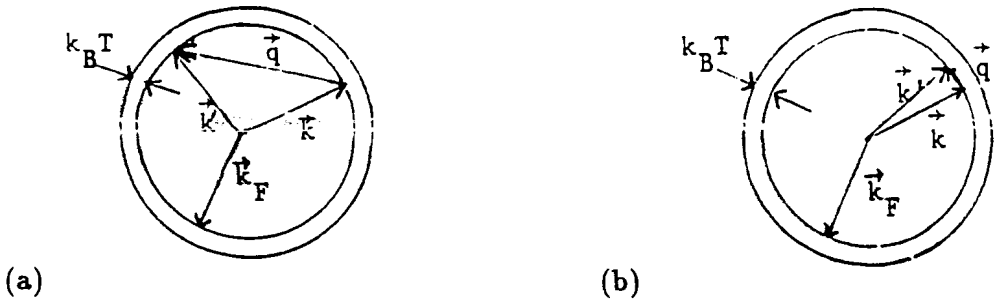


Figure 6.7: (a) Scattering of electrons on the Fermi surface of a metal. Large angle scattering dominates at high temperature ($T > \Theta_D$) and this regime is called the "quasi-elastic" limit. (b) Small angle scattering is important at low temperature ($T < \Theta_D$) and is in general an inelastic scattering process.

and the phonon occupation, since the electron-phonon coupling coefficient is essentially independent of T. The phonon concentration in the high temperature limit becomes

$$n(\vec{q}) = \frac{1}{\exp(\hbar\omega/k_BT) - 1} \approx \frac{k_BT}{\hbar\omega} \tag{6.44}$$

since $(\hbar\omega/k_BT) \ll 1$, so that from Eq. 6.44 we have $1/\tau \sim T$ and $\sigma = ne\mu \sim T^{-1}$. In this high temperature limit, the scattering is quasi-elastic and involves large-angle scattering, since phonon wave vectors up to the Debye wave vector q_D are involved in the electron scattering where q_D is related to the Debye frequency ω_D and to the Debye temperature Θ_D according to

$$\hbar\omega_D = k_B\Theta_D = \hbar q_D v_q \tag{6.45}$$

where v_q is the velocity of sound.

We can interpret q_D as the radius of a Debye sphere in $\vec{k}$-space which defines the range of accessible $\vec{q}$ vectors for scattering, i.e., $0 < q < q_D$. The magnitude of q_D is comparable to the Brillouin zone dimensions but the energy change of an electron (ΔE) on scattering by a phonon will be less than $k_B\Theta_D \simeq 1/40eV$ so that the restriction of $(\Delta E)_{max} \simeq k_B\Theta_D$ implies that the maximum electronic energy change on scattering will be small compared with the Fermi energy E_F. We thus obtain that for $T > \Theta_D$ (the high temperature regime), $\Delta E < k_BT$ and the scattering will be quasi-elastic as illustrated in Fig. 6.7a. In the opposite limit, $T \ll \Theta_D$, we have $\hbar\omega_q \simeq k_BT$ (because only low frequency acoustic phonons are available for scattering) and in the low temperature limit there is the possibility that $\Delta E > k_BT$, which implies inelastic scattering. In the low temperature limit, $T \ll \Theta_D$, the scattering is also small-angle scattering, since only low energy (low wave vector) phonons are available for scattering (as illustrated in Fig. 6.7b). At low temperature, the phonon density contributes a factor of T^3 to the scattering rate (Eq. 6.43) when the sum over phonon states is converted to an integral and q^2dq is written in terms of the dimensionless variable $\hbar\omega_q/k_BT$ with $\omega = v_qq$. Since small momentum transfer gives rise to small angle scattering, the diagram in Fig. 6.8 involves Eq. 6.7. Because of the small energy transfer

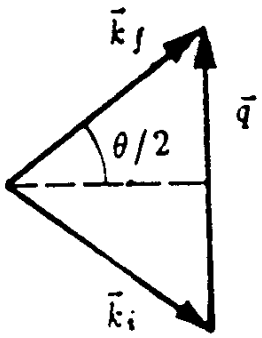


Figure 6.8: Geometry of the scattering process, where θ is the scattering angle between the incident and scattered wave vectors $\vec{k}_i$ and $\vec{k}_f$, respectively.

we can write,

$$|\vec{k_i} - \vec{k_f}| \sim k_f(1 - \cos\theta) \approx \frac{1}{2}k_f\theta^2 \approx \frac{1}{2}k_f(q/k_f)^2 \tag{6.46}$$

so that another factor of q^2 appears in the integration over $\vec{q}$ when calculating $(1/\tau_D)$. Thus, the electron scattering rate at low temperature is predicted to be proportional to T^5 so that $\sigma \sim T^{-5}$ (Bloch–Grüneisen formula). Thus, when phonon scattering is the dominant scattering mechanism in metals, the following results are obtained:

$$\sigma \sim \Theta_D/T \qquad T \gg \Theta_D \tag{6.47}$$

$$\sigma \sim (\Theta_D/T)^5 \qquad T \ll \Theta_D \tag{6.48}$$

In practice, the resistivity of metals at very low temperatures is dominated by other scattering mechanisms such as impurities, etc., and electron-phonon scattering (see Eq. 6.48) is relatively unimportant.

The possibility of umklapp processes further increases the range of phonon modes that can contribute to electron scattering in electron-phonon scattering processes. In an umklapp process, a non-vanishing reciprocal lattice vector can be involved in the momentum conservation relation as shown in the schematic diagram of Fig. 6.9.

In this diagram, the relation between the wave vectors for the phonon and for the incident and scattered electrons $\vec{G} = \vec{k} + \vec{q} + \vec{k'}$ is shown when crystal momentum is conserved for a non-vanishing reciprocal lattice vector $\vec{G}$. Thus, phonons involved in an umklapp process have large wave vectors with magnitudes of about 1/3 of the Brillouin zone dimensions. Therefore, substantial energies can be transferred on collision through an umklapp process. At low temperatures, normal scattering processes (i.e., normal as distinguished from umklapp processes) play an important part in completing the return to equilibrium of an excited electron in a metal, while at high temperatures, umklapp processes become more important.

This discussion is applicable to the creation or absorption of a single phonon in a particular scattering event. Since the restoring forces for lattice vibrations in solids are not strictly harmonic, anharmonic corrections to the restoring forces give rise to multiphonon

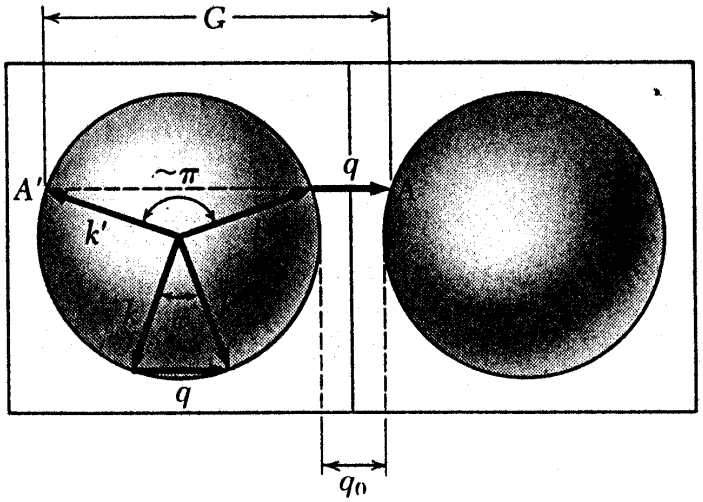


Figure 6.9: Schematic diagram showing the relation between the phonon wave vector $\vec{q}$ and the electron wave vectors $\vec{k}$ and $\vec{k}'$ in two Brillouin zones separated by the reciprocal lattice vector $\vec{G}$ (umklapp process).

processes where more than one phonon can be created or annihilated in a single scattering event. Experimental evidence for multiphonon processes is provided in both optical and transport studies. In some cases, more than one phonon at the same frequency are created (harmonics), while in other cases, multiple phonons at different frequencies (overtones) are involved.

6.3.2 Other Scattering Mechanisms in Metals

At very low temperatures where phonon scattering is of less importance, other scattering mechanisms become important, and

$$\frac{1}{\tau} = \sum_i \frac{1}{\tau_i} \tag{6.49}$$

where the sum is over all the scattering processes.

(a) Charged impurity scattering — The effect of charged impurity scattering (Z being the difference in the charge on the impurity site as compared with the charge on a regular lattice site) is of less importance in metals than in semiconductors because of screening effects by the free electrons.

(b) Neutral impurities — This process pertains to scattering centers having the same charge as the host. Such scattering has less effect on the transport properties than scattering by charged impurity sites, because of the much weaker scattering potential.

(c) Vacancies, interstitials, dislocations, size-dependent effects — the effects for these defects on the transport properties are similar to those for semiconductors.

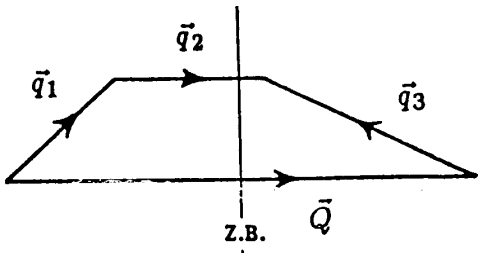


Figure 6.10: Phonon-phonon umklapp processes.

For most metals, phonon scattering is relatively unimportant at liquid helium temperatures, so that resistivity measurements at 4K provide a method for the detection of impurities and crystal defects. In fact, in characterizing the quality of a high purity metal sample, it is customary to specify the resistivity ratio $\rho(300\text{K})/\rho(4\text{K})$. This quantity is usually called the residual resistivity ratio (RRR), or the residual resistance ratio. In contrast, a typical semiconductor is characterized by its conductivity and Hall coefficient at room temperature and at 77 K.

6.4 Phonon Scattering

Whereas electron scattering is important in electronic transport properties, phonon scattering is important in thermal transport, particularly for the case of insulators where heat is carried mainly by phonons. The major scattering mechanisms for phonons are phonon-phonon scattering, phonon-boundary scattering and defect-phonon scattering, which are briefly discussed in the following subsections.

6.4.1 Phonon-phonon scattering

Phonons are scattered by other phonons because of anharmonic terms in the restoring potential. This scattering process permits:

- two phonons to combine to form a third phonon or
- one phonon to break up into two phonons.

In these anharmonic processes, energy and wavevector conservation apply:

$$\vec{q}_1 + \vec{q}_2 = \vec{q}_3 \qquad \text{normal processes} \tag{6.50}$$

or

$$\vec{q}_1 + \vec{q}_2 = \vec{q}_3 + \vec{Q} \qquad \text{umklapp processes} \tag{6.51}$$

where $\vec{Q}$ corresponds to a phonon wave vector of magnitude comparable to that of reciprocal lattice vectors. When umklapp processes (see Fig. 6.10) are present, the scattered phonon

wavevector $\vec{q}_3$ can be in a direction opposite to the energy flow, thereby giving rise to thermal resistance. Because of the high momentum transfer and the large phonon energies that are involved, umklapp processes dominate the thermal conductivity at high T.

The phonon density is proportional to the Bose factor so that the scattering rate is proportional to

$$\frac{1}{\tau_{\rm ph}} \sim \frac{1}{(e^{\hbar\omega/(k_B T)} - 1)}. \tag{6.52}$$

At high temperatures $T \gg \Theta_D$, the scattering time thus varies as T^{-1} since

$$\tau_{\rm ph} \sim (e^{\hbar\omega/k_B T} - 1) \sim \hbar\omega/k_B T \tag{6.53}$$

while at low temperatures $T \sim \Theta_D$, an exponential temperature dependence for $\tau_{\rm ph}$ is found

$$\tau_{\rm ph} \sim e^{\hbar\omega/k_B T} - 1. \tag{6.54}$$

These temperature dependences are important in considering the lattice contribution to the thermal conductivity (see §5.2.4).

6.4.2 Phonon-Boundary Scattering

Phonon-boundary scattering is important at low temperatures where the phonon density is low. In this regime, the scattering time is independent of T. The thermal conductivity in this range is proportional to the phonon density which is in turn proportional to T^3. This effect combined with phonon-phonon scattering results in a thermal conductivity κ for insulators with the general shape shown in Fig. 6.11 (see §5.2.4). The lattice thermal conductivity follows the relation

$$\kappa_L = C_p v_q \Lambda_{\rm ph}/3 \tag{6.55}$$

where the phonon mean free path $\Lambda_{\rm ph}$ is related to the phonon scattering probability $(1/\tau_{\rm ph})$ by

$$\tau_{\rm ph} = \Lambda_{\rm ph}/v_q \tag{6.56}$$

in which v_q is the velocity of sound and C_p is the heat capacity at constant pressure. Phonon-boundary scattering becomes more important as the crystallite size decreases.

6.4.3 Defect-Phonon Scattering

Defect-phonon scattering includes a variety of crystal defects, charged and uncharged impurity sites and different isotopes of the host constituents. The thermal conductivity curves in Fig. 6.11 show the scattering effects due to different isotopes of Li. The low mass of Li makes it possible to see such effects clearly. Isotope effects are also important in graphite and diamond which have the highest thermal conductivity of any solid.

6.4.4 Electron-Phonon Scattering

If electrons scatter from phonons, the reverse process also occurs. When phonons impart momentum to electrons, the electron distribution is affected. Thus, the electrons will also carry energy as they are dragged also by the stream of phonons. This phenomenon is called phonon drag. In the case of phonon drag we must simultaneously solve the Boltzmann equations for the electron and phonon distributions which are coupled by the phonon drag term.

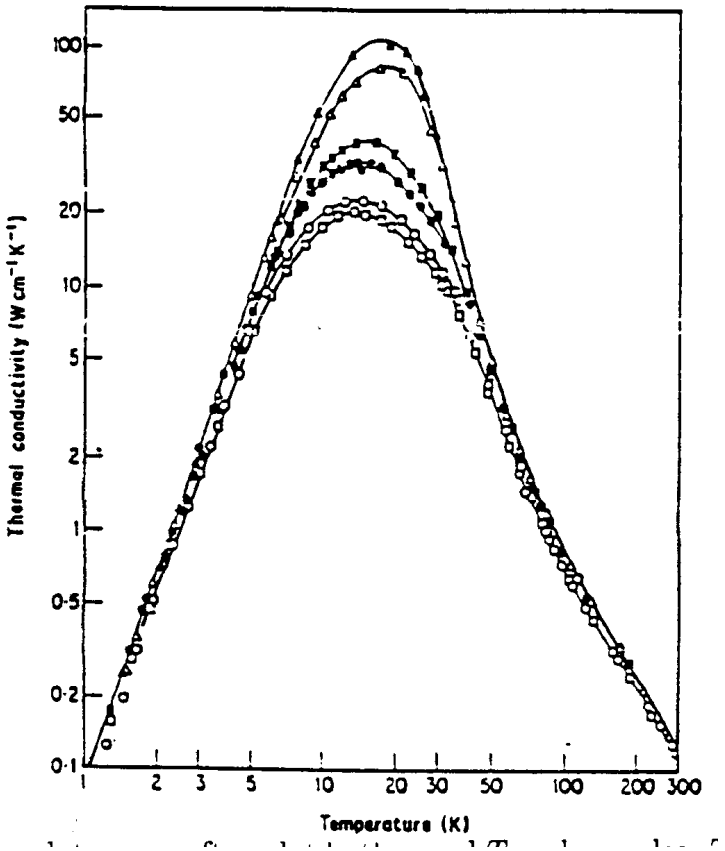


Figure 6.11: For insulators, we often plot both κ and T on log scales. The various curves are for LiF with different concentrations of Li isotopes ^{6}Li and ^{7}Li. For highly perfect crystals, it is possible to observe the scattering effects due to Li ions of different masses, which act as lattice defects.

6.5 Temperature Dependence of the Electrical and Thermal Conductivity

For the electrical conductivity, at very low temperatures, impurity, defect, and boundary scattering dominate. In this regime σ is independent of temperature. At somewhat higher temperatures but still far below Θ_D the electrical conductivity for metals exhibits a strong temperature dependence (see Eq. 6.48)

$$\sigma \propto (\Theta_D/T)^5 \qquad T \ll \Theta_D. \tag{6.57}$$

At higher temperatures where $T \gg \Theta_D$, scattering by phonons with any q vector is possible and the formula

$$\sigma \sim (\Theta_D/T) \qquad T \gg \Theta_D \tag{6.58}$$

applies. We now summarize the corresponding temperature ranges for the thermal conductivity.

Although the thermal conductivity was formally discussed in §5.2, a meaningful discussion of the temperature dependence of κ depends on scattering processes. The total thermal conductivity κ in general depends on the lattice and electronic contributions, κ_L and κ_e, respectively. The temperature dependence of the lattice contribution is discussed in §5.2.4 with regard to the various phonon scattering processes and their temperature dependence. For the electronic contribution, we must consider the temperature dependence of the electron scattering processes discussed in §5.2 and §6.2.

At very low temperatures, in the impurity scattering range, σ is independent of T and the same scattering processes apply for both the electronic thermal conductivity and the electrical conductivity so that $\kappa_e \propto T$ in the impurity scattering regime where $\sigma \sim$ constant and the Wiedemann–Franz law is applicable. From Fig. 5.1 we see that for copper, defect and boundary scattering are dominant below $\sim$ 20 K, while phonon scattering becomes important at higher T.

At low temperatures $T \ll \Theta_D$, but with T in a regime where phonon scattering has already become the dominant scattering mechanism, the thermal transport depends on the electron-phonon collision rate which in turn is proportional to the phonon density. At low temperatures the phonon density is proportional to T^3. This follows from the proportionality of the phonon density of states arising from the integration of $\int q^2 dq$, and from the dispersion relation for the acoustic phonons $\omega = qv_q$

$$q = \omega/v_q = xkT/\hbar v_q \tag{6.59}$$

where $x = \hbar\omega/k_B T$. Thus in the low temperature range of phonon scattering where $T \ll \Theta_D$ and the Wiedemann–Franz law is no longer satisfied, the temperature dependence of τ is found from the product $T(T^{-3})$ so that $\kappa_e \propto T^{-2}$. One reason why the Wiedemann–Franz law is not satisfied in this temperature regime is that κ_e depends on the collision rate τ_c while σ depends on the time to reach thermal equilibrium, τ_D. At low temperatures where only low q phonons participate in scattering events the times τ_c and τ_D are not the same.

At high T where $T \gg \Theta_D$ and the Wiedemann–Franz law applies, κ_e approaches a constant value corresponding to the regime where σ is proportional to $1/T$. This occurs at temperatures much higher than those shown in Fig. 5.1. The decrease in κ above the peak value at $\sim$17 K follows a $1/T^2$ dependence quite well.

In addition to the electronic thermal conductivity, there is heat flow due to lattice vibrations. The phonon thermal conductivity mechanism is in fact the principal mechanism operative in semiconductors and insulators, since the electronic contribution in this case is negligibly small. Since κ_L contributes also to metals the total measured thermal conductivity for metals should exceed the electronic contribution $(\pi^2 k_B^2 T\sigma)/(3e^2)$. In good metallic conductors of high purity, the electronic thermal conductivity dominates and the phonon contribution tends to be small. On the other hand, in conductors where the thermal conductivity due to phonons makes a significant contribution to the total thermal conductivity, it is necessary to separate the electronic and lattice contributions before applying the Wiedemann–Franz law to the total κ.

With regard to the lattice contribution, κ_L at very low temperatures is dominated by defect and boundary scattering processes. From the relation

$$\kappa_L = \frac{1}{3} C_p v_q \Lambda_{\rm ph} \tag{6.60}$$

we can determine the temperature dependence of κ_L, since $C_p \sim T^3$ at low T while the sound velocity v_q and phonon mean free path $\Lambda_{\rm ph}$ at very low T are independent of T. In this regime the number of scatterers is independent of T.

In the regime where only low q phonons contribute to transport and to scattering, only normal scattering processes contribute. In this regime C_p is still increasing as T^3, v_q is independent of T, but $1/\Lambda_{\rm ph}$ increases in proportion to the phonon density of states. With increasing T, the temperature dependence of C_p becomes less pronounced and that for $\Lambda_{\rm ph}$ becomes more pronounced as more scatters participate, leading eventually to a decrease in κ_L. We note that it is only the inelastic collisions that contribute to the decrease in $\Lambda_{\rm ph}$, since elastic phonon-phonon scattering has a similar effect as impurity scattering for phonons. The inelastic collisions are of course due to anharmonic forces.

Eventually phonons with wavevectors large enough to support umklapp processes are thermally activated. Umklapp processes give rise to thermal resistance and in this regime κ_L decreases as $\exp(-\Theta_D/T)$. In the high temperature limit $T \gg \Theta_D$, the heat capacity and phonon velocity are both independent of T. The $\kappa_L \sim 1/T$ dependence arises from the $1/T$ dependence of the mean free path, since in this limit the scattering rate becomes proportional to $k_B T$.

Chapter 7

Magneto-transport Phenomena

References:

- Ashcroft and Mermin, *Solid State Physics*, Holt, Rinehart and Winston, 1976, Chapter 12.
- Pippard, *Magnetoresistance in Metals*, Cambridge University Press, 1989
- Kittel, *Introduction to Solid State Physics*, 7^{th} Ed., Wiley, 1996, Chapter 6.

Since the electrical conductivity is sensitive to the product of the carrier density and the carrier mobility rather than each of these quantities independently as shown in Eq. 4.91, it is necessary to look for different transport techniques to provide information on the carrier density n and the carrier mobility μ separately. Magneto-transport provides us with such techniques, at least for simple cases, since the magnetoresistance is sensitive to the carrier mobility and the Hall effect is sensitive to the carrier density. In this chapter we consider magneto-transport in bulk solids. We return to the discussion of magneto-transport for lower dimensional systems later in the course particularly with regard to the quantum Hall effect and giant magnetoresistance.

7.1 Magneto-transport in the classical regime ($\omega_c\tau < 1$)

The magnetoresistance and Hall effect measurements, which are used to characterize semiconductors, are made in the weak magnetic field limit $\omega_c\tau \ll 1$ where the cyclotron frequency is given by

$$\omega_c = eB/(m^*c). \tag{7.1}$$

The cyclotron frequency ω_c is the angular frequency of rotation of a charged particle as it makes an orbit in a plane perpendicular to the magnetic field. In this chapter we explain the origin of magneto-transport effects and provide some insight into their measurement.

In the low field limit (defined by $\omega_c\tau \ll 1$) the carriers are scattered long before completing a single cyclotron orbit in real space, so that quantum effects are unimportant. In higher fields where $\omega_c\tau > 1$, quantum effects become important. In this limit (discussed in Part III of this course), the electrons complete cyclotron orbits and the resonance achieved

by tuning the microwave frequency of a resonant cavity to coincide with ω_c allows us to measure the effective mass of electrons in semiconductors.

A simplified version of the magnetoresistance phenomenon can be obtained in terms of the $\vec{F} = m\vec{a}$ approach and is presented in §7.1.1. The virtue of the simplified approach is to introduce the concept of the Hall field and the general form of the magneto-conductivity tensor. A more general version of these results will then be given using the Boltzmann equation formulation (§7.3). The advantage of the more general derivation is to put the derivation on a firmer foundation and to distinguish between the various effective masses which enter the transport equations: the cyclotron effective mass of Eq. 7.1, the longitudinal effective mass along the magnetic field direction, and the dynamical effective mass which describes transport in an electric field (see §7.5).

7.1.1 Classical Magneto-transport Equations

For the simplified $\vec{F} = m\vec{a}$ treatment, let the magnetic field $\vec{B}$ be directed along the z direction. Then writing $\vec{F} = m\vec{a}$ for the electronic motion in the plane perpendicular to $\vec{B}$ we obtain

$$\vec{F} = e(\vec{E} + \vec{v} \times \vec{B}/c) = m^*\dot{\vec{v}} + m^*\vec{v}/\tau \tag{7.2}$$

where $m^*\vec{v}/\tau$ is introduced to account for damping or electron scattering. For static electric and magnetic fields, there is no time variation in the problem so that $\dot{\vec{v}} = 0$ and thus the equation of motion (Eq.7.2) reduces to

$$\begin{aligned} m^*v_x/\tau &= e(E_x + v_yB/c) \\ m^*v_y/\tau &= e(E_y - v_xB/c) \end{aligned} \tag{7.3}$$

which can be written as

$$\frac{m^*}{\tau}(v_x + iv_y) = e(E_x + iE_y) - i\frac{eB}{c}(v_x + iv_y), \tag{7.4}$$

where i is the unit imaginary, so that $j_x + ij_y = ne(v_x + iv_y)$ becomes

$$(j_x + ij_y) = \left(\frac{ne^2\tau}{m^*}\right)\frac{(E_x + iE_y)}{1 + i\omega_c\tau} \tag{7.5}$$

where the cyclotron frequency is defined by Eq. 7.1. The unit imaginary i is introduced into Eqs. 7.4 and 7.5 because of the circular motion of the electron orbit in a magnetic field, suggesting circular polarization for fields and velocities.

Equating real and imaginary parts of Eq. 7.5 yields

$$\begin{aligned} j_x &= \left(\tfrac{ne^2\tau}{m^*}\right)\left[\tfrac{E_x}{1+(\omega_c\tau)^2} + \tfrac{\omega_c\tau E_y}{1+(\omega_c\tau)^2}\right] \\ j_y &= \left(\tfrac{ne^2\tau}{m^*}\right)\left[\tfrac{E_y}{1+(\omega_c\tau)^2} - \tfrac{\omega_c\tau E_x}{1+(\omega_c\tau)^2}\right]. \end{aligned} \tag{7.6}$$

Since $\vec{v} = v_z\hat{z}$ is parallel to $\vec{B}$ or $(\vec{v} \times \vec{B}) = 0$, the motion of an electron along the magnetic field experiences no force due to the magnetic field, so that

$$j_z = \frac{ne^2\tau}{m^*}E_z. \tag{7.7}$$

Equations 7.6 and 7.7 yield the magnetoconductivity tensor defined by $\vec{j} = \overleftrightarrow{\sigma}_B \cdot \vec{E}$ in the presence of a magnetic field in the low field limit where $\omega_c\tau \ll 1$ and the classical approach given here is applicable. In this limit an electron in a magnetic field is accelerated by an electric field and follows Ohm's law (as in the case of zero magnetic field):

$$\vec{j} = \overleftrightarrow{\sigma}_B \cdot \vec{E} \tag{7.8}$$

except that the magnetoconductivity tensor $\overleftrightarrow{\sigma}_B$ depends explicitly on magnetic field and in accordance with Eqs. 7.6 and 7.7 assumes the form

$$\overleftrightarrow{\sigma}_B = \frac{ne^2\tau/m^*}{1+(\omega_c\tau)^2} \begin{pmatrix} 1 & \omega_c\tau & 0 \\ -\omega_c\tau & 1 & 0 \\ 0 & 0 & 1+(\omega_c\tau)^2 \end{pmatrix}. \tag{7.9}$$

The magnetoresistivity tensor (which is more closely related to laboratory measurements) is defined as the inverse of the magnetoconductivity tensor

$$\overleftrightarrow{\rho}_B = [\overleftrightarrow{\sigma}_B]^{-1} = \frac{m^*}{ne^2\tau} \begin{pmatrix} 1 & -\omega_c\tau & 0 \\ \omega_c\tau & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{7.10}$$

7.1.2 Magnetoresistance

The magnetoresistance is defined in terms of the diagonal components of the magnetoresistivity tensor given by Eq. 7.10

$$\Delta\rho/\rho \equiv \Big(\rho(B) - \rho(0)\Big)/\rho(0) \tag{7.11}$$

and, in general, depends on $(\omega_c\tau)^2$ or on B^2. Since $\omega_c\tau = (e\tau/m^*c)B = \mu B/c$, the magnetoresistance provides information on the carrier mobility.

The longitudinal magnetoresistivity $\Delta\rho_{zz}/\rho_{zz}$ is measured with the electric field parallel to the magnetic field. On the basis of a spherical Fermi surface one carrier model, we have $E_z = j_z/\sigma_0$ from Eq. 7.10, so that there is no longitudinal magnetoresistivity in this case; that is, the resistivity is the same whether or not a magnetic field is present, since $\sigma_0 = ne^2\tau/m^*$. On the other hand, many semiconductors do exhibit longitudinal magnetoresistivity experimentally, and this effect arises from the non-spherical shape of their constant energy surfaces.

The transverse magnetoresistivity $\Delta\rho_{xx}/\rho_{xx}$ is measured with the current flowing in some direction (x) perpendicular to the magnetic field. With the direction of current flow along the x direction, then $j_y = 0$ and we can write from Eqs. 7.8 and 7.9 as

$$E_y = (\omega_c\tau)E_x \tag{7.12}$$

so that

$$j_x = \sigma_0 \left[\frac{E_x}{1+(\omega_c\tau)^2} + \frac{(\omega_c\tau)^2 E_x}{1+(\omega_c\tau)^2} \right] = \sigma_0 E_x \tag{7.13}$$

and again there is no transverse magnetoresistance for a material with a single carrier type having a spherical Fermi surface. Introduction of either a more complicated Fermi surface or more than one type of carrier results in a transverse magnetoresistance. When the velocity distribution of carriers at a finite temperature is taken into account, a finite transverse magnetoresistance is also obtained. In a similar way, multi-valley semiconductors (having several electron or hole constant energy surfaces some of which are equivalent by symmetry) can also display a transverse magnetoresistance. In all of these cases the magnetoresistance exhibits a B^2 dependence. The effect of two carrier types on the transverse magnetoresistance is discussed in §7.4 in some detail.

We note that the σ_{xy} and σ_{yx} terms arise from the presence of a magnetic field. The significance of these terms is further addressed in our discussion of the Hall effect (§7.2). We note that for non-spherical constant energy surfaces, Eqs. 7.6 and 7.7 must be rewritten to reflect the fact that m^* is a tensor so that $\vec{v}$ and $\vec{E}$ need not be parallel, even in the absence of a magnetic field (see Fig. 4.4). This point is clarified to some degree in the derivation of magneto-transport effects given in §7.3 using the Boltzmann Equation.

7.2 The Hall Effect

If an electric current is flowing in a semiconductor transverse to an applied magnetic field, an electric field is generated perpendicular to the plane containing $\vec{j}$ and $\vec{B}$. This is known as the Hall effect. Because the magnetic field acts to deflect the charge carriers transverse to their current flow, the Hall field is required to ensure that the transverse current vanishes. Let x be the direction of current flow and z the direction of the magnetic field. Then the boundary condition for the Hall effect is $j_y = 0$. From the magnetoconductivity tensor Eq. 7.9 we have

$$j_y = \frac{ne^2\tau}{m^*} \left(\frac{1}{1+(\omega_c\tau)^2} \right) (E_y - \omega_c\tau E_x) \tag{7.14}$$

so that a non-vanishing Hall field

$$E_y = \omega_c\tau E_x \tag{7.15}$$

must be present to ensure the vanishing of j_y (see Fig. 7.1). It is convenient to define the Hall coefficient

$$R_{\text{Hall}} \equiv \frac{E_y}{j_x B_z} = \frac{\tau E_x (eB_z/m^*c)}{j_x B_z}. \tag{7.16}$$

Substitution of the Hall field into the expression for j_x then yields

$$j_x = \frac{ne^2\tau}{m^*[1+(\omega_c\tau)^2]} [E_x + \omega_c\tau E_y] = \frac{ne^2\tau[1+(\omega_c\tau)^2]E_x}{m^*[1+(\omega_c\tau)^2]} \tag{7.17}$$

or

$$j_x = \frac{ne^2\tau}{m^*} E_x = \sigma_{dc} E_x. \tag{7.18}$$

Substitution of this expression into the Hall coefficient yields

$$R_{\text{Hall}} = \frac{e\tau}{m^* c(ne^2\tau/m^*)} = \frac{1}{nec}. \tag{7.19}$$

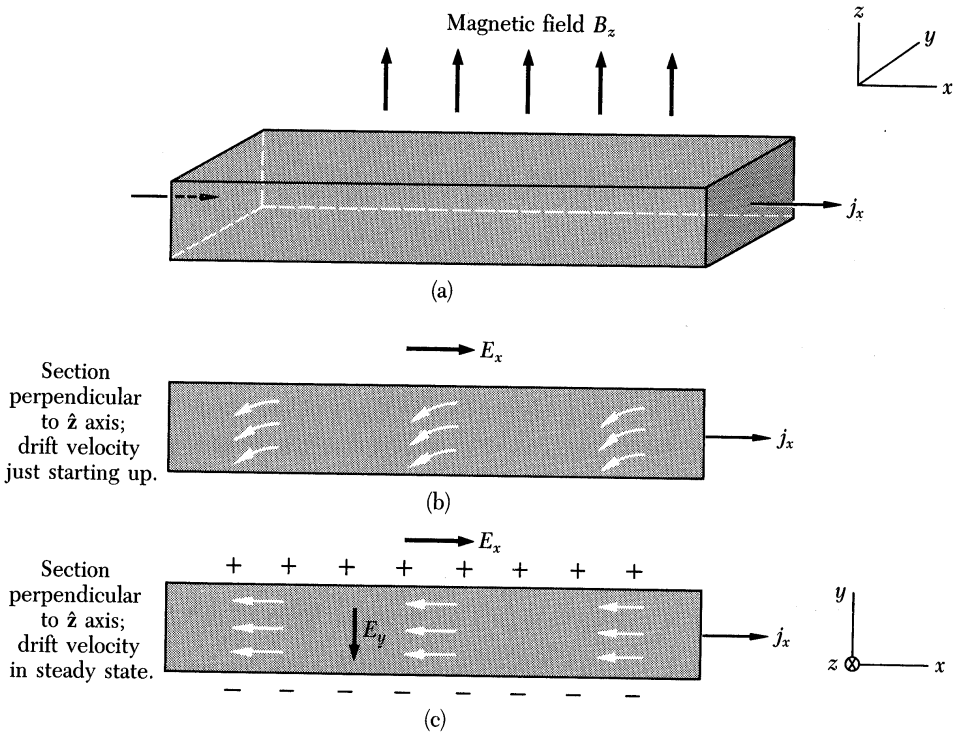


Figure 7.1: The standard geometry for the Hall effect: a specimen of rectangular cross-section is placed in a magnetic field B_z as in (a). An electric field E_x applied across the end electrodes causes an electric current density j_x to flow down the bar. The drift velocity of the electrons immediately after the electric field is applied is shown in (b). The deflection in the y direction is caused by the magnetic field. Electrons accumulate on one face of the bar and a positive ion excess is established on the opposite face until, as in (c), the transverse electric field (Hall field) just cancels the force due to the magnetic field.

The Hall coefficient is important because:

1. $R_{\rm Hall}$ depends only on the carrier density n, aside from universal constants
2. The sign of $R_{\rm Hall}$ determines whether conduction is by electrons ($R_{\rm Hall} < 0$) or by holes ($R_{\rm Hall} > 0$).

If the carriers are of one type we can relate the Hall mobility $\mu_{\rm Hall}$ to $R_{\rm Hall}$:

$$\mu = \frac{e\tau}{m^*} = \left(\frac{ne^2\tau}{m^*}\right) c \left(\frac{1}{nec}\right) = \sigma c R_{\rm Hall} \tag{7.20}$$

We define

$$\mu_{\rm Hall} \equiv \sigma c R_{\rm Hall} \tag{7.21}$$

and $\mu_{\rm Hall}$ carries the same sign as $R_{\rm Hall}$. The resistivity component $\rho_{xy} = ne\mu_{\rm Hall}$ is called the Hall resistivity.

A variety of new effects can occur in $R_{\rm Hall}$ when there is more than one type of carrier, as is commonly the case in semiconductors, and this is discussed in §7.4.

7.3 Derivation of the Magneto-transport Equations from the Boltzmann Equation

The corresponding results relating $\vec{j}$ and $\vec{E}$ will now be found using the Boltzmann equation (Eq. 4.4) in the absence of temperature gradients. We first use the linearized Boltzmann equation given by Eq. 4.18 to obtain the distribution function f_1. Then we will use f_1 to obtain the current density $\vec{j}$ in the presence of an electric field $\vec{E}$ and a magnetic field $\vec{B} = B\hat{z}$ in the z-direction.

In the presence of a magnetic field the equation of motion becomes

$$\hbar\dot{\vec{k}} = e\left(\vec{E} + \frac{1}{c}\vec{v} \times \vec{B}\right). \tag{7.22}$$

We use, as in Eq. 4.17, $f = f_0 + f_1$ with

$$\frac{\partial f_0}{\partial \vec{k}} = \frac{\partial f_0}{\partial E}\frac{\partial E(\vec{k})}{\partial \vec{k}} = \hbar\vec{v}\frac{\partial f_0}{\partial E}. \tag{7.23}$$

Substituting into the linearized form of the Boltzmann equation (Eq. 4.18) gives an equation for f_1:

$$\frac{e}{\hbar}(\vec{E} + \frac{1}{c}\vec{v} \times \vec{B})\cdot\left(\hbar\vec{v}\frac{\partial f_0}{\partial E} + \frac{\partial f_1}{\partial \vec{k}}\right) = -\frac{f_1}{\tau}. \tag{7.24}$$

In analogy with the case of zero field, we assume a solution for f_1 of the form

$$f_1 = -e\tau\vec{v}\cdot\vec{V}\frac{\partial f_0}{\partial E} \tag{7.25}$$

with (see Eq. 4.21) where $\vec{V}$ is a vector to be determined in analogy with the solution for $B = 0$. The form of Eq. 7.25 is motivated by the form suggested by the magnetoconductivity tensor in Eq. 7.9.

For a simple parabolic band, $\vec{v} = \hbar\vec{k}/m^*$, and substitution of Eq. 4.21 into Eq. 7.24 gives

$$\frac{e}{\hbar c}(\vec{v}\times\vec{B})\cdot\frac{\partial f_1}{\partial\vec{k}} = -\frac{e^2\tau}{m^*c}(\vec{v}\times\vec{B})\cdot\vec{V}\frac{\partial f_0}{\partial E}. \tag{7.26}$$

The following equation for $\vec{V}$ is then obtained from Eq. 7.24

$$\vec{v}\cdot\vec{E} - \frac{e\tau}{m^*c}(\vec{v}\times\vec{B})\cdot\vec{V} = \vec{v}\cdot\vec{V} \tag{7.27}$$

where we have neglected a term $\vec{E}\cdot\vec{V}$ which is small (of order $|\vec{E}|^2$ if $|\vec{E}|$ is small). Equation 7.27 is equivalent to

$$\begin{aligned} v_xE_x + \omega_c\tau v_xV_y &= v_xV_x \\ v_yE_y - \omega_c\tau v_yV_x &= v_yV_y \end{aligned} \tag{7.28}$$

which can be rewritten more compactly as:

$$\begin{aligned} \vec{V}_\perp &= \Big(\vec{E}_\perp - (e\tau/m^*c)[\vec{B}\times\vec{E}_\perp]\Big)\Big(1+(e\tau B/m^*c)^2\Big)^{-1} \\ V_z &= E_z \end{aligned} \tag{7.29}$$

where the notation "$\perp$" in Eq. 7.29 denotes the component in the $x-y$ plane, perpendicular to $\vec{B}$. This solves the problem of finding f_1.

Now we can carry out the calculation of $\vec{j}$ in Eq. 4.23, using the new expression for f_1 given by Eqs. 7.25, 4.13, and 7.29. With the more detailed calculation using the Boltzmann equation, it is clear that the cyclotron mass governs the cyclotron frequency while the dynamic effective mass controls the coefficients ($ne^2\tau/m^*$) in Eqs. 7.6 and 7.7. We discuss in §7.5 how to calculate the cyclotron effective mass.

7.4 Two Carrier Model

We calculate both the Hall effect and the transverse magnetoresistance for a two-carrier model. Referring to Fig. 7.1, the geometry under which transport measurements are made ($\vec{j} \parallel \hat{x}$) imposes the condition $j_y = 0$. From the magnetoconductivity tensor of Eq. 7.9

$$j_y = -\frac{\sigma_{01}\beta_1E_x}{1+\beta_1^2} + \frac{\sigma_{01}E_y}{1+\beta_1^2} - \frac{\sigma_{02}\beta_2E_x}{1+\beta_2^2} + \frac{\sigma_{02}E_y}{1+\beta_2^2} = 0 \tag{7.30}$$

where

$$\beta = \omega_c\tau \tag{7.31}$$

and the subscripts on σ_{0i} and β_i refer to the carrier index, $i = 1, 2$, so that $\sigma_{0i} = n_ie^2\tau_i/m_i^*$ and $\beta_i = \omega_{0c}\tau_i$. Solving Eq. 7.30 yields a relation between E_y and E_x which defines the Hall field

$$E_y = E_x\left[\frac{\frac{\sigma_{01}\beta_1}{1+\beta_1^2} + \frac{\sigma_{02}\beta_2}{1+\beta_2^2}}{\frac{\sigma_{01}}{1+\beta_1^2} + \frac{\sigma_{02}}{1+\beta_2^2}}\right] \tag{7.32}$$

for a two carrier system. This basic equation is applicable to two kinds of electrons, two kinds of holes, or a combination of electrons and holes. The generalization of Eq. 7.30 to

more than two types of carriers is immediate. The magnetoconductivity tensor is found by substitution of Eq. 7.32 into:

$$j_x = \frac{\sigma_{01}E_x}{1+\beta_1^2} + \frac{\sigma_{01}\beta_1 E_y}{1+\beta_1^2} + \frac{\sigma_{02}E_x}{1+\beta_2^2} + \frac{\sigma_{02}\beta_2 E_y}{1+\beta_2^2} \tag{7.33}$$

In general Eq. 7.33 is a complicated relation, but simplifications can be made in the low field limit $\beta \ll 1$, where we can neglect terms in β^2 relative to terms in β. Retaining the lowest power in terms in β then yields

$$E_y = E_x \left[\frac{\sigma_{01}\beta_1 + \sigma_{02}\beta_2}{\sigma_{01}+\sigma_{02}}\right] \tag{7.34}$$

and

$$j_x = (\sigma_{01}+\sigma_{02})E_x. \tag{7.35}$$

We thus obtain the following important relation for the Hall coefficient which is independent of magnetic field in this limit

$$R_{\rm Hall} \equiv \frac{E_y}{j_x B_z} = \frac{\beta_1\sigma_{01}+\beta_2\sigma_{02}}{(\sigma_{01}+\sigma_{02})^2 B} = \frac{\mu_1\sigma_{01}+\mu_2\sigma_{02}}{c(\sigma_{01}+\sigma_{02})^2} \tag{7.36}$$

where we have made use of the relation between β and the mobility μ

$$\beta = \mu B/c = e\tau B/(m_c^* c) = \omega_c \tau. \tag{7.37}$$

This allows us to write $R_{\rm Hall}$ in terms of the Hall coefficients R_i for each of the two types of carriers

$$R_{\rm Hall} = \frac{R_1\sigma_{01}^2 + R_2\sigma_{02}^2}{(\sigma_{01}+\sigma_{02})^2} \tag{7.38}$$

since

$$\frac{\beta}{B} = R_{\rm Hall}\sigma, \tag{7.39}$$

where for each carrier type we have

$$\sigma_{0i} = \frac{n_i e^2 \tau_i}{m_i^*} \tag{7.40}$$

and

$$R_i = \frac{1}{n_i e_i c} \tag{7.41}$$

where $i = 1, 2$. We note in Eq. 7.41 that $e_i = \pm|e|$ where $|e|$ is the magnitude of the charge on the electron. Thus electrons and holes contribute with opposite sign to $R_{\rm Hall}$ in Eq. 7.38. When more than one carrier type is present, it is not always the case that the sign of the Hall coefficient is the same as the sign of the majority carrier type. A minority carrier type may have a higher mobility, and the carriers with high mobility make a larger contribution per carrier to $R_{\rm Hall}$ than do the low mobility carriers.

The magnetoconductivity for two carrier types is obtained from Eq. 7.33 upon substitution of Eq. 7.34 into Eq. 7.33 and retaining terms in β^2. For the transverse magnetoconductance we obtain

$$\frac{\sigma_B(B)-\sigma_B(0)}{\sigma_B(0)} = \frac{2\sigma_{01}^2\beta_1^2 + 2\sigma_{02}^2\beta_2^2 + \sigma_{01}\sigma_{02}(\beta_1+\beta_2)^2}{(\sigma_{01}+\sigma_{02})^2} \tag{7.42}$$

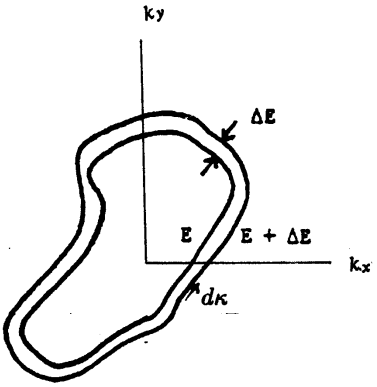


Figure 7.2: Contour of the constant energy surface in k space used to calculate the path integral in the evaluation of the cyclotron effective mass.

which is an average of β_1 and β_2 weighted by conductivity components σ_{01} and σ_{02}. But since $\Delta\rho/\rho = -\Delta\sigma/\sigma$ we obtain the following result for the transverse magnetoresistance

$$\frac{\Delta\rho}{\rho} = -\frac{2\sigma_{01}^2\beta_1^2 + 2\sigma_{02}^2\beta_2^2 + \sigma_{01}\sigma_{02}(\beta_1+\beta_2)^2}{(\sigma_{01}+\sigma_{02})^2}. \tag{7.43}$$

We note that the magnetoconductivity tensor (Eq. 7.9) yields no longitudinal magnetoresistance for a spherical two-carrier model.

7.5 Cyclotron Effective Mass

To calculate the magnetoresistance and Hall effect explicitly for non-spherical Fermi surfaces, we need to derive a formula for the cyclotron frequency $\omega_c = eB/(m_c^* c)$ which is generally applicable for non-spherical Fermi surfaces. The cyclotron effective mass can be determined in either of two ways. The first method is the *tube integral* method (see Fig. 7.2 for a schematic of the constant energy surfaces at energy E and $E + \Delta E$) which defines the cyclotron effective mass as

$$m_c^* = \frac{1}{2\pi}\oint \frac{\hbar d\kappa}{|v|} = \frac{\hbar^2}{2\pi}\oint \frac{d\kappa}{|\partial E/\partial k|} \tag{7.44}$$

where $d\kappa$ is an infinitesimal element of length along the contour and we can obtain m_c^* by direct integration. For the second method, we convert the line integral over an enclosed area, making use of $\Delta E = \Delta k(\partial E/\partial k)$ so that

$$m_c^* = \frac{\hbar^2}{2\pi}\frac{1}{\Delta E}\oint(\Delta k)d\kappa = \frac{\hbar^2}{2\pi}\frac{\Delta A}{\Delta E} \tag{7.45}$$

where ΔA is the area of the strip indicated in Fig. 7.2 by the separation ΔE. Therefore we obtain the relation

$$m_c^* = \frac{\hbar^2}{2\pi}\frac{\partial A}{\partial E} \tag{7.46}$$

which gives the second method for finding the cyclotron effective mass.

For a spherical constant energy surface, we have $A = \pi k^2$ and $E(\vec{k}) = (\hbar^2 k^2)/(2m^*)$ so that $m_c^* = m^*$. For an electron orbit described by an ellipse in reciprocal space (which is appropriate for the general orbit in the presence of a magnetic field on an ellipsoidal constant energy surface at wave vector k_B along the magnetic field) we write

$$E(\vec{k}_\perp) = \frac{\hbar^2 k_1^2}{2m_1} + \frac{\hbar^2 k_2^2}{2m_2} = \frac{\hbar^2 k_0^2}{2m_0} \tag{7.47}$$

which defines the area A enclosed by the constant energy surface as

$$A = \pi k_1 k_2 = \pi k_0^2 \sqrt{m_1 m_2}/m_0 \tag{7.48}$$

where $(k_i^2/m_i) = (k_0^2/m_0)$. Then substitution in Eq. 7.46 gives

$$m_c^* = \sqrt{m_1 m_2}. \tag{7.49}$$

This expression for m_c^* gives a clear physical picture of the relation between m_c^* and the electron orbit on a constant ellipsoidal energy surface in the presence of a magnetic field. Since finding the electron orbit requires geometrical calculation for a general magnetic field orientation, it is more convenient for computer computation to use the relation

$$m_c^* = \left(\frac{\det \overleftrightarrow{m^*}}{\hat{b} \cdot \overleftrightarrow{m^*} \cdot \hat{b}} \right)^{1/2} \tag{7.50}$$

for calculating m_c^* for ellipsoidal constant energy surfaces where $\hat{b} \cdot \overleftrightarrow{m^*} \cdot \hat{b}$ is the effective mass component along the magnetic field, $\det \overleftrightarrow{m^*}$ denotes the determinant of the effective mass tensor $\overleftrightarrow{m^*}$, and $\hat{b}$ is a unit vector along the magnetic field. One can show that for this case the Hall mobility is given by

$$\mu_{\text{Hall}} = \frac{e\tau}{m_c^*} \tag{7.51}$$

and

$$\mu_{\text{Hall}} \frac{B}{c} \equiv \omega_c \tau = \beta \tag{7.52}$$

so that μ_{Hall} involves the cyclotron effective mass.

7.6 Effective Masses for Ellipsoidal Fermi Surfaces

The effective mass of carriers in a magnetic field is complicated by the fact that several effective mass quantities are of importance. These include the cyclotron effective mass m_c^* for electron motion transverse to the magnetic field (§7.5) and the longitudinal effective mass m_B^* for electron motion along the magnetic field

$$m_B^* = \hat{b} \cdot \overleftrightarrow{m^*} \cdot \hat{b} \tag{7.53}$$

obtained by projecting the effective mass tensor along the magnetic field. These motions are considered in finding f_1, the change in the electron distribution function due to forces and fields.

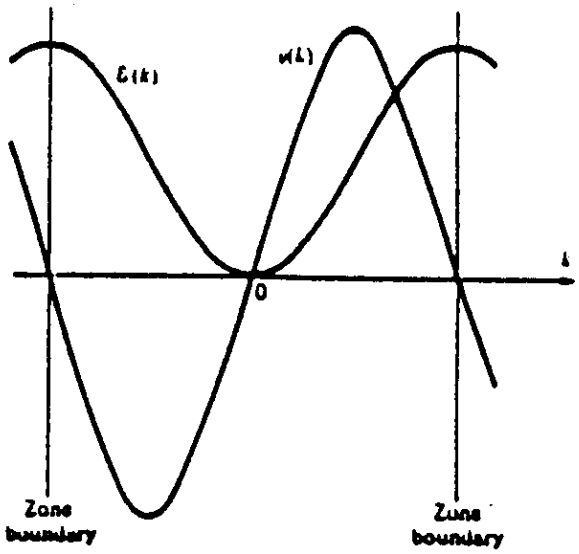


Figure 7.3: Schematic diagram of $E(\vec{k})$ and of the velocity $v(\vec{k})$ which is proportional to the derivative $\partial E(\vec{k})/\partial \vec{k}$ for an electron in a nearly free electron model.

Returning to Eqs. 4.9 and 4.10 the initial exposition for the current density calculated by the Boltzmann equation, we obtained the Drude formula

$$\overleftrightarrow{\sigma} = ne^2\tau \left(\overleftrightarrow{\frac{1}{m^*}}\right). \tag{7.54}$$

Thereby defining the drift mass tensor in an electric field. Referring to the magnetoresistance and magneto-conductance tensors (Eqs. 7.8 and 7.9), we see the drift term $(ne^2\tau/m^*)$ which utilizes the drift mass tensor and terms in $(\omega_c\tau)$ which utilizes the effective cyclotron mass m_c^* (see §7.5). Where the Fermi surface for a semiconductor consists of ellipsoidal carrier pockets, then the drift effective mass components are found in accordance with the procedure outlined in §4.5.1 for ellipsoidal carrier pockets. We conveniently use Eq. 7.50 to determine the cyclotron effective mass for ellipsoidal carrier pockets.

7.7 Dynamics of Electrons in a Magnetic Field

In this section we relate the electron motion on a constant energy surface in a magnetic field to real space orbits. Consider first the case of $B = 0$ shown in Fig. 7.3. At a given k value, $E(\vec{k})$ and $v(\vec{k})$ are specified and each of the quantities is a constant of the motion, where $\vec{v}(\vec{k}) = (1/\hbar)[\partial E(\vec{k})/\partial \vec{k}]$. If there are no forces on the system, $E(\vec{k})$ and $v(\vec{k})$ are unchanged with time. Thus, at any instant of time there is an equal probability that an electron will be found anywhere on a constant energy surface. The role of an external electric field $\vec{E}$ is to change the $\vec{k}$ vector on this constant energy surface according to the equation of motion

$$\hbar\dot{\vec{k}} = e\vec{E} \tag{7.55}$$

so that under a force $e\vec{E}$ the energy of the system is changed.

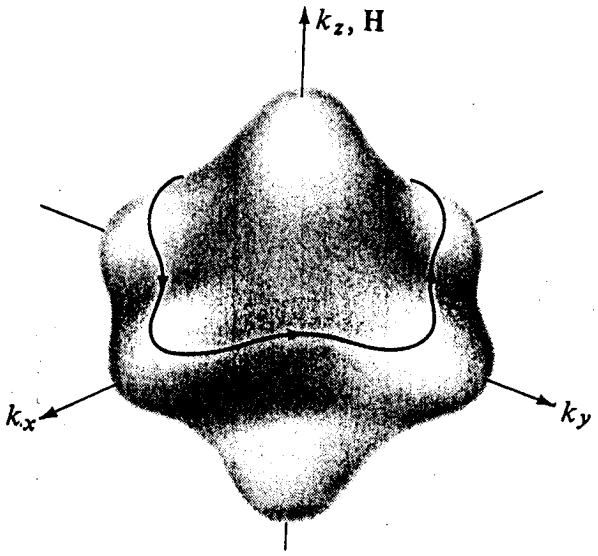


Figure 7.4: Schematic diagram of the motion of an electron along a constant energy (and constant k_z) trajectory in the presence of a magnetic field in the z-direction.

In a constant magnetic field (no electric field), the electron will move on a constant energy surface in $\vec{k}$ space in an orbit perpendicular to the magnetic field (see Fig. 7.4) and following the equation of motion

$$\hbar\dot{\vec{k}} = \frac{e}{c}(\vec{v} \times \vec{B}) \tag{7.56}$$

where we note that $|v_\perp|$ remains unchanged along the electron orbit. The electrons will execute the indicated orbit at a cyclotron frequency ω_c given by $\omega_c = eB/m_c^* c$ where m_c^* is the cyclotron effective mass (see §7.5).

For high magnetic fields, when $\omega_c \tau \gg 1$, an electron circulates many times around its semiclassical orbit before undergoing a collision. In this limit, there is interest in describing the orbits of carriers in $\vec{r}$ space. The solution to the semiclassical equations

$$\dot{\vec{r}} = \vec{v} = (1/\hbar)\partial E/\partial \vec{k} \tag{7.57}$$

$$\hbar\dot{\vec{k}} = e[\vec{E} + (1/c)\vec{v} \times \vec{B}] \tag{7.58}$$

is

$$\vec{r}_\perp = \vec{r} - \frac{\vec{B}}{B^2}(\vec{B} \cdot \vec{r}) \tag{7.59}$$

in which $\vec{r} = \vec{r}_\parallel + \vec{r}_\perp$. We also note that

$$\begin{aligned} \vec{B} \times \hbar\dot{\vec{k}} &= e\Big[\vec{B} \times \vec{E} + (1/c)\vec{B} \times (\vec{v} \times \vec{B})\Big] \\ &= e\Big[\vec{B} \times \vec{E} + (1/c)(B^2\vec{v} - (\vec{B} \cdot \vec{v})\vec{B})\Big]. \end{aligned} \tag{7.60}$$

Making use of Eqs. 7.59 and 7.60, we may write

$$\dot{\vec{r}}_\perp = \frac{c\hbar}{eB^2}\vec{B} \times \dot{\vec{k}} + \frac{c}{B^2}(\vec{E} \times \vec{B}) \tag{7.61}$$

which upon integration yields

$$\vec{r}_\perp(t) - \vec{r}_\perp(0) = \frac{\hbar}{\omega_c B m_c^*}\, \vec{B} \times [\vec{k}(t) - \vec{k}(0)] + \vec{w}\, t \tag{7.62}$$

where

$$\vec{w} = \frac{c}{B^2}\, (\vec{E} \times \vec{B}). \tag{7.63}$$

From Eq. 7.62 we see that the orbit in real space is $\pi/2$ out of phase with the orbit in reciprocal space. For the case of closed orbits, after a long time $t \approx \tau$. Then the second term $\vec{w}\, t$ of Eq. 7.62 will dominate, giving a transverse or Hall current

$$\vec{j}_\perp \rightarrow \frac{ne}{B^2}(\vec{E} \times \vec{B}) \tag{7.64}$$

where n is the electron density. Similarly the longitudinal current $\vec{j} \parallel \vec{E}$ will approach a constant value or saturate since $\vec{E} \perp \vec{j}$, and $\vec{E} \perp \vec{B}$.

The situation is very different for the magnetic and electric fields applied in special directions relative to the crystal axes. For these special direction open electron orbits can occur, as illustrated in Fig. 7.5 for copper. In this case $\vec{k}(t) - \vec{k}(0)$ has a component proportional to $\vec{E}t$ which is not negligible. When the term $[\vec{k}(t) - \vec{k}(0)]$ must be considered, it can be shown that the magnetoresistance does not saturate but instead increases as B^2.

Figure 7.6 shows the angular dependence of the magnetoresistance in copper, which exhibits both closed and open orbits depending on the direction of the magnetic field (see Fig. 7.6). The strong angular dependence is associated with the large difference in the magnitude of the magnetoresistance for closed and open orbits. Large values of $\omega_c\tau$ are needed to distinguish clearly between the open and closed orbits, thereby requiring the use of samples of very high purity and low temperature (e.g., 4.2 K) operation.

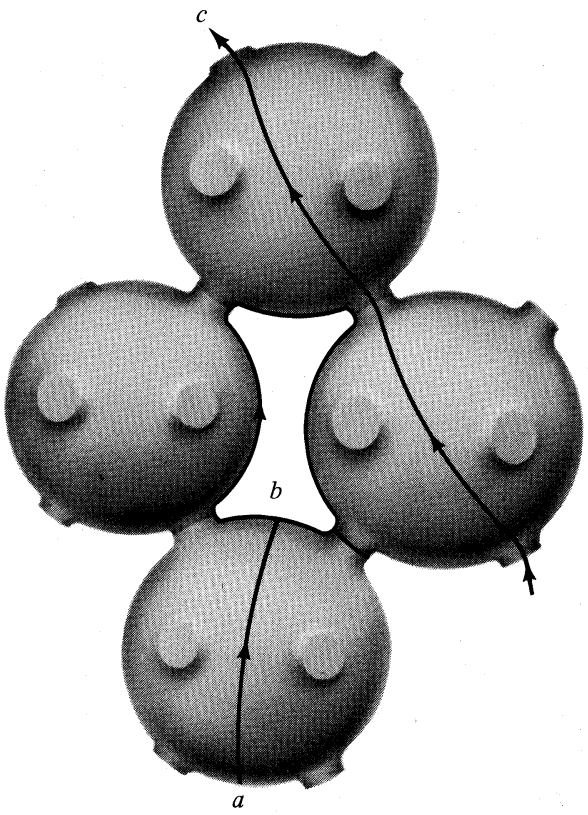


Figure 7.5: This diagram for the electron orbits in metallic copper indicates only a few of the many types of orbits an electron can pursue in k-space when a uniform magnetic field is applied to a noble metal. (Recall that the orbits are given by slicing the Fermi surface with planes perpendicular to the field.) The figure displays (a) a closed electron orbit; (b) a closed hole orbit; (c) an open hole orbit, which continues in the same general direction indefinitely in the repeated-zone scheme.

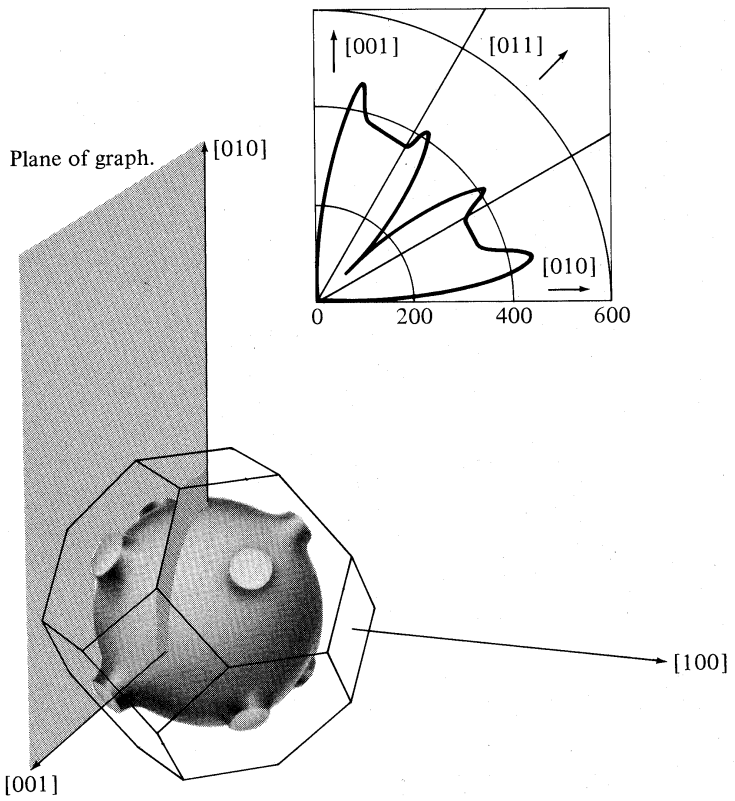


Figure 7.6: The spectacular directional dependence of the high-field magnetoresistance in copper is that characteristic of a Fermi surface supporting open orbits. The [001] and [010] directions of the copper crystal are as indicated in the figure, and the current flows in the [100] direction perpendicular to the graph. The magnetic field is in the plane of the graph. Its magnitude is fixed at 18 kilogauss, and its direction is varied continuously from [001] to [010]. The graph is a polar plot of the transverse magnetoresistance $[\rho(H) - \rho(0)]/\rho(0)$ vs. orientation of the field.

Chapter 9

Two Dimensional Electron Gas, Quantum Wells & Semiconductor Superlattices

References:

- Ando, Fowler and Stern, *Rev. Mod. Phys.* 54 437 (1982).
- R.F. Pierret, *Field Effect Devices*, Vol. IV of Modular Series on Solid State Devices, Addison-Wesley (1983).
- B.G. Streetman, *Solid State Electronic Devices*, Series in Solid State Physical Electronics, Prentice-Hall (1980).

9.1 Two-Dimensional Electronic Systems

One of the most important recent developments in semiconductors, both from the point of view of physics and for the purpose of device developments, has been the achievement of structures in which the electronic behavior is essentially two-dimensional (2D). This means that, at least for some phases of operation of the device, the carriers are confined in a potential such that their motion in one direction is restricted and thus is quantized, leaving only a two-dimensional momentum or k-vector which characterizes motion in a plane normal to the confining potential. The major systems where such 2D behavior has been studied are MOS structures, quantum wells and superlattices. More recently, quantization has been achieved in 1-dimension (the quantum wires) and "zero"–dimensions (the quantum dots). These topics are further discussed in Chapter 10 and in the course on semiconductor physics (6.735J).

9.2 MOSFETS

One of the most useful and versatile of these structures is the metal-insulator-semiconductor (MIS) layered structures, the most important of these being the metal-oxide-semiconductor

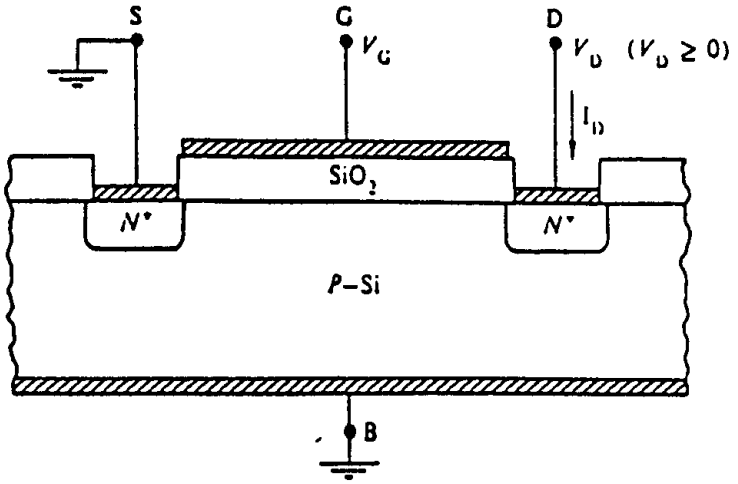


Figure 9.1: Cross-sectional view of the basic MOSFET structure showing the terminal designations and standard biasing conditions.

(MOS) structures. As shown in Fig. 9.1, the MOS device is fabricated from a substrate of usually moderately-doped p-type or n-type silicon which together with its grounded electrode is called the base and labeled B in the figure. On the top of the base is grown an insulating layer of silicon dioxide, followed by a metal layer; this structure is the gate (labeled G in the figure) and is used to apply an electric field through the oxide to the silicon. For the MOS device shown in the figure the base region is p-type and the source (S) and drain (D) regions are n-type. Measurements of the changes in the properties of the carriers in the silicon layer immediately below the gate (the conductance in the source-drain channel), in response to changes in the applied electric field at the gate electrode, are called field-effect measurements. As we show below, the field dramatically changes the conducting properties of the carriers beneath the gate. Use is made of this effect in the so-called metal-oxide-semiconductor field-effect transistor (MOSFET). To understand the operation of this device,we first consider the schematic energy band diagram of the MOS structure as shown in Fig. 9.2, for four different values of V_G, the gate potential relative to the substrate. For each V_G value, the diagram shows from left to right the metal (M) - oxide (O) - semiconductor (S) regions. In the semiconductor regions each of the diagrams show from top to bottom: the Si conduction band edge E_c, the "intrinsic" Fermi level for undoped Si as the dashed line, the Fermi level E_F in the p-type Si, and the valence band edge E_v. In each diagram, the central oxide region shows the valence band edge for the oxide. On the left hand side of each diagram, the Fermi level for the metal is shown and the dashed line gives the extension of the Si Fermi level. In the lower part of the figure, the charge layers of the interfaces for each case are illustrated.

We now explain the diagrams in Fig. 9.2 as a function of the gate voltage V_G. For $V_G = 0$ (the flat-band case), there are (ideally) no charge layers, and the energy levels of the metal (M) and semiconducting (S) regions line up to yield the same Fermi level (chemical potential). The base region is doped p-type. For a negative gate voltage ($V_G < 0$, the accumulation case), an electric field is set up in the oxide. The negative gate voltage

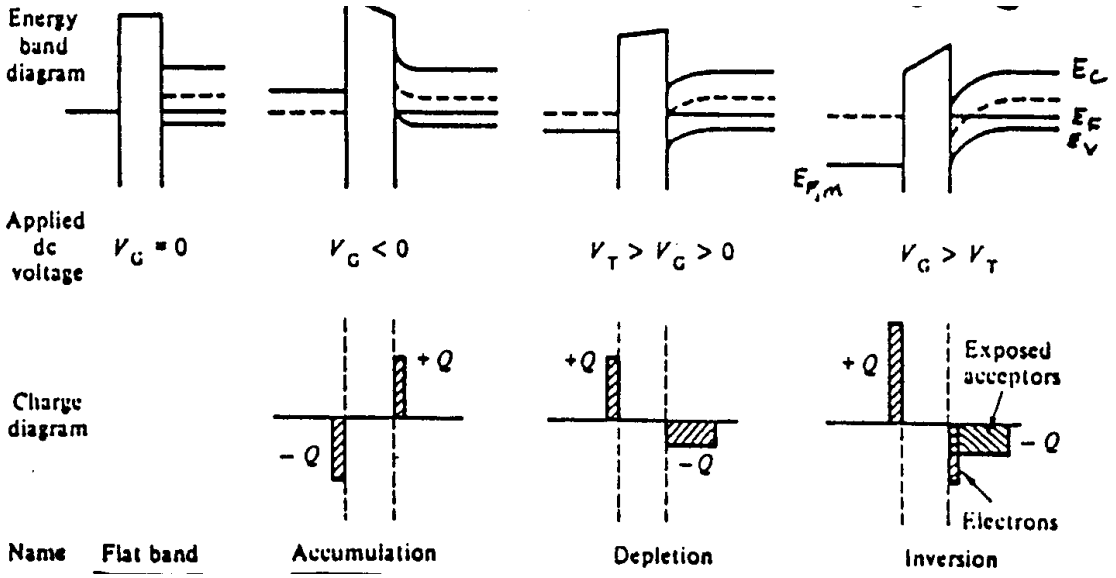


Figure 9.2: Energy band and block charge diagrams for a p–type device under flat band, accumulation, depletion and inversion conditions.

causes the Si bands to bend up at the oxide interface (see Fig. 9.2) so that the Fermi level is closer to the valence-band edge. Thus extra holes accumulate at the semiconductor-oxide interface and electrons accumulate at the metal-oxide interface (see lower part of Fig. 9.2). In the third (depletion) case, the gate voltage is positive but less than some threshold value V_T. The voltage V_T is defined as the gate voltage where the intrinsic Fermi level and the actual Fermi level are coincident at the interface (see lower part of Fig. 9.2). For the "depletion" regime, the Si bands bend down at the interface resulting in a depletion of holes, and a negatively charged layer of localized states is formed at the semiconductor-oxide interface. The size of this "depletion region" increases as V_G increases. The corresponding positively charged region at the metal-oxide interface is also shown. Finally, for $V_G > V_T$, the intrinsic Fermi level at the interface drops below the actual Fermi level, forming the "inversion layer", where mobile electrons reside. It is the electrons in this inversion layer which are of interest, both because they can be confined so as to exhibit two-dimensional behavior, and because they can be controlled by the gate voltage in the MOSFET (see Fig. 9.3)

The operation of a metal-oxide semiconductor field-effect transistor (MOSFET) is illustrated in Fig. 9.3, which shows the electron inversion layer under the gate for $V_G > V_T$ (for a p-type substrate), with the source region grounded, for various values of the drain voltage V_D. The inversion layer forms a conducting "channel" between the source and drain (as long as $V_G > V_T$). The dashed line in Fig. 9.3 shows the boundaries of the depletion region which forms in the p-type substrate adjoining the n^+ and p regions.

For $V_D = 0$ there is obviously no current between the source and the drain since both are at the same potential. For $V_D > 0$, the inversion layer or channel acts like a resistor, inducing the flow of electric current I_D. As shown in Fig. 9.3, increasing V_D imposes a reverse bias on the n^+-p drain-substrate junction, thereby increasing the width of the depletion region

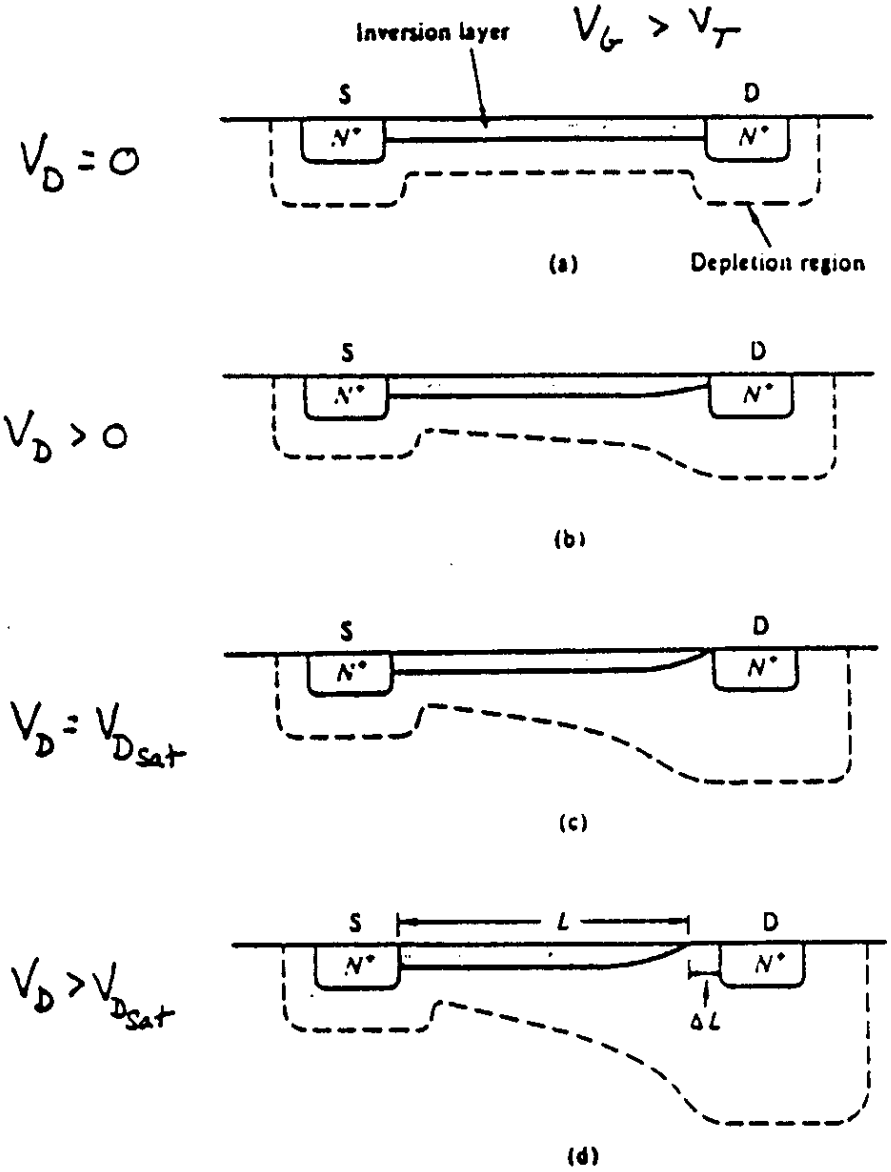


Figure 9.3: Visualization of various phases of $V_G > V_T$ MOSFET operation. (a) $V_D = 0$, (b) channel (inversion layer) narrowing under moderate V_D biasing, (c) pinch–off, and (d) post-pinch-off ($V_D > V_{Dsat}$) operation. (Note that the inversion layer widths, depletion widths, etc. are not drawn to scale.)

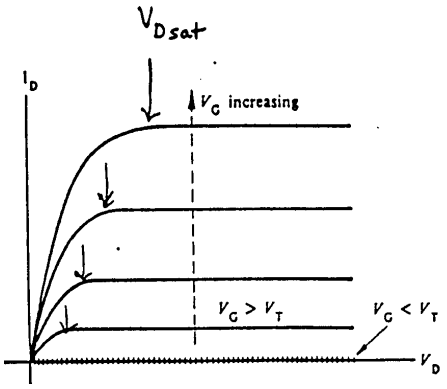


Figure 9.4: General form of the $I_D - V_D$ characteristics expected from a long channel ($\Delta L \ll L$) MOSFET.

and decreasing the number of carriers and narrowing the channel in the inversion layer as shown in Fig. 9.3. Finally as V_D increases further, the channel reaches the "pinched-off" condition V_{Dsat} shown in Fig. 9.3c. Further increase in V_D does not increase I_D but rather causes "saturation". We note that at saturation, $V_{Dsat} = V_G - V_T$. Saturation is caused by a decrease in the carrier density in the channel due to the pinch-off phenomena.

In Fig. 9.4 I_D vs. V_D curves are plotted for fixed values of $V_G > V_T$. We note that V_{Dsat} increases with increasing V_G. These characteristic curves are qualitatively similar to the curves for the bipolar junction transistor. The advantage of MOSFET devices lie in the speed of their operation and in the ease with which they can be fabricated into ultra-small devices.

The MOSFET device, or an array of a large number of MOSFET devices, is fabricated starting with a large Si substrate or "wafer". At each stage of fabrication, areas of the wafer which are to be protected are masked off using a light-sensitive substance called photoresist, which is applied as a thin film, exposed to light (or an electron or x-ray beam) through a mask of the desired pattern, then chemically developed to remove the photoresist from only the exposed (or, sometimes only the un–exposed) area. First the source and drain regions are formed by either diffusing or implanting (bombarding) donor ions into the p-type substrate. Then a layer of SiO_2 (which is an excellent and stable insulator) is grown by exposing the desired areas to an atmosphere containing oxygen; usually only a thin layer is grown over the gate regions and, in a separate step, thicker oxide layers are grown between neighboring devices to provide electrical isolation. Finally, the metal gate electrode, the source and drain contacts are formed by sputtering or evaporating a metal such as aluminum onto the desired regions.

9.3 Two-Dimensional Behavior

Other systems where two-dimensional behavior has been observed include heterojunctions of III-V compounds such as GaAs/$Ga_{1-x}Al_xAs$, layer compounds such as GaSe, $GaSe_2$ and related III-VI compounds, graphite and intercalated graphite, and electrons on the surface of liquid helium. The GaAs/$Ga_{1-x}Al_xAs$ heterojunctions are important for device

applications because the lattice constants and the coefficient of expansion of GaAs and $Ga_{1-x}Al_xAs$ are very similar. This lattice matching permits the growth of high mobility thin films of $Ga_{1-x}Al_xAs$ on a GaAs substrate.

The interesting physical properties of the MOSFET lie in the two-dimensional behavior of the electrons in the channel inversion layer at low temperatures. Studies of these electrons have provided important tests of modern theories of localization, electron-electron interactions and many-body effects. In addition, the MOSFETs have exhibited a highly unexpected property that, in the presence of a magnetic field normal to the inversion layer, the transverse or Hall resistance ρ_{xy} is quantized in integer values of e^2/h. This quantization is accurate to parts in 10^7 or 10^8 and provides the best measure to date of the fine structure constant $\alpha = e^2/hc$, when combined with the precisely-known velocity of light c. We will further discuss the quantized Hall effect later in the course (Part III).

We now discuss the two-dimensional behavior of the MOSFET devices in the absence of a magnetic field. The two-dimensional behavior is associated with the nearly plane wave electron states in the inversion layer. The potential $V(z)$ is associated with the electric field $V(z) = eEz$ and because of the negative charge on the electron, a potential well is formed containing bound states described by quantized levels. A similar situation occurs in the two–dimensional behavior for the case of electrons in quantum wells produced by molecular beam epitaxy. Explicit solutions for the bound states in quantum wells are given in §9.4. We discuss in the present section the form of the differential equation and of the resulting eigenvalues and eigenfunctions.

A single electron in a one-dimensional potential well $V(z)$ will, from elementary quantum mechanics, have discrete allowed energy levels E_n corresponding to bound states and usually a continuum of levels at higher energies corresponding to states which are not bound. An electron in a bulk semiconductor is in a three-dimensional periodic potential. In addition the potential causing the inversion layer of a MOSFET or a quantum well in $GaAs/Ga_{1-x}Al_xAs$ can be described by a one-dimensional confining potential $V(z)$ and can be written using the effective-mass theorem

$$[E(-i\vec{\nabla}) + \mathcal{H}']\Psi = i\hbar\left(\frac{\partial\Psi}{\partial t}\right) \tag{9.1}$$

where $\mathcal{H}' = V(z)$. The energy eigenvalues near the band edge can be written as

$$E(\vec{k}) = E(\vec{k}_0) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 E}{\partial k_i \partial k_j}\right)k_i k_j \tag{9.2}$$

so that the operator $E(-i\vec{\nabla})$ in Eq. 9.1 can be written as

$$E(-i\vec{\nabla}) = \sum_{i,j}\frac{p_i p_j}{2m_{i,j}} \tag{9.3}$$

where the p_i's are the operators

$$p_i = \frac{\hbar}{i}\frac{\partial}{\partial x_i} \tag{9.4}$$

which are substituted into Schrödinger's equation. The effect of the periodic potential is contained in the reciprocal of the effective mass tensor

$$\frac{1}{m_{ij}} = \frac{1}{\hbar^2}\frac{\partial^2 E(\vec{k})}{\partial k_i \partial k_j}\bigg|_{\vec{k}=\vec{k_0}} \tag{9.5}$$

where the components of $1/m_{ij}$ are evaluated at the band edge at $\vec{k}_0$.

If $1/m_{ij}$ is a diagonal matrix, the effective-mass equation $\mathcal{H}\Psi = E\Psi$ is solved by a function of the form

$$\Psi_{n,k_x,k_y} = e^{ik_x x} e^{ik_y y} f_n(z) \tag{9.6}$$

where $f_n(z)$ is a solution of the equation

$$-\frac{\hbar^2}{2m_{zz}} \frac{d^2 f_n}{dz^2} + V(z) f_n = E_{n,z} f_n \tag{9.7}$$

and the total energy is

$$E_n(k_x, k_y) = E_{n,z} + \frac{\hbar^2}{2m_{xx}} k_x^2 + \frac{\hbar^2}{2m_{yy}} k_y^2. \tag{9.8}$$

Since the $E_{n,z}$ energies (n=0,1,2,...) are discrete, the energies states $E_n(k_x, k_y)$ for each n value form a "sub-band". We give below (in §9.3.1) a simple derivation for the discrete energy levels by considering a particle in various potential wells (i.e., quantum wells). The electrons in these "sub–bands" form a 2D electron gas.

9.3.1 Quantum Wells and Superlattices

Many of the quantum wells and superlattices that are commonly studied today do not occur in nature, but rather are deliberately structured materials (see Fig. 9.5). In the case of superlattices formed by molecular beam epitaxy, the quantum wells result from the different bandgaps of the two constituent materials. The additional periodicity is in one–dimension (1–D) which we take along the z–direction, and the electronic behavior is usually localized on the basal planes (x–y planes) normal to the z–direction, giving rise to two–dimensional behavior.

A schematic representation of a semiconductor heterostructure superlattice is shown in Fig. 9.5 where d is the superlattice periodicity composed of a distance d_1, of semiconductor S_1, and d_2 of semiconductor S_2. Because of the different band gaps in the two semiconductors, potential wells and barriers are formed. For example in Fig. 9.5, the barrier heights in the conduction and valence bands are ΔE_c and ΔE_v respectively. In Fig. 9.5 we see that the difference in bandgaps between the two semiconductors gives rise to band offsets ΔE_c and ΔE_v for the conduction and valence bands. In principle, these band offsets are determined by matching the Fermi levels for the two semiconductors. In actual materials, the Fermi levels are highly sensitive to impurities, defects and charge transfer at the heterojunction interface.

The two semiconductors of a heterojunction superlattice could be different semiconductors such as InAs with GaP (see Table 9.1 for parameters related to these compounds) or a binary semiconductor with a ternary alloy semiconductor, such as GaAs with $Al_xGa_{1-x}As$ (sometimes referred to by their slang names "Gaas" and "Algaas"). In the typical semiconductor superlattices the periodicity $d = d_1 + d_2$ is repeated many times (e.g., 100 times). The period thicknesses typically vary between a few layers and many layers (10Å to 500Å). Semiconductor superlattices are today an extremely active research field internationally.

The electronic states corresponding to the heterojunction superlattices are of two fundamental types–bound states in quantum wells and nearly free electron states in zone–folded energy bands. In this course, we will limit our discussion to the bound states in a single infinite quantum well. Generalizations to multiple quantum wells will be made subsequently.

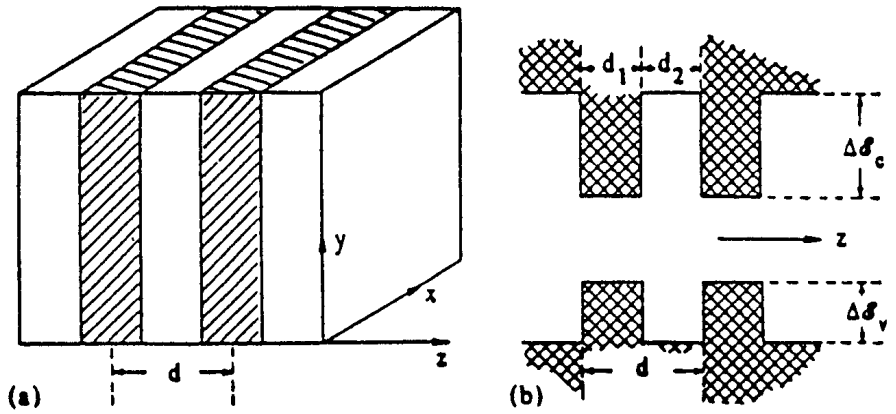


Figure 9.5: (a) A heterojunction superlattice of periodicity d. (b) Each superlattice unit cell consists of a thickness d_1 of material #1 and d_2 of material #2. Because of the different band gaps, a periodic array of potential wells and potential barriers is formed. When the band offsets are both positive as shown in this figure, the structure is called a type I superlattice.

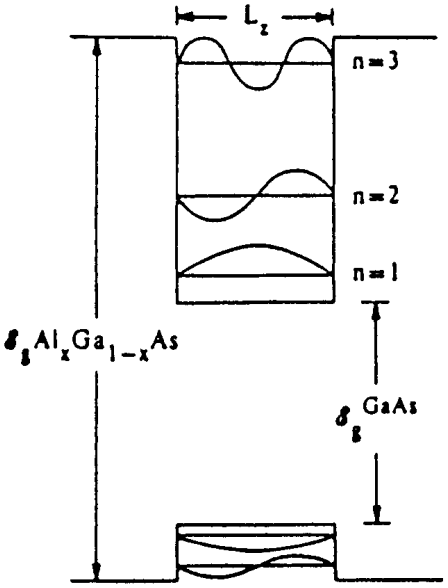


Figure 9.6: The eigenfunctions and bound state energies of an infinitely deep potential well used as an approximation to the states in two finite wells. The upper well applies to electrons and the lower one to holes. This diagram is a schematic representation of a quantum well in the GaAs region formed by the adjacent wider gap semiconductor $Al_xGa_{1-x}As$.

Table 9.1: Material parameters of GaAs, GaP, InAs, and InP.[1]

Property	Parameter (units)	GaAs	GaP	InAs	InP
Lattice constant	a(Å)	5.6533	5.4512	6.0584	5.8688
Density	$g(\mathrm{g/cm^3})$	5.307	4.130	5.667	4.787
Thermal expansion	$\alpha_{th}(\times 10^{-6}/^\circ\mathrm{C})$	6.63	5.91	5.16	4.56
Γ point band gap	E_0(eV)	1.42	2.74	0.36	1.35
plus spin orbit	$E_0+\Delta_0$(eV)	1.76	2.84	0.79	1.45
L point band gap	E_1(eV)	2.925	3.75	2.50	3.155
plus spin orbit	$E_1+\Delta_1$(eV)	3.155	...	2.78	3.305
Γ point band gap	$E_0{}'$(eV)	4.44	4.78	4.44	4.72
Δ axis band gap	E_2(eV)	4.99	5.27	4.70	5.04
plus spin orbit	$E_2+\delta$(eV)	5.33	5.74	5.18	5.60
Gap pressure coefficient	$\partial E_0/\partial P(\times 10^{-6}\mathrm{eV/bar})$	11.5	11.0	10.0	8.5
Gap temperature coefficient	$\partial E_0/\partial T(\times 10^{-4}\mathrm{eV}/^\circ\mathrm{C})$	−3.95	−4.6	−3.5	−2.9
Electron mass	m^*/m_0	0.067	0.17	0.023	0.08
light hole	${m_{\ell h}}^*/m_0$	0.074	0.14	0.027	0.089
heavy hole	${m_{hh}}^*/m_0$	0.62	0.79	0.60	0.85
spin orbit hole	${m_{so}}^*/m_0$	0.15	0.24	0.089	0.17
Dielectric constant: static	ϵ_s	13.1	11.1	14.6	12.4
Dielectric constant: optic	ϵ_∞	11.1	8.46	12.25	9.55
Ionicity	f_1	0.310	0.327	0.357	0.421
Polaron coupling	α_F	0.07	0.20	0.05	0.08
Elastic constants	$c_{11}(\times 10^{11}\mathrm{dyn/cm^2})$	11.88	14.120	8.329	10.22
	$c_{12}(\times 10^{11}\mathrm{dyn/cm^2})$	5.38	6.253	4.526	5.76
	$c_{44}(\times 10^{11}\mathrm{dyn/cm^2})$	5.94	7.047	3.959	4.60
Young's modulus	$Y(\times 10^{11}\mathrm{dyn/cm^2})$	8.53	10.28	5.14	6.07
	P	0.312	0.307	0.352	0.360
Bulk modulus	$B(\times 10^{11}\ \mathrm{dyn/cm^2})$	7.55	8.88	5.79	7.25
	A	0.547	0.558	0.480	0.485
Piezo–electric coupling	$e_{14}(\mathrm{C/m^2})$	−0.16	−0.10	−0.045	−0.035
	$K_{[110]}$	0.0617	0.0384	0.0201	0.0158
Deformation potential	a(eV)	2.7	3.0	2.5	2.9
	b(eV)	−1.7	−1.5	−1.8	−2.0
	d(eV)	−4.55	−4.6	−3.6	−5.0
Deformation potential	Ξ_{eff}(eV)	6.74	6.10	6.76	7.95
Donor binding	G(meV)	4.4	10.0	1.2	5.5
Donor radius	a_B(Å)	136	48	406	106
Thermal conductivity	$\kappa(\mathrm{watt/deg-cm})$	0.46	0.77	0.273	0.68
Electron mobility	$\mu_n(\mathrm{cm^2/V-sec})$	8000	120	30000	4500
Hole mobility	$\mu_p(\mathrm{cm^2/V-sec})$	300	–	450	100

[1]Table from *J. Appl. Physics* 53, 8777 (1982).

9.4 Bound Electronic States

From the diagram in Fig. 9.5 we see that the heterojunction superlattice consists of an array of potential wells. The interesting limit to consider is the case where the width of the potential well contains only a small number of crystallographic unit cells ($L_z < 100$ Å), in which case the number of bound states in the well is a small number.

From a mathematical standpoint, the simplest case to consider is an infinitely deep rectangular potential well. In this case, a particle of mass m^* in a well of width L_z in the z direction satisfies the free particle Schrödinger equation

$$-\frac{\hbar^2}{2m^*}\frac{d^2\psi}{dz^2} = E\psi \tag{9.9}$$

with eigenvalues

$$E_n = \frac{\hbar^2}{2m^*}\left(\frac{n\pi}{L_z}\right)^2 = \left(\frac{\hbar^2\pi^2}{2m^*L_z^2}\right)n^2 \tag{9.10}$$

and the eigenfunctions

$$\psi_n = A\sin(n\pi z/L_z) \tag{9.11}$$

where n = 1,2,3.... are the plane wave solutions that satisfy the boundary conditions that the wave functions in Eq. 9.11 must vanish at the walls of the quantum wells ($z = 0$ and $z = L_z$).

We note that the energy levels are not equally spaced, but have energies $E_n \sim n^2$, though the spacings $E_{n+1} - E_n$ are proportional to n. We also note that $E_n \sim {L_z}^{-2}$, so that as L_z becomes large, the levels become very closely spaced as expected for a 3D semiconductor. However when L_z decreases, the number of states in the quantum well decreases, so that for a well depth E_d it would seem that there is a critical width ${L_z}^c$ below which there would be no bound states

$${L_z}^c = \frac{\hbar\pi}{(2m^*E_d)^{\frac{1}{2}}}. \tag{9.12}$$

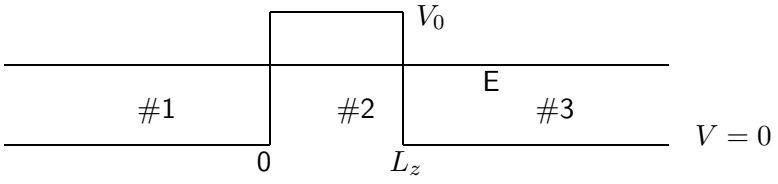
An estimate for ${L_z}^c$ is obtained by taking $m^* = 0.1m_0$ and $E_d = 0.1$ eV to yield ${L_z}^c = 61$Å. There is actually a theorem in quantum mechanics that says that there will be at least one bound state for an arbitrarily small potential well. More exact calculations considering quantum wells of finite thickness have been carried out, and show that the infinite well approximation gives qualitatively correct results.

The closer level spacing of the valence band bound states in Fig. 9.6 reflects the heavier masses in the valence band. Since the states in the potential well are quantized, the structures in Figs. 9.5 and 9.6 are called quantum well structures.

If the potential energy of the well V_0 is not infinite but finite, the wave functions are similar to those given in Eq. 9.11, but will have decaying exponentials on either side of the potential well walls. The effect of the finite size of the well on the energy levels and wave functions is most pronounced near the top of the well. When the particle has an energy greater than V_0, its eigenfunction corresponds to a continuum state $\exp(ik_z z)$.

In the case of MOSFETs, the quantum well is not of rectangular shape as shown in Fig. 9.7, but rather is approximated as a triangular well. The solution for the bound states in a triangular well cannot be solved exactly, but can only be done approximately, as for example using the WKB approximation described in §9.6.

Figure 9.7: Schematic of a potential barrier.



9.5 Review of Tunneling Through a Potential Barrier

When the potential well is finite, the wave functions do not completely vanish at the walls of the well, so that tunneling through the potential well becomes possible. We now briefly review the quantum mechanics of tunneling through a potential barrier. We will return to tunneling in semiconductor heterostructures after some introductory material.

Suppose that the potential V shown in Fig. 9.7 is zero ($V = 0$) in regions #1 and #3, while $V = V_0$ in region #2. Then in regions #1 and #3

$$E = \frac{\hbar^2 k^2}{2m^*} \tag{9.13}$$

$$\psi = e^{ikz} \tag{9.14}$$

while in region #2 the wave function is exponentially decaying

$$\psi = \psi_0 e^{-\beta z} \tag{9.15}$$

so that substitution into Schrödinger's equation gives

$$\frac{-\hbar^2}{2m^*}\beta^2\psi + (V_0 - E)\psi = 0 \tag{9.16}$$

or

$$\beta^2 = \frac{2m^*}{\hbar^2}(V_0 - E). \tag{9.17}$$

The probability that the electron tunnels through the rectangular potential barrier is then given by

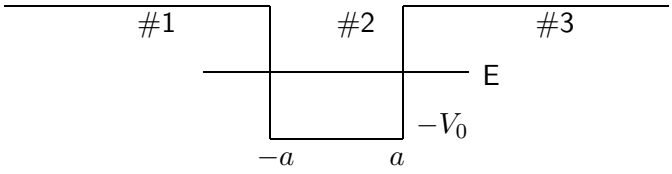
$$\mathcal{P} = \exp\Big\{ -2\int_0^{L_z} \beta(z)dz \Big\} = \exp\Big\{ -2\Big(\frac{2m^*}{\hbar^2}\Big)^{\frac{1}{2}}(V_0 - E)^{\frac{1}{2}}L_z \Big\} \tag{9.18}$$

As L_z increases, the probability of tunneling decreases exponentially. Electron tunneling phenomena frequently occur in solid state physics.

9.6 Quantum Wells of Different Shape and the WKB Approximation

With the sophisticated computer control available with state of the art molecular beam epitaxy systems it is now possible to produce quantum wells with specified potential profiles $V(z)$ for semiconductor heterojunction superlattices. Potential wells with non–rectangular

Figure 9.8: Schematic of a rectangular well.



profiles also occur in the fabrication of other types of superlattices (e.g., by modulation doping). We therefore briefly discuss (in the recitation class) bound states in general potential wells.

In the general case where the potential well has an arbitrary shape, solution by the WKB (Wentzel–Kramers–Brillouin) approximation is very useful (see for example, Shanker, "Principles of Quantum Mechanics", Plenum press, chapter 6). According to this approximation, the energy levels satisfy the Bohr–Sommerfeld quantization condition

$$\int_{z_1}^{z_2} p_z dz = \hbar\pi(r + c_1 + c_2) \tag{9.19}$$

where $p_z = (2m^*[E - V])^{\frac{1}{2}}$ and the quantum number r is an integer $r = 0, 1, 2, \ldots$ while c_1 and c_2 are the phases which depend on the form of $V(z)$ at the turning points z_1 and z_2 where $V(z_i) = E$. If the potential has a sharp discontinuity at a turning point, then $c = 1/2$, but if V depends linearly on z at the turning point then $c = 1/4$.

For example for the infinite rectangular well (see Fig. 9.8)

$$V(z) = 0 \quad \text{for } \mid z \mid < a \quad \text{(inside the well)} \tag{9.20}$$

$$V(z) = \infty \quad \text{for } \mid z \mid > a \quad \text{(outside the well)} \tag{9.21}$$

By the WKB rules, the turning points occur at the edges of the rectangular well and therefore $c_1 = c_2 = 1/2$. In this case p_z is a constant, independent of z so that $p_z = (2m^*E)^{\frac{1}{2}}$ and Eq. 9.19 yields

$$(2m^*E)^{\frac{1}{2}} L_z = \hbar\pi(r + 1) = \hbar\pi n \tag{9.22}$$

where $n = r + 1$ and

$$E_n = \frac{\hbar^2\pi^2}{2m^* {L_z}^2} n^2 \tag{9.23}$$

in agreement with the exact solution given by Eq. 9.10. The finite rectangular well shown in Fig. 9.8 is thus approximated as an infinite well with solutions given by Eq. 9.10.

As a second example consider a harmonic oscillator potential well shown in Fig. 9.9, where $V(z) = m^*\omega^2 z^2/2$. The harmonic oscillator potential well is typical of quantum wells in periodically doped (nipi which is n-type; insulator; p-type; insulator) superlattices. In this case

$$p_z = (2m^*)^{1/2}\left(E - \frac{m^*\omega^2}{2} z^2\right)^{\frac{1}{2}}. \tag{9.24}$$

The turning points occur when $V(z) = E$ so that the turning points are given by $z = \pm(2E/m^*\omega^2)^{\frac{1}{2}}$. Near the turning points $V(z)$ is approximately linear in z, so the phase

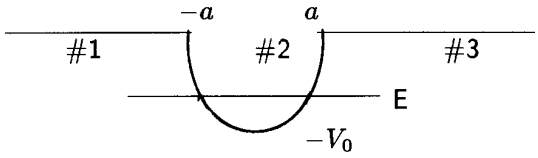
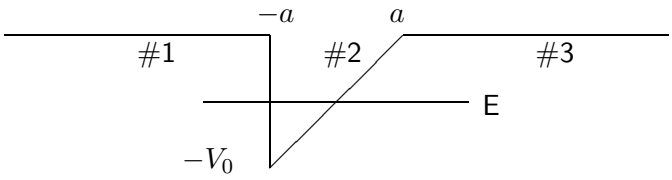


Figure 9.9: Schematic of a harmonic oscillator well.

Figure 9.10: Schematic of a triangular well.



factors become $c_1 = c_2 = \frac{1}{4}$. The Bohr–Sommerfeld quantization thus yields

$$\int_{z_1}^{z_2} p_z dz = \int_{z_1}^{z_2} (2m^*)^{\frac{1}{2}} \left(E - \frac{m^*\omega^2}{2} z^2 \right)^{\frac{1}{2}} dz = \hbar\pi(r + \frac{1}{2}). \tag{9.25}$$

Making use of the integral relation

$$\int \sqrt{a^2 - u^2}\, du = \frac{u}{2}\sqrt{a^2 - u^2} + \frac{a^2}{2}\sin^{-1}\frac{u}{a} \tag{9.26}$$

we obtain upon substitution of Eq. 9.26 into 9.25:

$$(2m^*)^{\frac{1}{2}} \left(\frac{m^*\omega^2}{2} \right)^{\frac{1}{2}} \left(\frac{E_r}{m^*\omega^2} \right) \pi = \frac{E_r \pi}{\omega} = \hbar\pi(r + \frac{1}{2}) \tag{9.27}$$

which simplifies to the familiar relation for the harmonic oscillator energy levels:

$$E_r = \hbar\omega(r + \frac{1}{2}) \quad \text{where} \quad r = 0, 1, 2... \tag{9.28}$$

another example of an exact solution. For homework, you will use the WKB method to find the energy levels for an asymmetric triangular well. Such quantum wells are typical of the interface of metal–insulator–semiconductor (MOSFET) device structures (see Fig. 9.10).

9.7 The Kronig–Penney Model

We review here the Kronig–Penney model which gives an explicit solution for a one–dimensional array of finite potential wells shown in Fig. 9.11. Starting with the one dimensional Hamiltonian with a periodic potential (see Eq. 9.7)

$$-\frac{\hbar^2}{2m^*}\frac{d^2\psi}{dz^2} + V(z)\psi = E\psi \tag{9.29}$$

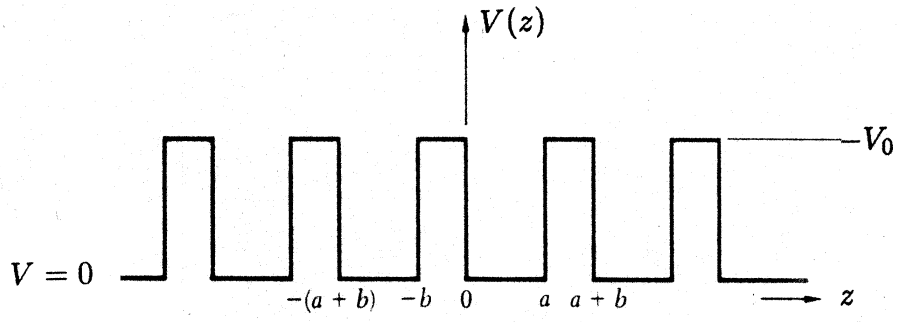


Figure 9.11: Kronig–Penney square well periodic potential

we obtain solutions in the region $0 < z < a$ where $V(z) = 0$

$$\psi(z) = Ae^{iKz} + Be^{-iKz} \tag{9.30}$$

$$E = \frac{\hbar^2 K^2}{2m^*} \tag{9.31}$$

and in the region $-b < z < 0$ where $V(z) = V_0$ (the barrier region)

$$\psi(z) = Ce^{\beta z} + De^{-\beta z} \tag{9.32}$$

where

$$\beta^2 = \frac{2m^*}{\hbar^2}[V_0 - E]. \tag{9.33}$$

Continuity of $\psi(z)$ and $d\psi(z)/dz$ at $z = 0$ and $z = a$ determines the coefficients A, B, C, D. At $z = 0$ we have:

$$\begin{aligned} A + B &= C + D \\ iK(A - B) &= \beta(C - D) \end{aligned} \tag{9.34}$$

At $z = a$, we apply Bloch's theorem (see Fig. 9.11), introducing a factor $\exp[ik(a + b)]$ to obtain $\psi(a) = \psi(-b)\exp[ik(a + b)]$

$$\begin{aligned} Ae^{iKa} + Be^{-iKa} &= (Ce^{-\beta b} + De^{\beta b})e^{ik(a+b)} \\ iK(Ae^{iKa} - Be^{-iKa}) &= \beta(Ce^{-\beta b} - De^{\beta b})e^{ik(a+b)}. \end{aligned} \tag{9.35}$$

These 4 equations (Eqs. 9.34 and 9.35) in 4 unknowns determine A, B, C, D. The vanishing of the coefficient determinant restricts the conditions under which solutions to the Kronig–Penney model are possible, leading to the algebraic equation

$$\frac{\beta^2 - K^2}{2\beta K}\sinh\beta b \sin Ka + \cosh\beta b \cos Ka = \cos k(a + b) \tag{9.36}$$

which has solutions for a limited range of β values.

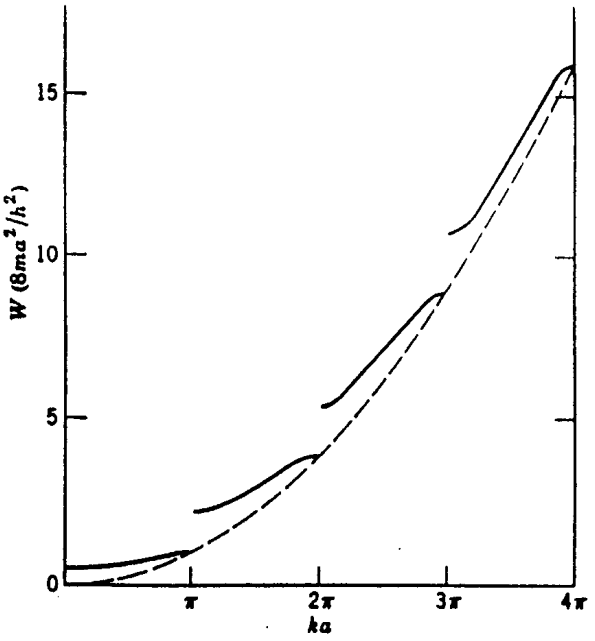


Figure 9.12: Plot of energy vs. k for the Kronig–Penney model with $P = 3\pi/2$. (After Sommerfeld and Bethe.)

Normally the Kronig–Penney model in the textbooks is solved in the limit $b \to 0$ and $V_0 \to \infty$ in such a way that $[\beta^2 ba/2] = P$ remains finite. The restricted solutions in this limit lead to the energy bands shown in Fig. 9.12.

For the superlattice problem we are interested in solutions both within the quantum wells and in the continuum. This is one reason for discussing the Kronig–Penny model. Another reason for discussing this model is because it provides a review of boundary conditions and the application of Bloch's theorem. In the quantum wells, the permitted solutions give rise to narrow bands with large band gaps while in the continuum regions the solutions correspond to wide bands and small band gaps.

9.8 3D Motion within a 1–D Rectangular Well

The thin films used for the fabrication of quantum well structures (see §9.4) are very thin in the z–direction but have macroscopic size in the perpendicular x–y plane. An example of a quantum well structure would be a thin layer of GaAs sandwiched between two thicker $Al_xGa_{1-x}As$ layers, as shown in the Fig. 9.5. For the thin film, the motion in the x and y directions is similar to that of the corresponding bulk solid which can be treated by the conventional 1–electron approximation and the Effective Mass Theorem. Thus the potential can be written as a sum of a periodic term $V(x, y)$ and the quantum well term $V(z)$. The electron energies thus are superimposed on the quantum well energies, the periodic solutions obtained from solution of the 2–D periodic potential

$$E_n(k_x, k_y) = E_{n,z} + \frac{\hbar^2(k_x^2 + k_y^2)}{2m^*} = E_{n,z} + E_\perp \tag{9.37}$$

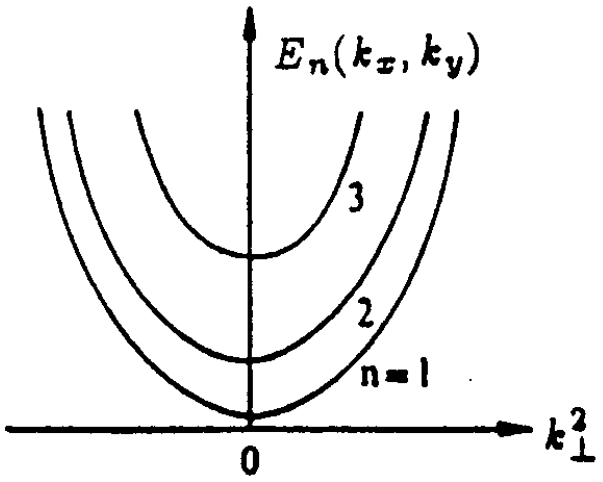


Figure 9.13: Subbands associated with bound states for the 2D electron gas.

in which the quantized bound state energies $E_{n,z}$ are given by Eq. 9.10. A plot of the energy levels is given in Fig. 9.13. At $(k_x, k_y) = (0,0)$ the energy is precisely the quantum well energy E_n for all n. The band of energies associated with each state n is called a subband.

Of particular interest is the density of states for the quantum well structures. Associated with each two–dimensional subband is a constant density of states, as derived below. From elementary considerations the number of electrons per unit area in a 2–dimensional circle is given by

$$N_{2D} = \frac{2}{(2\pi)^2}\pi k_\perp^2 \tag{9.38}$$

where $k_\perp^2 = k_x^2 + k_y^2$ and

$$E_\perp = \frac{\hbar^2 k_\perp^2}{2m^*} \tag{9.39}$$

so that for each subband the density of states $g_{2D}(E)$ contribution becomes

$$\frac{\partial N_{2D}}{\partial E} = g_{2D}(E) = \frac{m^*}{\pi\hbar^2}. \tag{9.40}$$

If we now plot the density of states corresponding to the 3D motion in a 1–D rectangular well, we have $g_{2D}(E) = 0$ until the bound state energy E_1 is reached, when a step function contribution of $(m^*/\pi\hbar^2)$ is made. The density of states $g_{2D}(E)$ will then remain constant until the minimum of subband E_2 is reached when an additional step function contribution of $(m^*/\pi\hbar^2)$ is made, hence yielding the staircase density of states shown in Fig. 9.14. Two generalizations of Eq. 9.40 for the density of states for actual quantum wells are needed, as we discuss below. The first generalization takes into account the finite size L_z of the quantum well, so that the system is not completely two dimensional and some k_z dispersion must occur. Secondly, the valence bands of typical semiconductors are degenerate so that coupling between the valence band levels occurs, giving rise to departures from the simple parabolic bands discussed below.

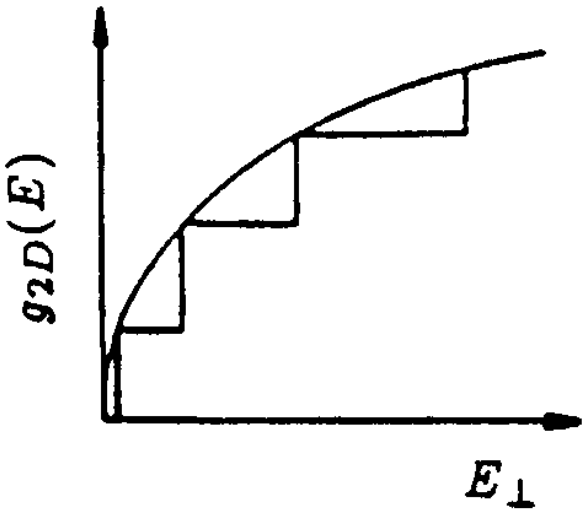


Figure 9.14: Two dimensional density of states $g_{2D}(E)$ for rectangular quantum well structures.

A comparison between the energy dispersion relation $E(\vec{k})$ and the density of states $g(E)$ in two dimensions and three dimensions is shown in Fig. 9.15 together with a quasi two–dimensional case, typical of actual quantum well samples. In the quasi–two dimensional case, the $E(\vec{k})$ relations exhibit a small degree of dispersion along k_z, leading to a corresponding width in the steps of the density of states function shown in Fig. 9.15(b).

A generalization of the simple 2D density of states in Fig. 9.14 is also necessary to treat the complex valence band of a typical III-V compound semiconductor. The $E(\vec{k})$ diagram (where $k_\perp$ is normal to k_z) for the heavy hole and light hole levels can be calculated using $\vec{k} \cdot \vec{p}$ perturbation theory to be discussed later in the course.

The most direct evidence for bound states in quantum wells comes from optical absorption measurements (to be discussed later in the course) and resonant tunneling effects which we discuss below.

9.9 Resonant Tunneling in Quantum Wells

Resonant tunneling (see Fig. 9.18) provides direct evidence for the existence of bound states in quantum wells. We review first the background material for tunneling across potential barriers in semiconductors and then apply these concepts to the resonant tunneling phenomenon.

The carriers in the quantum well structures are normally described in terms of the effective mass theorem where the wave functions for the carriers are given by the one electron approximation. The effective mass equation is written in terms of slowly varying wavefunctions corresponding to a slowly varying potential which satisfies Poisson's equation when an electric field is applied (e.g., a voltage is imposed across the quantum well structure).

Further simplifications that are made in treating the tunneling problem include:

1. The wavefunctions for the tunneling particle are expanded in terms of a single band

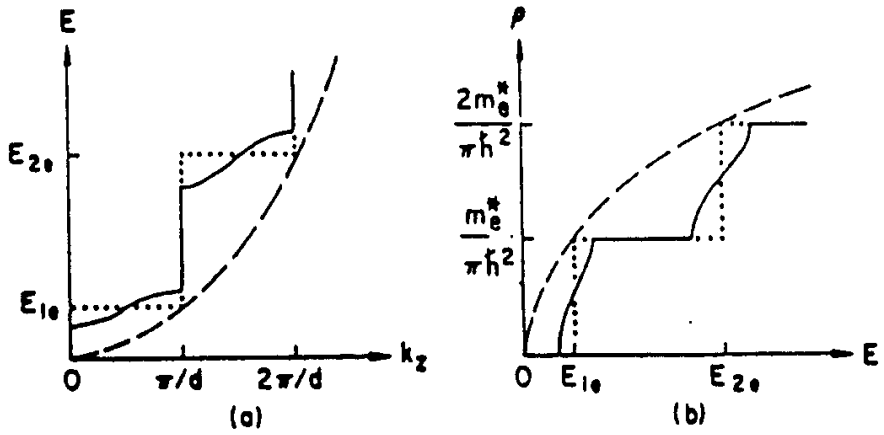


Figure 9.15: Schematic diagrams of (a) energy dispersion and (b) density of states. Indicated are the two–dimensional (dotted), three–dimensional (dashed), and intermediate (solid) cases.

on either side of the junction.

2. Schrödinger's equation is separated into two components, parallel and perpendicular to the junction plane, leading to a 1–dimensional tunneling problem.

3. The eigenstates of interest have energies sufficiently near those of critical points in the energy band structure on both sides of the interface so that the simplified form of the effective mass theorem can be used.

4. The total energy, E, and the momentum parallel to the interface or perpendicular to the layering direction, $k_\perp$, are conserved in the tunneling process. Since the potential acts only in the z–direction, the 1–dimensional Schrödinger equation becomes:

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dz^2} + V(z) - E\right]\psi_e = 0 \tag{9.41}$$

where $V(z)$ is the electrostatic potential, and ψ_e is an envelope function. The wave function ψ_e is subject, at an interface $z = z_1$ (see Fig. 9.16), to the following boundary conditions that guarantee current conservation:

$$\psi_e(z_1^-) = \psi_e(z_1^+) \tag{9.42}$$

$$\frac{1}{m_1}\left.\frac{d}{dz}\psi_e\right|_{z_1^-} = \frac{1}{m_2}\left.\frac{d}{dz}\psi_e\right|_{z_1^+} \tag{9.43}$$

The current density for tunneling through a barrier becomes

$$J_z = \frac{e}{4\pi^3\hbar}\int dk_z d^2k_\perp f(E)T(E_z)\frac{dE}{dk_z} \tag{9.44}$$

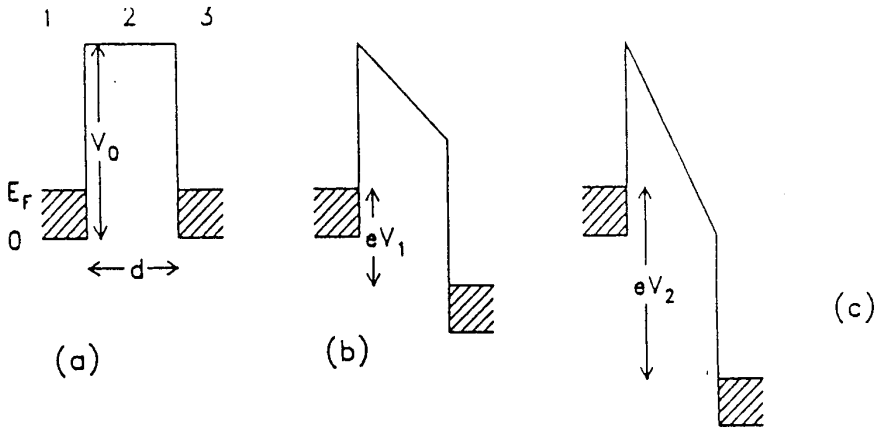


Figure 9.16: Rectangular–potential model (a) used to describe the effect of an insulator, 2, between two metals, 1 and 3. When a negative bias is applied to 1, electrons, with energies up to the Fermi energy E_F, can tunnel through the barrier. For small voltages, (b), the barrier becomes trapezoidal, but at high bias (c), it becomes triangular.

where $f(E)$ is the Fermi–Dirac distribution, and $T(E_z)$ is the probability of tunneling through the potential barrier. Here $T(E_z)$ is expressed as the ratio between the transmitted and incident probability currents.

If an external bias V is applied to the barrier (see Fig. 9.16), the net current flowing through it is the difference between the current from left to right and that from right to left. Thus, we obtain:

$$J_z = \frac{e}{4\pi^3\hbar}\int dE_z d^2k_\perp [f(E) - f(E+eV)]T(E_z) \tag{9.45}$$

where E_z represents the energy from the k_z component of crystal momentum, i.e., $E_z = \hbar^2 k_z^2/(2m)$. Since the integrand is not a function of $k_\perp$ in a plane normal to k_z, we can integrate over $d^2k_\perp$ by writing

$$dk_x dk_y = d^2k_\perp = \frac{2m}{\hbar^2} dE_\perp \tag{9.46}$$

where $E_\perp = \hbar^2 k_\perp^2/(2m)$ and after some algebra, the tunneling current can be written as,

$$\begin{aligned} J_z &= \tfrac{em}{2\pi^2\hbar^3}\left[eV\int_0^{E_F-eV} dE_z T(E_z) + \int_{E_F-eV}^{E_F} dE_z (E_F - E_z)T(E_z)\right] \quad \text{if } eV \le E_F \\ J_z &= \tfrac{em}{2\pi^2\hbar^3}\int_0^{E_F} dE_z (E_F - E_z)T(E_z) \qquad\qquad\qquad\qquad\qquad\quad \text{if } eV \ge E_F \end{aligned} \tag{9.47}$$

(see Fig. 9.16 for the geometry of the model) which can be evaluated as long as the tunneling probability through the barrier is known. We now discuss how to find the tunneling probability.

An enhanced tunneling probability occurs for certain voltages as a consequence of the constructive interference between the incident and the reflected waves in the barrier region between regions 1 and 3. To produce an interference effect the wavevector $\vec{k}$ in the plane

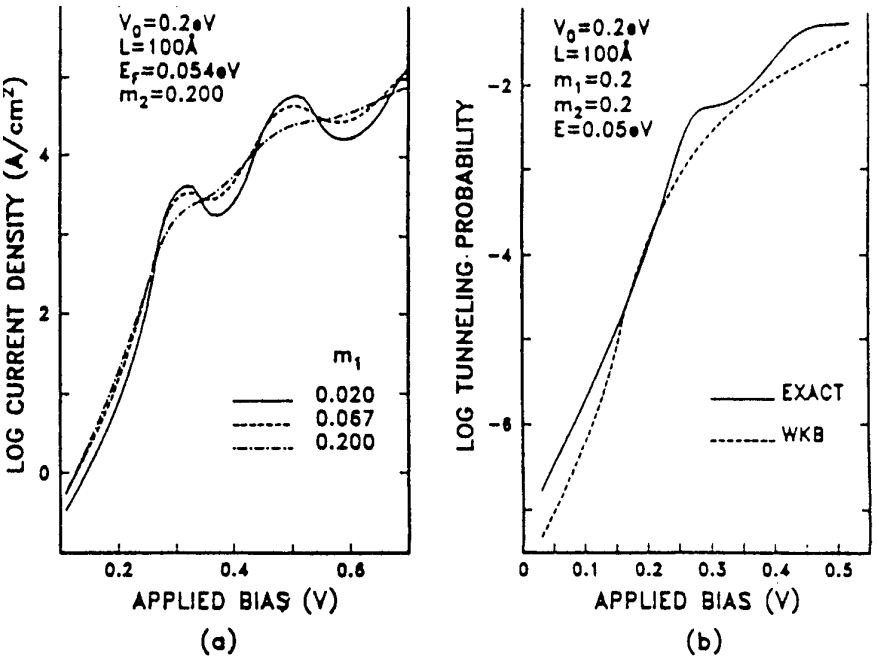


Figure 9.17: (a) Tunneling current through a rectangular barrier (like the one of Fig. 9.16a) calculated as a function of bias for different values of m_1, in the quantum well. (b) Comparison of an exact calculation of the tunneling probability through a potential barrier under an external bias with an approximate result obtained using the WKB method. The barrier parameters are the same as in (a), and the energy of an incident electron, of mass $0.2m_0$, is 0.05eV. (From the book of E.E. Mendez and K. von Klitzing, "Physics and Applications of Quantum Wells and Superlattices", NATO ASI Series, Vol. 170, p.159 (1987)).

wave solution e^{ikz} must have a real component so that an oscillating (rather than a decaying exponential) solution is possible. To accomplish this, it is necessary for a sufficiently high electric field to be applied (as in Fig. 9.16(c)) so that a virtual bound state is formed. As can be seen in Fig. 9.17a, the oscillations are most pronounced when the difference between the electronic mass at the barrier and at the electrodes is the largest. This interference phenomenon is frequently called resonant Fowler–Nordheim tunneling and has been observed in metal–oxide–semiconductor (MOS) heterostructures and in GaAs/$Ga_{1-x}Al_xAs$/GaAs capacitors. Since the WKB method is semiclassical, it does not give rise to the resonant tunneling phenomenon, which is a quantum interference effect.

For the calculation of the resonant tunneling phenomenon, we must therefore use the quantum mechanical solution. In this case, it is convenient to use the transfer–matrix method to find the tunneling probability. In region (#1) of Fig. 9.16, the potential $V(z)$ is constant and solutions to Eq. 9.41 have the form

$$\psi_e(z) = A\exp(ikz) + B\exp(-ikz) \tag{9.48}$$

with

$$\frac{\hbar^2 k^2}{2m} = E - V. \tag{9.49}$$

When $E-V>0$, then k is real and the wave functions are plane waves. When $E-V<0$, then k is imaginary and the wave functions are growing or decaying waves. The boundary conditions Eqs. 9.42 and 9.43 determine the coefficients A and B which can be described by a (2×2) matrix R such that

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = R \begin{pmatrix} A_2 \\ B_2 \end{pmatrix} \tag{9.50}$$

where the subscripts on A and B refer to the region index and R can be written as

$$R = \frac{1}{2k_1m_2} \begin{pmatrix} (k_1m_2+k_2m_1)\exp[i(k_2-k_1)z_1] & (k_1m_2-k_2m_1)\exp[-i(k_2+k_1)z_1] \\ (k_1m_2-k_2m_1)\exp[i(k_2+k_1)z_1] & (k_1m_2+k_2m_1)\exp[-i(k_2-k_1)z_1] \end{pmatrix} \tag{9.51}$$

and the terms in R of Eq. 9.51 are obtained by matching boundary conditions as given in Eqs. 9.42 and 9.43.

In general, if the potential profile consists of n regions, characterized by the potential values V_i and the masses m_i $(i=1,2,\ldots n)$, separated by $n-1$ interfaces at positions z_i $(i=1,2,\ldots(n-1))$, then

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = (R_1R_2\ldots R_{n-1}) \begin{pmatrix} A_n \\ B_n \end{pmatrix}. \tag{9.52}$$

The matrix elements of R_i are

$$\begin{aligned} (R_i)_{1,1} &= \left(\tfrac{1}{2}+\tfrac{k_{i+1}m_i}{2k_im_{i+1}}\right)\exp[i(k_{i+1}-k_i)z_i] \\ (R_i)_{1,2} &= \left(\tfrac{1}{2}-\tfrac{k_{i+1}m_i}{2k_im_{i+1}}\right)\exp[-i(k_{i+1}+k_i)z_i] \\ (R_i)_{2,1} &= \left(\tfrac{1}{2}-\tfrac{k_{i+1}m_i}{2k_im_{i+1}}\right)\exp[i(k_{i+1}+k_i)z_i] \\ (R_i)_{2,2} &= \left(\tfrac{1}{2}+\tfrac{k_{i+1}m_i}{2k_im_{i+1}}\right)\exp[-i(k_{i+1}-k_i)z_i] \end{aligned} \tag{9.53}$$

where the k_i are defined by Eq. 9.49. If an electron is incident from the left (region #1) only a transmitted wave will appear in the last region #n, and therefore $B_n=0$. The transmission probability is then given by

$$T = \left(\frac{k_1m_n}{k_nm_1}\right)\frac{|A_n|^2}{|A_1|^2}. \tag{9.54}$$

This is a general solution to the problem of transmission through multiple barriers. Under certain conditions, a particle incident on the left can appear on the right essentially without attenuation. This situation, called resonant tunneling, corresponds to a constructive interference between the two plane waves coexisting in the region between the barriers (quantum well).

The tunneling probability through a double rectangular barrier is illustrated in Fig. 9.18. In this figure, the mass of the particle is taken to be $0.067m_0$, the height of the barriers is 0.3eV, their widths are 50Å and their separations are 60Å. As observed in the figures, for

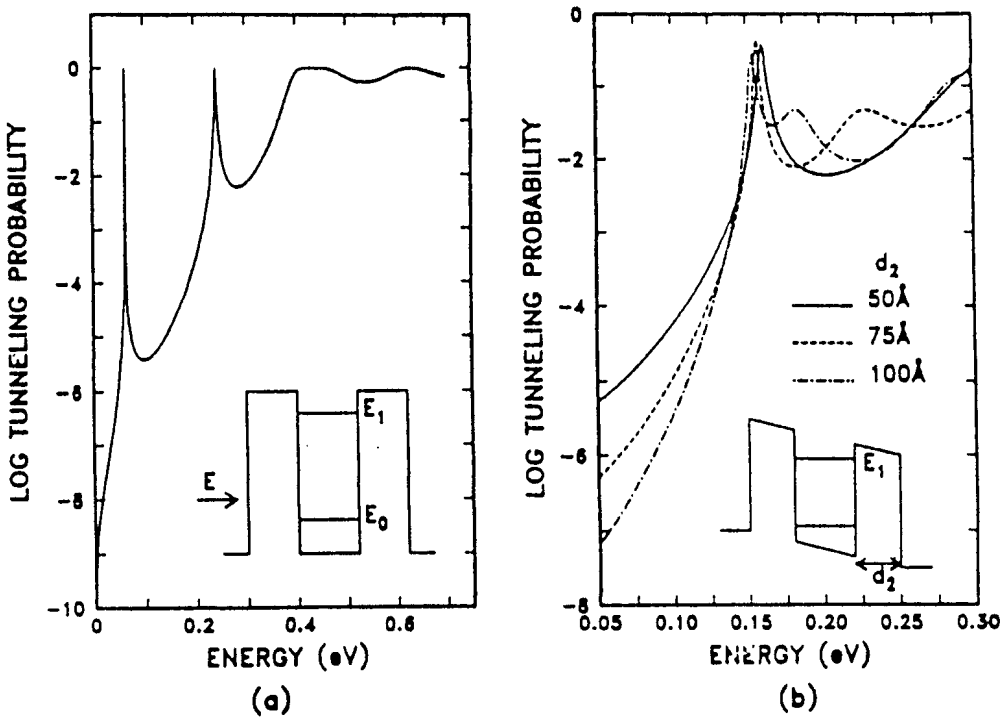


Figure 9.18: (a) Probability of tunneling through a double rectangular barrier as a function of energy. The carrier mass is taken to be $0.1m_0$ in the barrier and $0.067m_0$ outside, and the width of the quantum well is 60Å. (b) Tunneling probability through a double–barrier structure, subject to an electric field of 1×10^5V/cm. The width of the left barrier is 50Å, while that of the right barrier is varied between 50Å and 100Å. The peak at ~ 0.16eV corresponds to resonant tunneling through the first excited state (E_1) of the quantum well. The optimum transmission is obtained when the width of the right barrier is ~ 75Å.

certain energies below the barrier height, the particle can tunnel without attenuation. These energies correspond precisely to the eigenvalues of the quantum well; this is understandable, since the solutions of Schrödinger's equation for an isolated well are standing waves. When the widths of the two barriers are different (see Fig. 9.18b), the tunneling probability does not reach unity, although the tunneling probability shows maxima for incident energies corresponding to the bound and virtual states.

Chapter 10

Transport in Low Dimensional Systems

References:

- *Solid State Physics*, Volume **44**, Semiconductor Heterostructures and Nanostructures. Edited by H. Ehrenreich and D. Turnbull, Academic Press (1991).
- *Electronic transport in mesoscopic systems*, Supriyo Datta, Cambridge University Press, 1995.

10.1 Introduction

Transport phenomena in low dimensional systems such as in quantum wells (2D), quantum wires (1D), and quantum dots (0D) are dominated by quantum effects not included in the classical treatments based on the Boltzmann equation and discussed in Chapters 4-6. With the availability of experimental techniques to synthesize materials of high chemical purity and of nanometer dimensions, transport in low dimensional systems has become an active current research area. In this chapter we consider some highlights on the subject of transport in low dimensional systems.

10.2 Observation of Quantum Effects in Reduced Dimensions

Quantum effects dominate the transport in quantum wells and other low dimensional systems such as quantum wires and quantum dots when the de Broglie wavelength of the electron

$$\lambda_{\rm dB} = \frac{\hbar}{(2m^*E)^{1/2}} \tag{10.1}$$

exceeds the dimensions of a quantum structure of characteristic length L_z ($\lambda_{\rm dB} > L_z$) or likewise for tunneling through a potential barrier of length L_z. To get some order of magnitude estimates of the electron kinetic energies E below which quantum effects become important we look at Fig. 10.1 where a log-log plot of $\lambda_{\rm dB}$ vs E in Eq. 10.1 is shown for

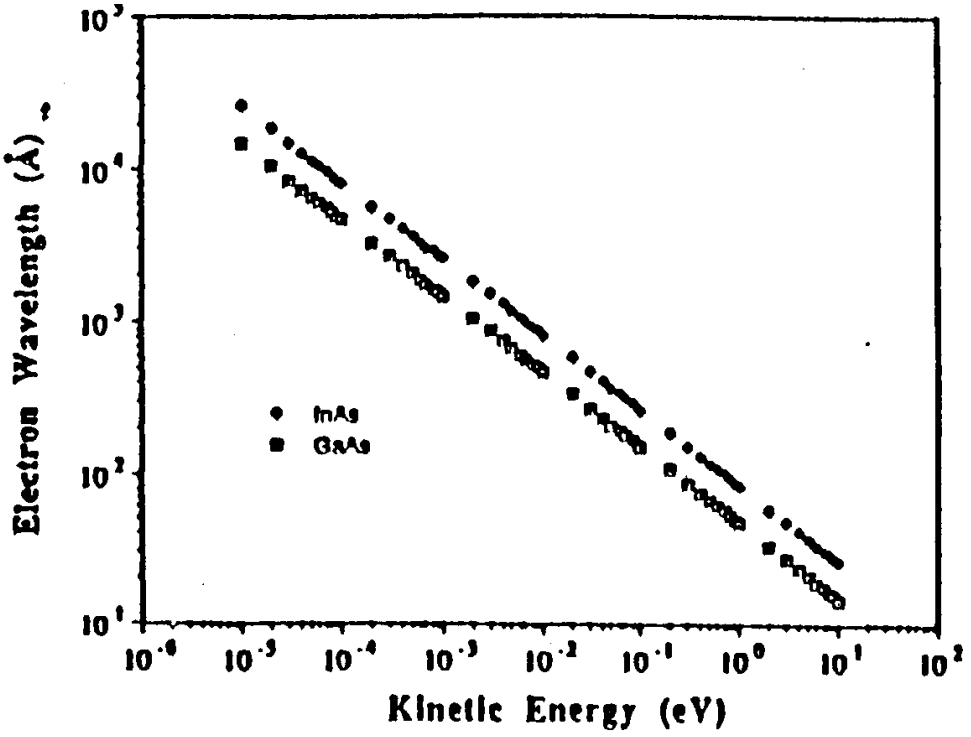


Figure 10.1: Plot of the electron de Broglie wavelength $\lambda_{\rm dB}$ vs the electron kinetic energy E for GaAs ($\Box$) and InAs ($\diamond$).

GaAs and InAs. From the plot we see that an electron energy of $E \sim 0.1$ eV for GaAs corresponds to a de Broglie wavelength of 200 Å. Thus wave properties for electrons can be expected for structures smaller than $\lambda_{\rm dB}$.

To observe quantum effects, the thermal energy must also be less than the energy level separation, $k_B T < \Delta E$, where we note that room temperature corresponds to 25 meV. Since quantum effects depend on the phase coherence of electrons, scattering can also destroy quantum effects. The observation of quantum effects thus requires that the carrier mean free path be much larger than the dimensions of the quantum structures (wells, wires or dots).

The limit where quantum effects become important has been given the name of **mesoscopic physics**. Carrier transport in this limit exhibits both particle and wave characteristics. In this ballistic transport limit, carriers can in some cases transmit charge or energy without scattering.

The small dimensions required for the observation of quantum effects can be achieved by the direct fabrication of semiconductor elements of small dimensions (quantum wells, quantum wires and quantum dots). Another approach is the use of gates on a field effect transistor to define an electron gas of reduced dimensionality. In this context, negatively charged metal gates can be used to control the source to drain current of a 2D electron gas formed near the GaAs/AlGaAs interface as shown in Fig. 10.2. Between the dual gates shown on this figure, a thin conducting wire is formed out of the 2D electron gas. Controlling the gate voltage controls the amount of charge in the depletion region under the gates, as well as the charge in the quantum wire. Thus lower dimensional channels can be made in a 2D electron gas by using metallic gates. In the following sections a number of important applications are made of this concept.

10.3 Density of States in Low Dimensional Systems

We showed in Eq. 8.40 that the density of states for a 2D electron gas is a constant for each 2D subband

$$g_{\rm 2D} = \frac{m^*}{\pi \hbar^2}. \tag{10.2}$$

This is shown in Fig. 10.3(a) where the inset is appropriate to the quantum well formed near a modulation doped GaAs-AlGaAs interface. In the diagram only the lowest bound state is occupied.

Using the same argument, we now derive the density of states for a 1D electron gas

$$N_{\rm 1D} = \frac{2}{2\pi}(k) = \frac{1}{\pi}(k) \tag{10.3}$$

which for a parabolic band $E = E_n + \hbar^2 k^2/(2m^*)$ becomes

$$N_{\rm 1D} = \frac{2}{2\pi}(k) = \frac{1}{\pi}\left(\frac{2m^*(E-E_n)}{\hbar^2}\right)^{1/2} \tag{10.4}$$

yielding an expression for the density of states $g_{\rm 1D}(E) = \partial N_{\rm 1D}/\partial E$

$$g_{\rm 1D}(E) = \frac{1}{2\pi}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}(E-E_n)^{-1/2}. \tag{10.5}$$

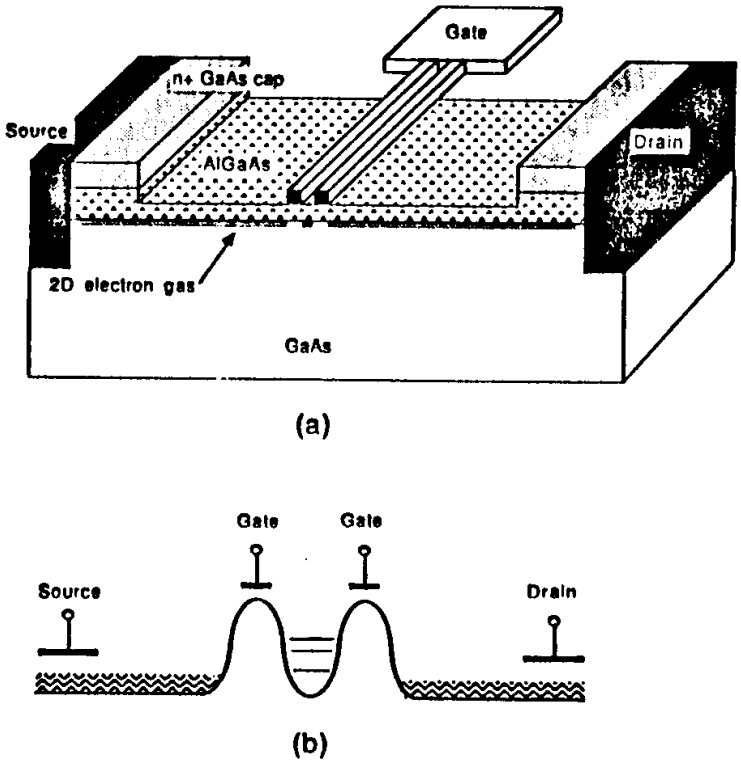
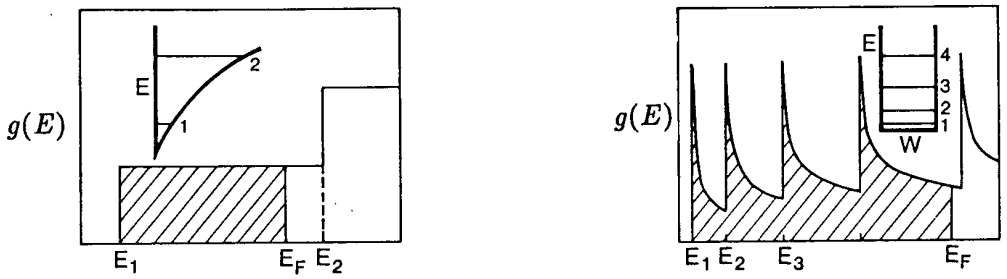


Figure 10.2: (a) Schematic diagram of a lateral resonant tunneling field-effect transistor which has two closely spaced fine finger metal gates; (b) schematic of an energy band diagram for the device. A 1D quantum wire is formed in the 2D electron gas between the gates.



(a) (b)

Figure 10.3: Density of states $g(E)$ as a function of energy. (a) Quasi-2D density of states, with only the lowest subband occupied (hatched). Inset: Confinement potential perpendicular to the plane of the 2DEG. The discrete energy levels correspond to the bottoms of the first and second 2D subbands. (b) Quasi-1D density of states, with four 1D subbands occupied. Inset: Square-well lateral confinement potential with discrete energy levels indicating the 1D subband extrema.

The interpretation of this expression is that at each doubly confined bound state level E_n there is a singularity in the density of states, as shown in Fig. 10.3(b) where the first four levels are occupied.

10.3.1 Quantum Dots

This is an example of a zero dimensional system. Since the levels are all discrete any averaging would involve a sum over levels and not an integral over energy. If, however, one chooses to think in terms of a density of states, then the DOS would be a delta function positioned at the energy of the localized state. For more extensive treatment see the review by Marc Kastner (Appendix C).

10.4 The Einstein Relation and the Landauer Formula

In the classical transport theory (Chapter 4) we related the current density $\vec{j}$ to the electric field $\vec{E}$ through the conductivity σ using the Drude formula

$$\sigma = \frac{ne^2\tau}{m^*}. \tag{10.6}$$

This equation is valid when many scattering events occur within the path of an electron through a solid, as shown in Fig. 10.4(a). As the dimensions of device structures become smaller and smaller, other regimes become important, as shown in Figs. 10.4(b) and 10.4(c).

To relate transport properties to device dimensions it is often convenient to rewrite the Drude formula by explicitly substituting for the carrier density n and for the relaxation time τ in Eq. 10.6. Writing $\tau = \ell/v_F$ where ℓ is the mean free path and v_F is the Fermi

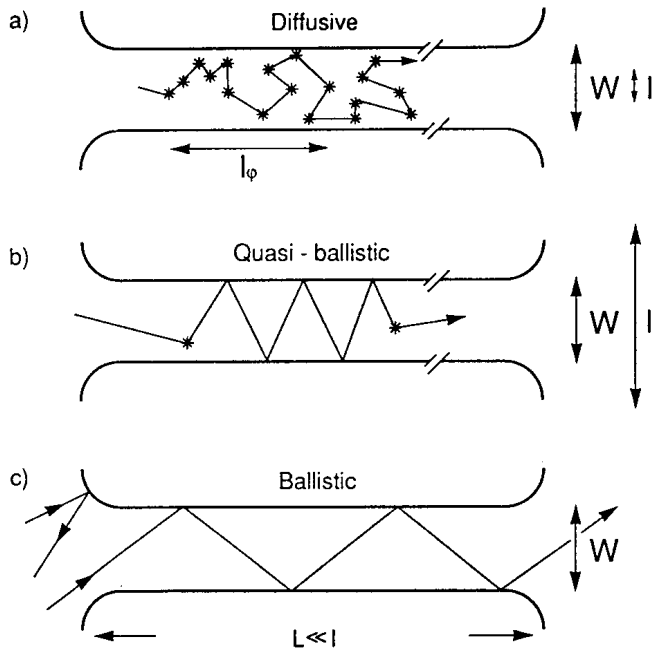


Figure 10.4: Electron trajectories characteristic of the diffusive ($\ell < W, L$), quasi-ballistic ($W < \ell < L$), and ballistic ($W, L < \ell$) transport regimes, for the case of specular boundary scattering. Boundary scattering and internal impurity scattering (asterisks) are of equal importance in the quasi-ballistic regime. A nonzero resistance in the ballistic regime results from backscattering at the connection between the narrow channel and the wide 2DEG regions. Taken from H. Van Houten et al. in "Physics and Technology of Submicron Structures" (H. Heinrich, G. Bauer and F. Kuchar, eds.) Springer, Berlin, 1988.

velocity, and writing $n = k_F^2/(2\pi)$ for the carrier density for a 2D electron gas (2DEG) we obtain

$$\sigma = \frac{k_F^2}{2\pi} e^2 \frac{\ell}{m^* v_F} = \frac{k_F^2 e^2 \ell}{2\pi\hbar k_F} = \frac{e^2}{h} k_F \ell \tag{10.7}$$

where e^2/h is a universal constant and is equal to $\sim (26\ \mathrm{k}\Omega)^{-1}$.

Two general relations that are often used to describe transport in situations where collisions are not important within device dimensions are the Einstein relation and the Landauer formula. We discuss these relations below. The Einstein relation follows from the continuity equation

$$\vec{j} = eD\vec{\nabla}n \tag{10.8}$$

where D is the diffusion coefficient and $\vec{\nabla}n$ is the gradient of the carrier density involved in the charge transport. In equilibrium the gradient in the electrochemical potential $\vec{\nabla}\mu$ is zero and is balanced by the electrical force and the change in Fermi energy

$$\vec{\nabla}\mu = 0 = -e\vec{E} + \vec{\nabla}n \frac{dE_F}{dn} = -e\vec{E} + \vec{\nabla}n/g(E_F) \tag{10.9}$$

where $g(E_F)$ is the density of states at the Fermi energy. Substitution of Eq. 10.9 into Eq. 10.8 yields

$$\vec{j} = eDg(E_F)e\vec{E} = \sigma\vec{E} \tag{10.10}$$

yielding the Einstein relation

$$\sigma = e^2 D g(E_F) \tag{10.11}$$

which is a general relation valid for 3D systems as well as systems of lower dimensions.

The Landauer formula is an expression for the conductance G which is the proportionality between the current I and the voltage V,

$$I = GV. \tag{10.12}$$

For 2D systems the conductance and the conductivity have the same dimensions, and for a large 2D conductor we can write

$$G = (W/L)\sigma \tag{10.13}$$

where W and L are the width and length of the conducting channel in the current direction, respectively. If W and L are both large compared to the mean free path ℓ, then we are in the diffusive regime (see Fig. 10.4(a)). However when we are in the opposite regime, the ballistic regime, where $\ell > W, L$, then the conductance is written in terms of the Landauer formula which is obtained from Eqs. 10.7 and 10.13. Writing the number of quantum modes N, then $N\pi = k_F W$ or

$$k_F = \frac{N\pi}{W} \tag{10.14}$$

and noting that the quantum mechanical transition probability coupling one channel to another in the ballistic limit $|t_{\alpha,\beta}|^2$ is $\pi\ell/(2LN)$ per mode, we obtain the general Landauer formula

$$G = \frac{2e^2}{h} \sum_{\alpha,\beta}^{N} |t_{\alpha,\beta}|^2. \tag{10.15}$$

We will obtain the Landauer formula below for some explicit examples, which will make the derivation of the normalization factor more convincing.

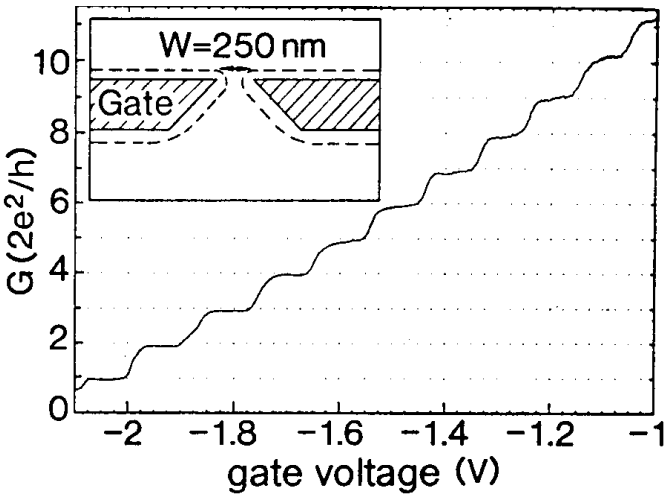


Figure 10.5: Point contact conductance as a function of gate voltage at 0.6 K, obtained from the raw data after subtraction of the background resistance. The conductance shows plateaus at multiples of $e^2/\pi\hbar$. Inset: Point-contact layout [from B.J. van Wees, et al., Phys. Rev. Lett. 60, 848 (1988)].

10.5 One Dimensional Transport and Quantization of the Ballistic Conductance

In the last few years one dimensional ballistic transport has been demonstrated in a two dimensional electron gas (2DEG) of a GaAs-GaAlAs heterojunction by constricting the electron gas to flow in a very narrow channel (see Fig. 10.5). Ballistic transport refers to carrier transport without scattering. As we show below, in the ballistic regime, the conductance of the 2DEG through the constriction shows quantized behavior with the conductance changing in quantized steps of $(e^2/\pi\hbar)$ when the effective width of the constricting channel is varied by controlling the voltages of the gate above the 2DEG. We first give a derivation of the quantization of the conductance.

The current I_x flowing between source and drain (see Fig. 10.2) due to the contribution of one particular 1D electron subband is given by

$$I_x = ne\delta v \tag{10.16}$$

where n is the carrier density (i.e., the number of carriers per unit length of the channel) and δv is the increase in electron velocity due to the application of a voltage V. The carrier density in 1D is

$$n = \frac{2}{2\pi}k_F = \frac{k_F}{\pi} \tag{10.17}$$

and the gain in velocity δv resulting from an applied voltage V is

$$eV = \frac{1}{2}m^*(v_F + \delta v)^2 - \frac{1}{2}m^* v_F^2 = m^* v_F \delta v + \frac{1}{2}m^*(\delta v)^2. \tag{10.18}$$

Retaining only the first order term in Eq. 10.18 yields

$$\delta v = eV/m^* v_F \tag{10.19}$$

so that from Eq. 10.16 we get for the source-drain current (see Fig. 10.5)

$$I_x = \frac{k_F}{\pi}\, e\, \frac{eV}{m^* v_F} = \frac{e^2}{\pi\hbar} V \tag{10.20}$$

since $\hbar k_F = m^* v_F$. This yields a conductance per 1D electron subband G_i of

$$G_i = \frac{e^2}{\pi\hbar} \tag{10.21}$$

or summing over all occupied subbands i we obtain

$$G = \sum_i \frac{e^2}{\pi\hbar} = \frac{ie^2}{\pi\hbar}. \tag{10.22}$$

Two experimental observations of these phenomena were simultaneously published [D.A. Wharam, T.J. Thornton, R. Newbury, M. Pepper, H. Ahmed, J.E.F. Frost, D.G. Hasko, D.C. Peacock, D.A. Ritchie, and G.A.C. Jones, J. Phys. C: Solid State Phys. 21, L209 (1988); and B.J. van Wees, H. van Houten, C.W.J. Beenakker, J.G. Williamson, L.P. Kouvenhoven, D. van der Marel, and C.T. Foxon, Phys. Rev. Lett. 60, 848 (1988)]. The experiments by Van Wees et al. were done using ballistic point contacts on a gate structure placed over a two-dimensional electron gas as shown schematically in the inset of Fig. 10.5. The width W of the gate (in this case 2500 Å) defines the effective width W' of the conducting electron channel, and the applied gate voltage is varied in order to control the effective width W'. Superimposed on the raw data for the resistance vs gate voltage is a collection of periodic steps as shown in Fig. 10.5 after subtracting off the background resistance of 400 Ω.

There are several conditions necessary to observe perfect $2e^2/h$ quantization of the 1D conductance. One requirement is that the electron mean free path l_e be much greater than the length of the channel L. This limits the values of channel lengths to L <5,000 Å even though mean free path values are much larger, l_e=8.5 μm. It is important to note, however, that l_e=8.5 μm is the mean free path for the 2D electron gas. When the channel is formed, the screening effect of the 2D electron gas is no longer present and the effective mean free path becomes much shorter. A second condition is that there are adiabatic transitions at the inputs and outputs of the channel. This minimizes reflections at these two points, an important condition for the validity of the Landauer formula to be discussed later in this section. A third condition requires the Fermi wavelength $\lambda_F = 2\pi/k_F$ (or $k_F L > 2\pi$) to satisfy the relation $\lambda_F < L$ by introducing a sufficient carrier density ($3.6\times10^{11}\mathrm{cm}^{-2}$) into the channel. Finally, as discussed earlier, it is necessary that the thermal energy $k_B T \ll E_j - E_{j-1}$ where $E_j - E_{j-1}$ is the subband separation between the j and $j-1$ one dimensional energy levels. Therefore, the quantum conductance measurements are done at low temperatures (T <1 K).

The point contacts in Fig. 10.5 were made on high-mobility molecular-beam-epitaxy-grown GaAs-AlGaAs heterostructures using electron beam lithography. The electron density of the material is $3.6 \times 10^{11}/\mathrm{cm}^2$ and the mobility is $8.5 \times 10^5 \mathrm{cm}^2/\mathrm{V\ s}$ (at 0.6 K). These values were obtained directly from measurements of the devices themselves. For the transport measurements, a standard Hall bar geometry was defined by wet etching. At a gate voltage of $V_g = -0.6$ V the electron gas underneath the gate is depleted, so that conduction takes place through the point contact only. At this voltage, the point contacts

have their maximum effective width $W'_{\max}$, which is about equal to the opening W between the gates. By a further decrease (more negative) of the gate voltage, the width of the point contacts can be reduced, until they are fully pinched off at $V_g = -2.2$ V.

The results agree well with the appearance of conductance steps that are integral multiples of $e^2/\pi h$, indicating that the conductance depends directly on the number of 1D subbands that are occupied with electrons. To check the validity of the proposed explanation for these steps in the conductance (see Eq. 10.22), the effective width W' for the gate was estimated from the voltage $V_g = -0.6$ V to be 3600 Å, which is close to the geometric value for W. In Fig. 10.5 we see that the average conductance varies linearly with V_g which in turn indicates a linear relation between the effective point contact width W' and V_g. From the 16 observed steps and a maximum effective point contact width $W'_{\max} = 3600$ Å, an estimate of 220 Å is obtained for the increase in width per step, corresponding to $\lambda_F/2$. Theoretical work done by Rolf Landauer nearly 20 years ago shows that transport through the channel can be described by summing up the conductances for all the possible transmission modes, each with a well defined transmission coefficient t_{nm}. The conductance of the 1D channel can then be described by the Landauer formula

$$G = \frac{e^2}{\pi\hbar} \sum_{n,m=1}^{N_c} |t_{nm}|^2 \tag{10.23}$$

where N_c is the number of occupied subbands. If the conditions for perfect quantization described earlier are satisfied, then the transmission coefficient reduces to $|t_{nm}|^2 = \delta_{nm}$. This corresponds to purely ballistic transport with no scattering or mode mixing in the channel (i.e., no back reflections).

A more explicit derivation of the Landauer formula for the special case of a 1D system can be done as follows. The current flowing in a 1D channel can be written as

$$I_j = \int_{E_i}^{E_f} e g_j(E) v_z(E) T_j(E) dE \tag{10.24}$$

where the electron velocity is given by

$$m^* v_z = \hbar k_z \tag{10.25}$$

and

$$E = E_j + \frac{\hbar^2 k_z^2}{2m^*} \tag{10.26}$$

while the 1D density of states is from Eq. 10.5 given by

$$g_j(E) = \frac{(2m^*)^{1/2}}{h(E - E_g)^{1/2}} \tag{10.27}$$

and $T_j(E)$ is the probability that an electron injected into subband j with energy E will get across the 1D wire ballistically. Substitution of Eqs. 10.25, 10.26 and 10.27 into Eq. 10.24 then yields

$$I_j = \frac{2e}{h} \int_{E_i}^{E_j} T_j(E) dE = \frac{2e^2}{h} T_j \Delta V \tag{10.28}$$

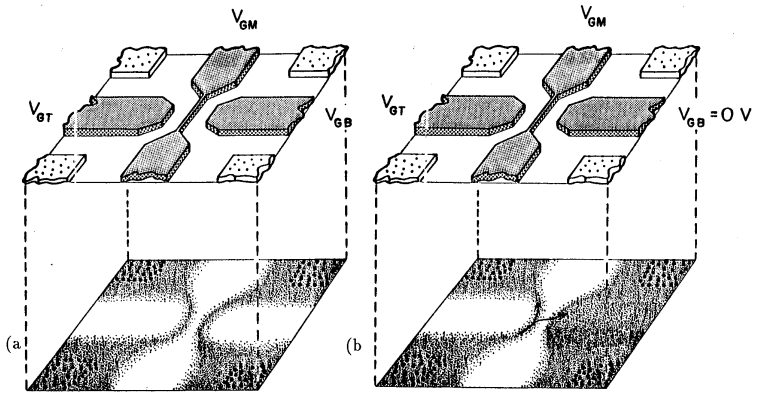


Figure 10.6: (a) Schematic illustration of the split-gate dual electron waveguide device. The top plane shows the patterned gates at the surface of the MODFET structure. The bottom plane shows the implementation of two closely spaced electron waveguides when the gates (indicated by V_{GT}, V_{GB} and V_{GM}) are properly biased. Shading represents the electron concentration. Also shown are the four ohmic contacts which allow access to the inputs and outputs of each waveguide. (b) Schematic of the "leaky" electron waveguide implementation. The bottom gate is grounded (V_{GB}=0) so that only one waveguide is in an "on" state. V_{GM} is fixed such that only a small tunneling current crosses it. The current flowing through the waveguide as well as the tunneling current (depicted by arrows) are monitored simultaneously.

where we have noted that the potential energy difference is the difference between initial and final energies $e\Delta V = E_f - E_i$. Summing over all occupied states j we then obtain the Landauer formula

$$G = \frac{2e^2}{h} \sum_j T_j. \tag{10.29}$$

10.6 Ballistic Transport in 1D Electron Waveguides

Another interesting quantum-effect structure proposed and implemented by C. Eugster and J. del Alamo is a split-gate dual electron waveguide device shown in Fig. 10.6 [C.C. Eugster, J.A. del Alamo, M.J. Rooks and M.R. Melloch, Applied Physics Letters **60**, 642 (1992)]. By applying the appropriate negative biases on patterned gates at the surface, two electron waveguides can be formed at the heterointerface of a MODFET structure. As shown in Fig. 10.6, the two electron waveguides are closely spaced over a certain length and their separation is controlled by the middle gate bias (V_{GM}). The conductance of each waveguide, shown in Fig. 10.7, is measured simultaneously and independently as a function of the side gate biases (V_{GT} and V_{GB}) and each show the quantized $2e^2/h$ conductance steps. Such a device can be used to study 1D coupled electron waveguide interactions. An electron directional coupler based on such a structure has also been proposed [J. del Alamo and C. Eugster, Appl. Phys. Lett. **56**, 78 (1990)]. Since each gate can be independently accessed, various other regimes can be studied in addition to the coupled waveguide regime.

One interesting regime is that of a "leaky" electron waveguide [C.C. Eugster and J.A. del

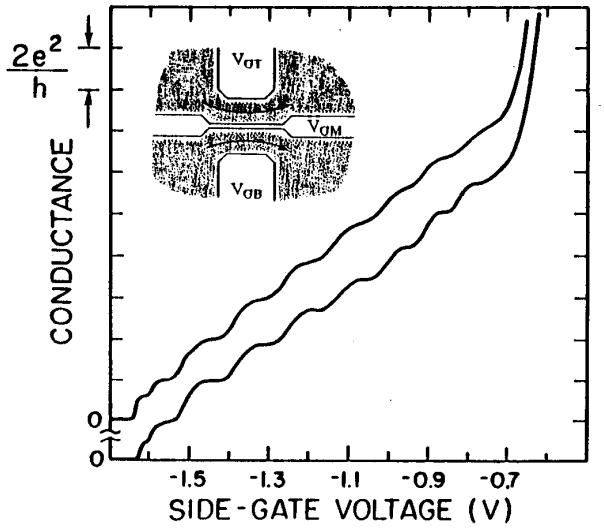


Figure 10.7: Conductance of each waveguide in Fig. 10.6(a) as a function of side gate bias of an $L = 0.5\ \mu$m, $W = 0.3\ \mu$m split-gate dual electron waveguide device. The inset shows the biasing conditions (i.e., in the depletion regime) and the direction of current flow for the measurements.

Alamo, Physical Review Letters **67**, 3586 (1991)]. For such a scheme, one of the side gates is grounded so that the 2D electron gas underneath it is unaffected, as shown in Fig. 10.6(b). The middle gate is biased such that only a small tunneling current can flow from one waveguide to the other in the 2D electron gas. The other outer gate bias V_{GT} is used to sweep the subbands in the waveguide through the Fermi level. In such a scheme, there is only one waveguide which has a thin side wall barrier established by the middle gate bias. The current flowing through the waveguide as well as the current leaking out of the thin middle barrier are independently monitored. Figure 10.8 shows the $I - V_{GS}$ characteristics for the leaky electron waveguide implementation. As discussed earlier, conductance steps of order $2e^2/h$ are observed for the current flowing through the waveguide. However what is unique to the leaky electron experiment is that the tunneling current leaking from the thin side wall is monitored. As seen in Fig. 10.8, very strong oscillations in the tunneling current are observed as the Fermi level is modulated in the waveguide. We show below that the tunneling current is directly **tracing** out the 1D density of states of the waveguide.

An expression for the tunneling current I_{S2} flowing through the sidewall of the waveguide can be obtained by the following integral,

$$I_{S2} = e\sum_j \int_{-\infty}^{\infty} v_{\perp j}(E-E_j) g_{1D,j}(E-E_j) T_j(E-E_b) \Big[f(E-E_F-e\Delta V, T) - f(E-E_F, T) \Big] dE \tag{10.30}$$

where we have accounted for the contribution to the current from each occupied subband j. Here E_j is the energy at the bottom of the j^{th} subband and E_b is the height of the tunneling energy barrier. The normal velocity against the tunneling barrier is $v_{\perp j}(E) = \hbar k_{\perp j}/m^*$. The transmission coefficient, $T_j(E - E_b)$, to first order, is the same for the different 1D electron subbands since the barrier height relative to the Fermi level E_F is fixed (see Fig. 10.9). The Fermi function, f, gives the distribution of electrons as a function of temperature and applied bias ΔV between the input of the waveguide and the 2DEG on

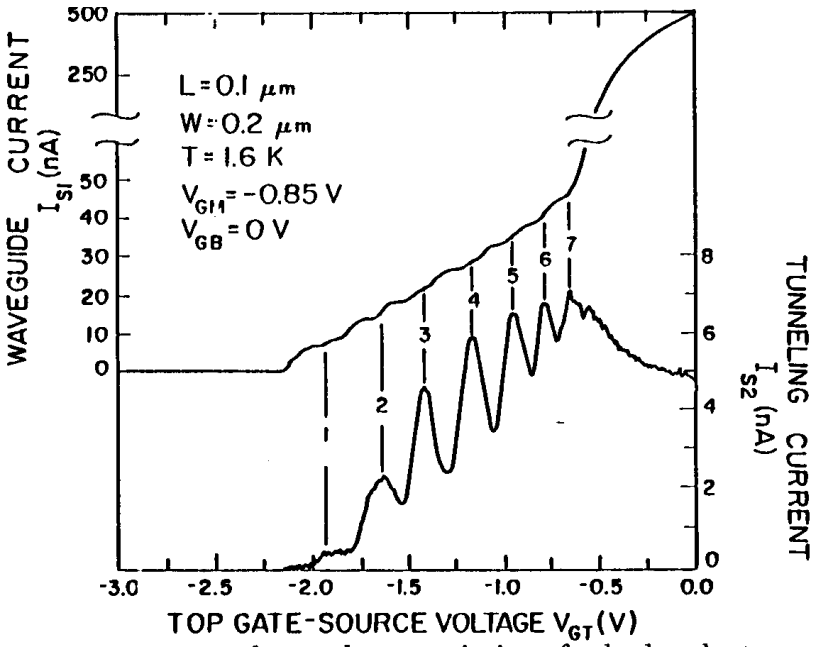


Figure 10.8: The waveguide current vs gate-source voltage characteristics of a leaky electron waveguide implemented with the proper biases for the device. For this case one of the side-gates in Fig. 10.6(b) is grounded so that only one leaky electron waveguide is on. I_{S1} is the current flowing through the waveguide and I_{S2} is the current tunneling through the thin middle side barrier. The bias voltage V_{DS} between the contacts is 100 μV.

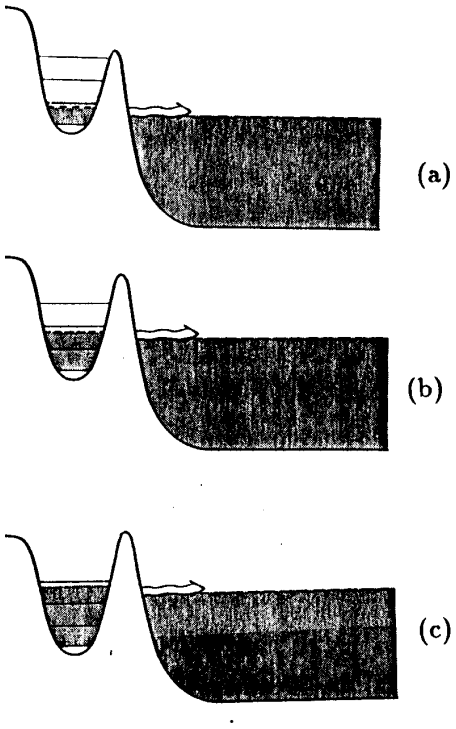


Figure 10.9: Cross-section of a leaky electron waveguide. Shaded regions represent electrons. The dashed line is the Fermi level and the solid lines depict the energy levels in the waveguide. The three figures are from top to bottom (a), (b), (c) at increasingly negative gate-source voltage V_{GT}.

the other side of the tunneling barrier (see Fig. 10.9).

For low enough temperatures and small ΔV, we can approximate Eq. 10.30 by

$$I_{S2} \cong \frac{e^2\hbar}{m^*} \sum_j k_{\perp j} T_j(E_F - E_b) g_{1D,j}(E_F - E_j) \Delta V \tag{10.31}$$

where $k_{\perp j}$ is a constant for each subband and has a value determined by the confining potential. As seen by Eq. 10.31, the tunneling current I_{S2} is proportional to $g_{1D,j}(E_F - E_j)$, the 1D density of states (see Fig. 10.10a). Increasing V_{GT} to less negative values sweeps the subbands through the Fermi level as shown in Fig. 10.9 and since $k_{\perp j}$ and $T_j(E_F - E_b)$ are constant for a given subband (see Fig. 10.10d), the 1D density of states $g_{1D,j}$ summed over each subband, can be extracted from measurement of I_{S2} (see Fig. 10.10e).

Figure 10.9 shows a schematic of the cross-section of the waveguide for three different values of the gate source voltage V_{GT}. The middle gate bias V_{GM} is fixed and is the same for all three cases. The parabolic potential in Fig. 10.9 is a result of the fringing fields and is characteristic of the quantum well for split-gate defined channels. By making V_{GT} more negative, the sidewall potential of the waveguide is raised with respect to the Fermi level. As seen in Fig. 10.9, this results in having fewer energy levels below the Fermi level (i.e., fewer occupied subbands). In Fig. 10.9a, which corresponds to a very negative V_{GT}, only the first subband is occupied. At less negative values of V_{GT}, the second subband becomes occupied (Fig. 10.9b) and then the third (Fig. 10.9c) and so on. There is no tunneling current until the first level has some carrier occupation. The tunneling current increases as a new subband crosses below the Fermi-level. The difference between E_F in the quantum well (on the left) and in the metallic contact (on the right) is controlled by the bias voltage between the input and output contacts of the waveguide V_{DS}. In fact, a finite voltage V_{DS} and a finite temperature gives rise to lifetime effects and a broadening of the oscillations in I_{S2} (see Fig. 10.10e).

A summary of the behavior of a leaky electron waveguide is shown in Fig. 10.10. Included in this figure are (a) the 1D density of states g_{1D}. The measurement of I_{S1} as shown in Fig. 10.8 is modeled in terms of $\sum k_{\|} g_{1D,j}$ and the results for $I_{S1}(V_{GT})$ which relate to the steps in the conductance can be used to extract g_{1D} as indicated in Figs. 10.10b and 10.10c. In contrast, Figs. 10.10d and 10.10e indicate the multiplication of $k_{\perp,j}$ and $g_{1D,j}$ to obtain $\sum k_{\perp,j} g_{1D,j}$ summed over occupied levels which is measured by the tunneling current I_{S2} in Fig. 10.8. The 1D density of states g_{1D} can then be extracted from either I_{S1} or I_{S2} as indicated in Fig. 10.10.

10.7 Single Electron Charging Devices

By making even narrower channels it has been possible to observe single electron charging in a nanometer field-effect transistor, shown schematically in Fig. 10.11. In studies, a metal barrier is placed in the middle of the channel and the width of the metal barrier and the gap between the two constricted gates are of very small dimensions.

The first experimental observation of single electron charging was by Meirav et al. [U. Meirav, M.A. Kastner and S.J. Wind, Phys. Rev. Lett. **65**, 771 (1990); see also M.A. Kastner, Physics Today, page 24, January 1993], working with a double potential barrier GaAs device, as shown in Fig. 10.12. The two dimensional gas forms near the GaAs-

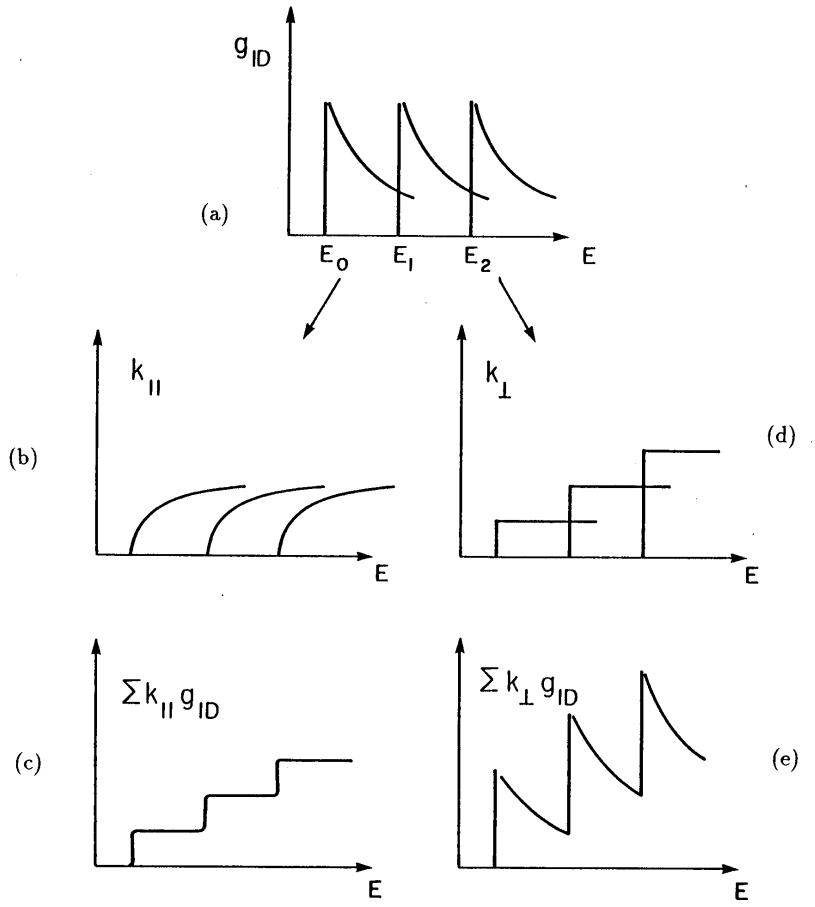


Figure 10.10: Summary of leaky electron waveguide phenomena. Top picture (a) represents the 1D density of states for the waveguide. The two left pictures (b) and (c) are for the current flowing through the waveguide ($k_{\|}$ is the wavevector along waveguide). Quantized conductance steps result from sweeping subbands through the Fermi level as I_{S1} in Fig. 10.8. The two right hand pictures (d) and (e) are for the tunneling current ($k_{\perp}$ is transverse wavevector). Oscillations in the tunneling current I_{S2} in Fig. 10.8 arise from sweeping each subband through the Fermi level.

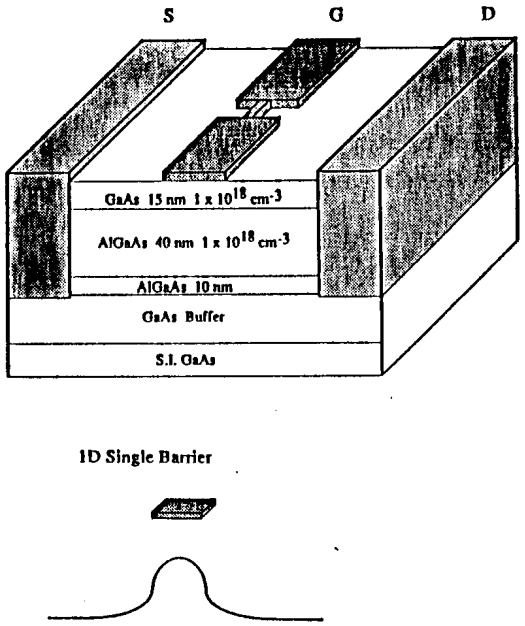


Figure 10.11: A split gate nanometer field-effect transistor, shown schematically. In the narrow channel a 1D electron gas forms when the gate is biased negatively. The potential of the 1D barrier is shown.

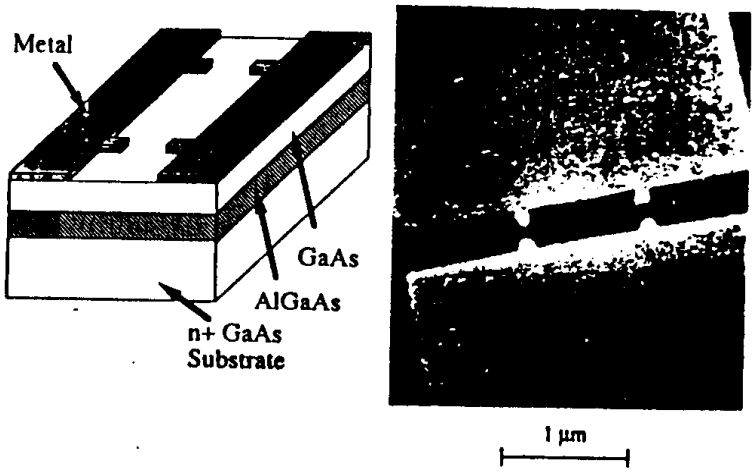


Figure 10.12: Schematic drawing of the device structure along with a scanning electron micrograph of one of the double potential barrier samples. An electron gas forms at the top GaAs-AlGaAs interface, with an electron density controlled by the gate voltage V_g. The patterned metal electrodes on top define a narrow channel with two potential barriers.

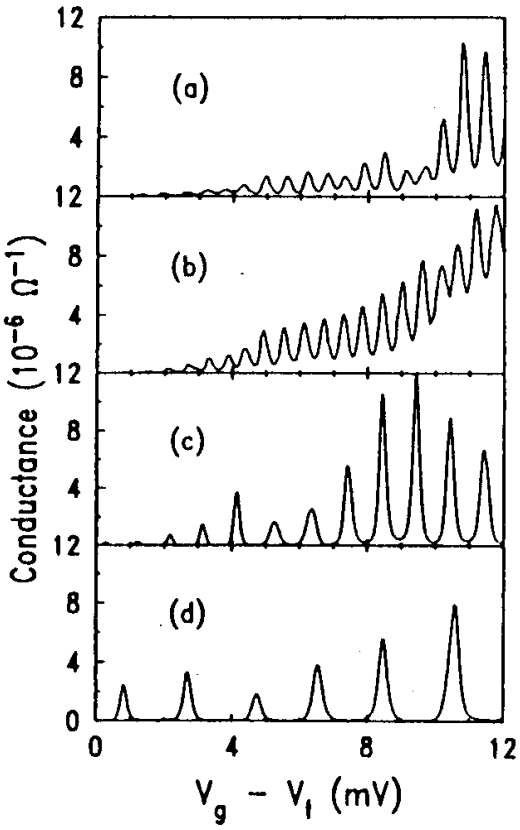


Figure 10.13: Periodic oscillations of the conductance vs gate voltage V_g, measured at $T \approx 50$mK on a sample-dependent threshold V_t. Traces (a) and (b) are for two samples with the same electrode geometry and hence show the same period. Traces (c) and (d) show data for progressively shorter distances between the two constrictions, with a corresponding increase in period. Each oscillation corresponds to the addition of a single electron between the barriers.

GaAlAs interface. Each of the constrictions in Fig. 10.12 is 1000 Å long and the length of the channel between constrictions is 1μm. When a negative bias voltage ($V_b \sim -0.5$V) is applied to the gate, the electron motion through the gates is constrained. At a threshold gate voltage of V_t, the current in the channel goes to zero. This is referred to as Coulomb blockade. If the gate voltage is now increased (positively) above V_t, a series of periodic oscillations are observed, as shown in Fig. 10.13. The correlation between the period of the conductance oscillations and the electron density indicates that a single electron is flowing through the double gated structure per oscillation. The oscillations in Fig 10.13 show the same periodicity when prepared under the same conditions [as in traces (a) and (b)]. As the length L_0 between the constrictions is reduced below 1μm, the oscillation period gets longer. To verify that they had seen single electron charging behavior, Meirav et al. fit the experimental lineshape for a single oscillation to the functional form for the conductance

$$G(\mu) \sim \frac{\partial F}{\partial E} \sim \cosh^{-2}\left[\frac{E-\mu}{2k_BT}\right]. \tag{10.32}$$

where μ is the chemical potential, $F(E-\mu,T)$ is the Fermi function and E is the single electron energy in the 2DEG.

Chapter 11

Ion Implantation and Rutherford Backscattering Spectroscopy

References:

- S.T. Picraux, *Physics Today*, November (1984), p. 38.
- S.T. Picraux and P.S. Peercy, *Scientific American*, March (1985), p. 102.
- J.W. Mayer, L. Eriksson and J.A. Davies, *Ion Implantation of Semiconductors*, Stanford University Press (1970).
- G. Carter and W.A. Grant, *Ion Implantation of Semiconductors*, Edward Arnold Publishers (1976).
- J.K. Hirvonen, *Ion Implantation, Treatise on Materials Science and Technology*, Vol. 18, Academic Press (1980).
- W.K. Chu, J.W. Mayer and M.A. Nicolet, *Backscattering Spectrometry* (Academic Press, 1978)

11.1 Introduction to the Technique

Ions of all energies incident on a solid influence its materials properties. Ions are used in many different ways in research and technology. We here review some of the physics of the interaction of ion beams with solids and some of the uses of ion beams in semiconductors. Some useful reviews are listed above.

We start by noting that the interaction of ion beams with solids depends on the energy of the incident ion. A directed low energy ion ($\sim$ 10–100 eV) comes to rest at or near the surface of a solid, possibly growing into a registered epitaxial layer upon annealing (Fig. 11.1). A 1–keV heavy ion beam is the essential component in the sputtering of surfaces. For this application, a large fraction of the incident energy is transferred to the atoms of the solid resulting in the ejection of surface atoms into the vacuum. The surface is left in a disordered state. Sputtering is used for removal of material from a sample surface on an almost layer–by–layer basis. It is used both in semiconductor device fabrication, ion

etching or ion milling, and more generally in materials analysis, through depth profiling. The sputtered atoms can also be used as a source for the sputter deposition technique discussed in connection with the growth of superlattices.

At higher energies, $\sim$ 100–300 keV, energetic ions are used as a source of atoms to modify the properties of materials. Low concentrations, $\ll 0.1$ atomic percent, of implanted atoms are used to change and control the electrical properties of semiconductors. The implanted atom comes to rest $\sim$ 1000Å below the surface in a region of disorder created by the passage of the implanted ion. The electrical properties of the implanted layer depend on the species and concentration of impurities, the lattice position of the impurities and the amount of lattice disorder that is created.

Lattice site and lattice disorder as well as epitaxial layer formation can readily be analyzed by the channeling of high energy light ions (such as H^+ and He^+), using the Rutherford backscattering technique. In this case MeV ions are used because they penetrate deeply into the crystal (microns) without substantially perturbing the lattice. This is an attractive ion energy regime because scattering cross sections, flux distributions in the crystal, and the rate of energy loss are quantitatively established. Particle–solid interactions in the range from about 0.1 MeV to 5 MeV are understood and one can use this well–characterized tool for investigations of solid state phenomena. Outside of this range many of the concepts discussed below remain valid: however, the use of MeV ions is favored in solid state characterization applications because of both experimental convenience and the ability to probe both surface and bulk properties.

In § 6.2 and § 6.3, we will review the two last energy regimes, namely ion implantation in the 100 keV region and ion beam analysis (Rutherford backscattering and channeling) by MeV light mass particles (H^+, He^+). We start by describing the most important features of ion implantation. Then we will review a two atom collision in order to introduce the basic atomic scattering concepts needed to describe the slowing down of ions in a solid. Then we will state the results of the LSS theory (J. Lindhard, M. Scharff and H. Schiøtt, *Mat. Fys. Medd. Dan. Vid. Selsk.* 33, No. 14 (1963)) which is the most successful basic theory for describing the distribution of ion positions and the radiation–induced disorder in the implanted sample. A number of improvements to this model have been made in the last two decades and are used for current applications. We will conclude this introduction to ion implantation with a discussion of the lattice damage caused by the slowing down of the energetic ions in the solid.

We then go on and describe the use of a beam of energetic (1–2 MeV) light mass particles (H^+, He^+) to study material properties in the near surface region ($< \mu$m) of a solid. We will then see how to use Rutherford backscattering spectrometry (RBS) to determine the stoichiometry of a sample composed of multiple chemical species and the depth distribution of implanted ions. Furthermore, we see that with single–crystal targets, the effect of channeling also allows investigation of the crystalline perfection of the sample as well as the lattice location of the implanted atomic species. Finally, we will review some examples of the modification of material properties by ion implantation, including the use of ion beam analysis in the study of these materials modifications.

11.2 Ion Implantation

Ion implantation is an important technique for introducing impurity atoms in a controlled way, thus leading to the synthesis of new classes of materials, including metastable materials. The technique is important in the semiconductor industry for making p–n junctions by, for example, implanting n–type impurities into a p–type host material. From a materials science point of view, ion implantation allows essentially any element of the periodic table to be introduced into the near surface region of essentially any host material, with quantitative control over the depth and composition profile by proper choice of ion energy and fluence. More generally, through ion implantation, materials with increased strength and corrosion resistance or other desirable properties can be synthesized.

A schematic diagram of an ion implanter is shown in Fig. 11.2. In this diagram the "target" is the sample that is being implanted. In the implantation process, ions of energy E and beam current i_b are incident on a sample surface and come to rest at some characteristic distance R_p with a Gaussian distribution of half width at half maximum ΔR_p. Typical values of the implantation parameters are: ion energies $E \sim 100$ keV, beam currents $i_b \sim 50$ μA, penetration depths $R_p \sim 1000$ Å and half-widths $\Delta R_p \sim 300$ Å (see Fig. 11.3). In the implanted regions, implant concentrations of 10^{-3} to 10^{-5} relative to the host materials are typical. In special cases, local concentrations as high as 20 at.% (atomic percent) of implants have been achieved.

Some characteristic features of ion implantation are the following:

1. The ions characteristically only penetrate the host material to a depth $R_p \leq 1\mu$m. Thus ion implantation is a near surface phenomena. To achieve a large percentage of impurity ions in the host, the host material must be thin (comparable to R_p), and high fluences of implants must be used ($\phi > 10^{16}/\text{cm}^2$).

2. The depth profile of the implanted ions (R_p) is controlled by the ion energy. The impurity content is controlled by the ion fluence ϕ.

3. Implantation is a non–equilibrium process. Therefore there are no solubility limits on the introduction of dopants. With ion implantation one can thus introduce high concentrations of dopants, exceeding the normal solubility limits. For this reason ion implantation permits the synthesis of metastable materials.

4. The implantation process is highly directional with little lateral spread. Thus it is possible to implant materials according to prescribed patterns using masks. Implantation proceeds in the regions where the masks are not present. An application of this technology is to the ion implantation of polymers to make photoresists with sharp boundaries. Both positive and negative photoresists can be prepared using ion implantation, depending on the choice of the polymer. These masks are widely used in the semiconductor industry.

5. The diffusion process is commonly used for the introduction of impurities into semiconductors. Efficient diffusion occurs at high temperatures. With ion implantation, impurities can be introduced at much lower temperatures, as for example room temperature, which is a major convenience to the semiconductor industry.

6. The versatility of ion implantation is another important characteristic. With the same implanter, a large number of different implants can be introduced by merely changing the ion source. The technique is readily automated, and thus is amenable for use by technicians in the semiconductor industry. The implanted atoms are introduced in an atomically dispersed fashion, which is also desirable. Furthermore, no oxide or interfacial barriers are formed in the implantation process.

7. The maximum concentration of ions that can be introduced is limited. As the implantation process proceeds, the incident ions participate in both implantation into the bulk and the sputtering of atoms off the surface. Sputtering occurs because the surface atoms receive sufficient energy to escape from the surface during the collision process. The dynamic equilibrium between the sputtering and implantation processes limits the maximum concentration of the implanted species that can be achieved. Sputtering causes the surface to recede slowly during implantation.

8. Implantation causes radiation damage. For many applications, this radiation damage is undesirable. To reduce the radiation damage, the implantation can be carried out at elevated temperatures or the materials can be annealed after implantation. In practice, the elevated temperatures used for implantation or for post–implantation annealing are much lower than typical temperatures used for the diffusion of impurities into semiconductors.

A variety of techniques are used to characterize the implanted alloy. Ion backscattering of light ions at higher energies (e.g., 2 MeV He^+) is used to determine the composition versus depth with $\sim$ 10nm depth resolution. Depth profiling by sputtering in combination with Auger or secondary ion mass spectroscopy is also used. Lateral resolution is provided by analytical transmission electron microscopy. Electron microscopy, glancing angle x–ray analysis and ion channeling in single crystals provide detailed information on the local atomic structure of the alloys formed. Ion backscattering and channeling are discussed in Section 6.3.

11.2.1 Basic Scattering Equations

An ion penetrating into a solid will lose its energy through the Coulomb interaction with the atoms in the target. This energy loss will determine the final penetration of the projectile into the solid and the amount of disorder created in the lattice of the sample. When we look more closely into one of these collisions (Fig. 11.4) we see that it is a very complicated event in which :

- The two nuclei with masses M_1 and M_2 and charges Z_1 and Z_2, respectively, repel each other by a Coulomb interaction, screened by the respective electron clouds.

- Each electron is attracted by the two nuclei (again with the corresponding screening) and is repelled by all other electrons.

In addition, the target atom is bonded to its neighbors through bonds which usually involve its valence electrons. The collision is thus a many–body event described by a complicated Hamiltonian. However experience has shown that very accurate solutions for the trajectory of the nuclei can be obtained by making some simplifying assumptions. The

most important assumption for us, is that the interaction between the two atoms can be separated into two components:

- ion (projectile)–nucleus (target) interaction
- ion (projectile)–electron (target) interaction.

Let us now use the very simple example of a collision between two masses M_1 and M_2 to determine the relative importance of these two processes and to introduce the basic atomic scattering concepts required to describe the stopping of ions in a solid. Figure 11.5 shows the classical collision between the incident mass M_1 and the target mass M_2 which can be a target atom or a nearly free target electron.

By applying conservation of energy and momentum, the following relations can be derived (you will do it as homework).

- The energy transferred (T) (Ref. H. Goldstein, *Classical Mechanics*, Academic Press (1950)) in the collision from the incident projectile M_1 to the target particle M_2 is given by

$$T = T_{max} \quad \sin^2\left(\frac{\Theta}{2}\right) \tag{11.1}$$

  where

$$T_{max} = 4E_0 \quad \frac{M_1 M_2}{(M_1 + M_2)^2} \tag{11.2}$$

  is the maximum possible energy transfer from M_1 to M_2.

- The scattering angle of the projectile in the laboratory system of coordinates is given by

$$\cos\theta = \frac{1 - (1 + M_2/M_1)(T/2E)}{\sqrt{(1 - T/E)}} \tag{11.3}$$

- The energies of the projectile before (E_0) and after (E_1) scattering are related by

$$E_1 = k^2 E_0 \tag{11.4}$$

  where the kinematic factor k is given by

$$k = \left(\frac{M_1 \cos\theta \pm ({M_2}^2 - {M_1}^2 \sin^2\theta)^{1/2}}{M_1 + M_2}\right) \tag{11.5}$$

These relations are absolutely general no matter how complex the force between the two particles, so long as the force acts along the line joining the particles and the electron is nearly free so that the collision can be taken to be elastic. In reality, the collisions between the projectile and the target electrons are inelastic because of the binding energy of the electrons. The case of inelastic collisions will be considered in what follows.

Using Eqs. (11.2) and (11.3) and assuming either that M_2 is of the same atomic species as the projectile ($M_2 = M_1$), or that M_2 is a nearly free electron ($M_2 = m_e$) we construct Table 11.1.

With the help of the previous example, we can formulate a qualitative picture of the slowing down process of the incident energetic ions. As the incident ions penetrate into the solid, they lose energy. There are two dominant mechanisms for this energy loss:

Table 11.1: Maximum energy transfer T_{max} and scattering angle θ for nuclear ($M_2 = M_1$) and electronic ($M_2 = m_e$) collisions.

	"nuclear" collision	"electronic" collision
T_{max}	$T_{max} \simeq E_0$	$T_{max} \simeq (4m/M_1)\, E_0$
θ	$0 < \theta \leq \pi/2$	$\theta = 0^\circ$

1. The interaction between the incident ion and the electrons of the host material. This inelastic scattering process gives rise to electronic energy loss.

2. The interaction between the incident ions and the nuclei of the host material. This is an elastic scattering process which gives rise to nuclear energy loss.

The ion–electron interaction (see Table 11.1) induces small losses in the energy of the incoming ion as the electrons in the atom are excited to higher bound states or are ionized. These interactions do not produce significant deviations in the projectile trajectory. In contrast, the ion–nucleus interaction results in both energy loss and significant deviation in the projectile trajectory. In the ion–nucleus interaction, the atoms of the host are also significantly dislodged from their original positions giving rise to lattice defects, and the deviations in the projectile trajectory will give rise to the lateral spread of the distribution of implanted species.

Let us now further develop our example of a "nuclear" collision. For a given interaction potential $V(r)$, each ion coming into the annular ring of area $2\pi pdp$ with energy E, will be deflected through an angle θ where p is the impact parameter (see Fig. 11.6). We define $T = E_0 - E_1$ as the energy transfer from the incoming ion to the host and we define $2\pi pdp = d\sigma$ as the differential cross section. When the ion moves a distance Δx in the host material, it will interact with $N\Delta x 2\pi pdp$ atoms where N is the atom density of the host.

The energy ΔE lost by an ion traversing a distance Δx will be

$$\Delta E = N\Delta x \int T\, 2\pi pdp \tag{11.6}$$

so that as $\Delta x \to 0$, we have for the stopping power

$$\frac{dE}{dx} = N \int T d\sigma \tag{11.7}$$

where σ denotes the cross sectional area. We thus obtain for the stopping cross section E

$$E = \frac{1}{N}\frac{dE}{dx} = \int T d\sigma. \tag{11.8}$$

The total stopping power is due to both electronic and nuclear processes

$$\frac{dE}{dx} = \left(\frac{dE}{dx}\right)_e + \left(\frac{dE}{dx}\right)_n = N(E_e + E_n) \tag{11.9}$$

where N is the target density and E_e and E_n are the electronic and nuclear stopping cross sections, respectively. Likewise for the stopping cross section E we can write

$$E = E_e + E_n. \tag{11.10}$$

Ion	E_1(keV)	E_2(keV)	E_3(keV)
B	3	17	3000
P	17	140	$\sim 3 \times 10^4$
As	73	800	$> 10^5$
Sb	180	2000	$> 10^5$

Table 11.2: Typical Values of E_1, E_2, E_3 for silicon.[a]

[a]See Fig. 11.7 for the definition of the notation.

From the energy loss we can obtain the ion range or penetration depth

$$R = \int \frac{dE}{dE/dx}. \tag{11.11}$$

Since we know the energy transferred to the lattice (including both phonon generation and displacements of the host ions), we can calculate the energy of the incoming ions as a function of distance into the medium $E(x)$.

At low energies of the projectile ion, nuclear stopping is dominant, while electron stopping dominates at high energies as shown in the characteristic stopping power curves of Figs. 11.7 and 11.8. Note the three important energy parameters on the curves shown in Fig. 11.7: E_1 is the energy where the nuclear stopping power is a maximum, E_3 where the electronic stopping power is a maximum, and E_2 where the electronic and nuclear stopping powers are equal. As the atomic number of the ion increases for a fixed target, the scale of E_1, E_2 and E_3 increases. Also indicated on the diagram is the functional form of the energy dependence of the stopping power in several of the regimes of interest. Typical values of the parameters E_1, E_2 and E_3 for various ions in silicon are given in Table 11.2. Ion implantation in semiconductors is usually done in the regime where nuclear energy loss is dominant. The region in Fig. 11.7 where $(dE/dx) \sim 1/E$ corresponds to the regime where light ions like H^+ and He^+ have incident energies of 1–2 MeV and is therefore the region of interest for Rutherford backscattering and channeling phenomena.

11.2.2 Radiation Damage

The energy transferred from the projectile ion to the target atom is usually sufficient to result in the breaking of a chemical bond and the permanent displacement of the target atom from its original site (see Fig. 11.9). The condition for this process is that the energy transfer per collision T is greater than the binding energy E_d.

Because of the high incident energy of the projectile ions, each incident ion can dislodge multiple host ions. The damage profile for low dose implantation gives rise to isolated regions of damage as shown in Fig. 11.10.

As the fluence is increased, these damaged regions coalesce as shown in Fig. 11.10. The damage profile also depends on the mass of the projectile ion, with heavy mass ions of a given energy causing more local lattice damage as the ions come to rest. Since $(dE/dx)_n$ increases as the energy decreases, more damage is caused as the ions are slowed down and come to rest. The damage pattern is shown in Fig. 11.11 schematically for light ions (such

as boron in silicon) and for heavy ions (such as antimony in silicon). Damage is caused both by the incident ions and by the displaced energetic (knock-on) ions.

A schematic diagram of the types of defects caused by ion implantation is shown in Fig. 11.12. Here we see the formation of vacancies and interstitials, Frenkel pairs (the pair formed by the Coulomb attraction of a vacancy and an interstitial). The formation of multiple vacancies leads to a depleted zone while multiple interstitials lead to ion crowding.

11.2.3 Applications of Ion Implantation

For the case of semiconductors, ion implantation is dominantly used for doping purposes, to create sharp *p*-*n* junctions in the near-surface region. To reduce radiation damage, implantation is sometimes done at elevated temperatures. Post implantation annealing is also used to reduce radiation damage, with elevated temperatures provided by furnaces, lasers or flash lamps. The ion implanted samples are characterized by a variety of experimental techniques for the implant depth profile, the lattice location of the implant, the residual lattice disorder subsequent to implantation and annealing, the electrical properties (Hall effect and conductivity) and the device performance.

A major limitation of ion implantation for modifying metal surfaces has been the shallow depth of implantation. In addition, the sputtering of atoms from the surface sets a maximum concentration of elements which can be added to a solid, typically $\sim$ 20 to 40 at.%. To form thicker layers and higher concentrations, combined processes involving ion implantation and film deposition are being investigated. Intense ion beams are directed at the solid to bring about alloying while other elements are simultaneously brought to the surface, for example by sputter deposition, vapor deposition or the introduction of reactive gases. One process of interest is ion beam mixing, where thin films are deposited onto the surface first and then bombarded with ions. The dense collision cascades of the ions induce atomic–scale mixing between elements. Ion beam mixing is also a valuable tool to study metastable phase formation.

With regard to polymers, ion implantation can enhance the electrical conductivity by many orders of magnitude, as is for example observed (see Fig. 11.13) for ion implanted polyacrylonitrile (PAN, a graphite fiber precursor). Some of the attendant property changes of polymers due to ion implantation include cross-linking and scission of polymer chains, gas evolution as volatile species are released from polymer chains and free radical formation when vacancies or interstitials are formed. Implantation produces solubility changes in polymers and therefore can be utilized for the patterning of resists for semiconductor mask applications. For the positive resists, implantation enhances the solubility, while for negative resists, the solubility is reduced. The high spatial resolution of the ion beams makes ion beam lithography a promising technique for sub–micron patterning applications. For selected polymers (such as PAN), implantation can result in transforming a good insulator into a conducting material with an increase in conductivity by more than 10 orders of magnitude upon irradiation. Thermoelectric power measurements on various implanted polymers show that implantation can yield either *p*–type or *n*–type conductors and in fact a *p*–*n* junction has recently been made in a polymer through ion implantation (T. Wada, A. Takeno, M. Iwake, H. Sasabe, and Y. Kobayashi, *J. Chem. Soc. Chem. Commun.*, **17**, 1194 (1985)). The temperature dependence of the conductivity for many implanted

polymers is of the form $\sigma = \sigma_o \exp(T_0/T)^{1/2}$ which is also the relation characteristic of the one–dimensional hopping conductivity model for disordered materials. Also of interest is the long term chemical stability of implanted polymers.

Due to recent developments of high brightness ion sources, focused ion beams to submicron dimensions can now be routinely produced, using ions from a liquid metal source. Potential applications of this technology are to ion beam lithography, including the possibility of maskless implantation doping of semiconductors. Instruments based on these ideas may be developed in the future. The applications of ion implantation represent a rapidly growing field.

Instruments based on these ideas may be developed in the future. The applications of ion implantation represent a rapidly growing field.

Further discussion of the application of ion implantation to the preparation of metastable materials is presented after the following sections on the characterization of ion implanted materials by ion backscattering and channeling.

11.3 Ion Backscattering

In Rutherford backscattering spectrometry (RBS), a beam of mono-energetic (1–2 MeV), collimated light mass ions (H^+, He^+) impinges (usually at near normal incidence) on a target and the number and energy of the particles that are scattered backwards at a certain angle θ are monitored (as shown in Fig. 11.14) to obtain information about the composition of the target (host species and impurities) as a function of depth. With the help of Fig. 11.15 we will review the fundamentals of the RBS analysis.

Particles scattered at the surface of the target will have the highest energy E upon detection. Here the energy of the backscattered ions E is given by the relation

$$E = k^2 E_0 \tag{11.12}$$

where

$$k = \left(\frac{M_1 \cos\theta \pm (M_2{}^2 - M_1{}^2 \sin^2\theta)^{1/2}}{M_1 + M_2} \right) \tag{11.13}$$

as discussed in Section 11.2. For a given mass species, the energy E_s of particles scattered from the surface corresponds to the edge of the spectrum (see Fig. 11.15). In addition, the scattered energy depends through k on the mass of the scattering atom. Thus different species will appear displaced on the energy scale of Fig. 11.15, thereby allowing for their chemical identification. We next show that the displacement along the energy scale from the surface contribution gives information about the depth where the backscattering took place. Thus the energy scale is effectively a depth scale.

The height H of the RBS spectrum corresponds to the number of detected particles in each energy channel ΔE.

11.4 Channeling

If the probing beam is aligned nearly parallel to a close-packed row of atoms in a single crystal target, the particles in the beam will be steered by the potential field of the rows of

atoms, resulting in an undulatory motion in which the "channeled" ions will not approach the atoms in the row to closer than 0.1–0.2 Å. This is called the channeling effect (D.V. Morgan, *Channeling* (Wiley, 1973)). Under this channeling condition, the probability of large angle scattering is greatly reduced. As a consequence, there will be a drastic reduction in the scattering yield from a channeled probing beam relative to the yield from a beam incident in a random direction (see Fig. 11.16). Two characteristic parameters for channeling are the normalized minimum yield $\chi_{min} = H_A/H$ which is a measure of the crystallinity of the target, and the critical angle for channeling $\psi_{1/2}$ (the halfwidth at half maximum intensity of the channeling resonance) which determines the degree of alignment required to observed the channeling effect.

The RBS–channeling technique is frequently used to study radiation–induced lattice disorder by measuring the fraction of atom sites where the channel is blocked.

In general, the channeled ions are steered by the rows of atoms in the crystal. However if some portion of the crystal is disordered and lattice atoms are displaced so that they partially block the channels, the ions directed along nominal channeling directions can now have a close collision with these displaced atoms, so that the resulting scattered yield will be increased above that for an undisturbed channel. Furthermore, since the displaced atoms are of equal mass to those of the surrounding lattice, the increase in the yield occurs at a position in the yield *vs.* energy spectrum corresponding to the depth at which the displaced atoms are located. The increase in the backscattering yield from a given depth will depend upon the number of displaced atoms, so the depth (or equivalently, the backscattering energy E) dependence of the yield, reflects the depth dependence of displaced atoms, and integrations over the whole spectrum will give a measure of the total number of displaced atoms. This effect is shown schematically in Fig. 11.17.

Another very useful application of the RBS–channeling technique is in the determination of the location of foreign atoms in a host lattice. Since channeled ions cannot approach the rows of atoms which form the channel closer than ~ 0.1Å, we can think of a "forbidden region", as a cylindrical region along each row of atoms with radius ~ 0.1Å, such that there are no collisions between the channeled particles and atoms located within the forbidden zone. In particular, if an impurity is located in a forbidden region it will not be detected by the channeled probing beam. On the other hand, any target particle can be detected by (i.e., will scatter off) probing particles from a beam which impinges in a random direction. Thus, by comparing the impurity peak observed for channeling and random alignments, the fraction of impurities sitting in the forbidden region of a particular channel (high symmetry crystallographic axis) can be determined. Repeating the procedure for other crystallographic directions allows the identification of the lattice location of the impurity atom in many cases.

Rutherford backscattering will not always reveal impurities embedded in a host matrix, in particular if the mass of the impurities is smaller than the mass of the host atoms. In such cases, ion induced x–rays and ion induced nuclear reactions are used as signatures for the presence of the impurities inside the crystal, and the lattice location is derived from the changes in yield of these processes for random and channeled impingement of the probing beam.

Ion markers consist of a very thin layer of a guest atomic species are embedded in an otherwise uniform host material of a different species to establish reference distances. Backscattering spectra are taken before and after introduction of the marker. The RBS

spectrum taken after insertion of the marker can be used as a reference for various applications. Some examples where marker references are useful include:

- Estimation of surface sputtering by ion implantation. In this case, recession of the surface from the reference position set by the marker (see Fig. 11.18) can be measured by RBS and can be analyzed to yield the implantation–induced surface sputtering.
- Estimation of surface material vaporized through laser annealing, rapid thermal annealing or laser melting of a surface.
- Estimation of the extent of ion beam mixing.

Figure 11.1: Schematic illustrating the interactions of ion beams with a single–crystal solid. Directed beams of $\sim$ 10 eV are used for film deposition and epitaxial formation. Ion beams of energy $\sim$ 1 keV are employed in sputtering applications; $\sim$ 100 keV ions are used in ion implantation. Both the sputtering and implantation processes damage and disorder the crystal. Higher energy light ions are used for ion beam analysis.

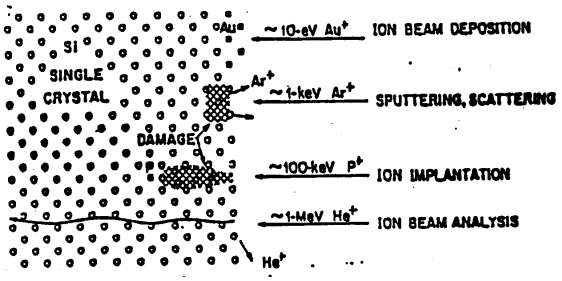
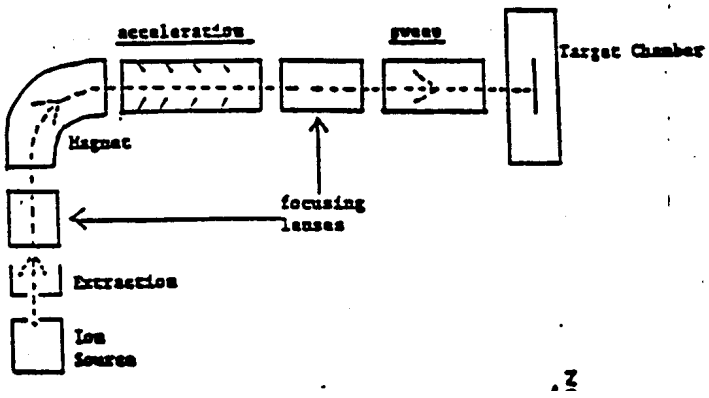


Figure 11.2: Schematic Diagram of Ion Implanter.



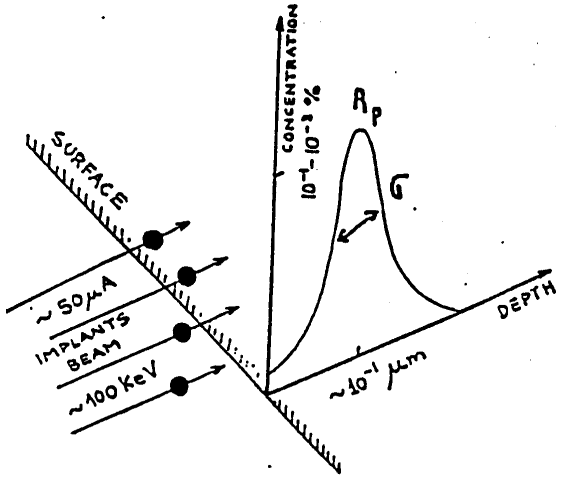


Figure 11.3: Typical ion implantation parameters.

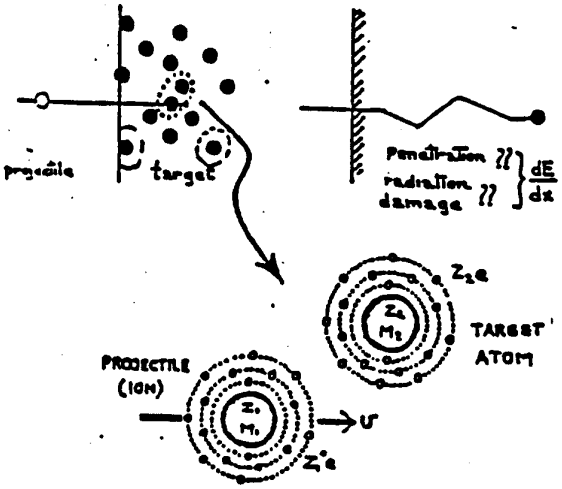


Figure 11.4: Penetration of ions into solids.

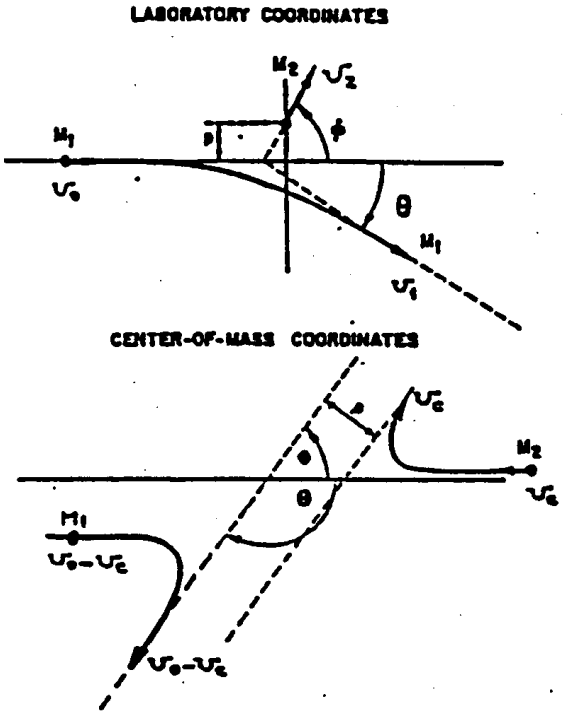


Figure 11.5: The upper figure defines the scattering variables in a two–body collision. The projectile has mass M_1 and an initial velocity v_0, and an impact parameter, p, with the target particle. The projectile's final angle of deflection is θ and its final velocity is v_1. The target particle with mass, M_2, recoils at an angle ϕ with velocity v_2. The lower figure is the same scattering event in the center–of–mass (CM) coordinates in which the *total momentum of the system is zero.* The coordinate system moves with velocity v_c relative to the laboratory coordinates, and the angles of scatter and recoil are Θ and Φ in the center of mass system.

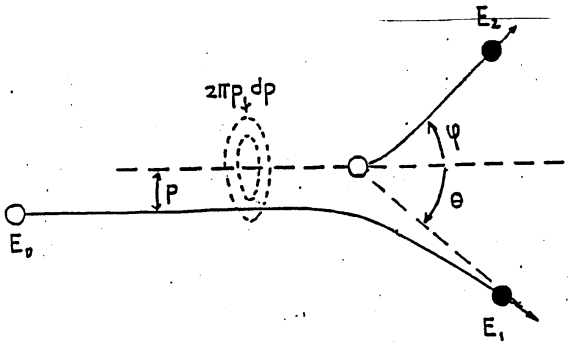


Figure 11.6: Classical model for the collision of a projectile of energy E with a target at rest. The open circles denote the initial state of the projectile and target atoms, and the full circles denote the two atoms after the collision.

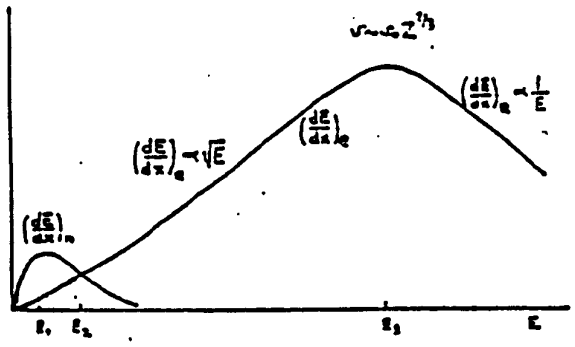


Figure 11.7: Nuclear and electronic energy loss (stopping power) *vs.* Energy.

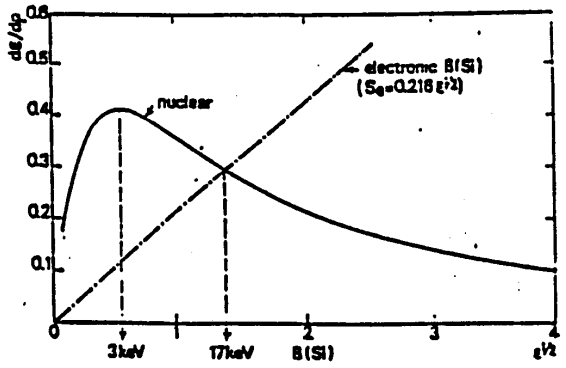


Figure 11.8: Nuclear and electronic energy loss (stopping power) *vs.* $\epsilon^{1/2}$ (reduced units of LSS theory).

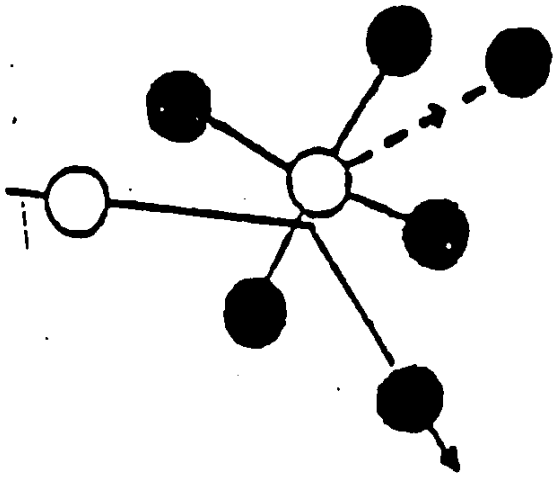


Figure 11.9: Energy transfer from projectile ion to target atoms for a single scattering event for the condition $T > E_d$ where E_d is the binding energy and T is the energy transfer per collision.

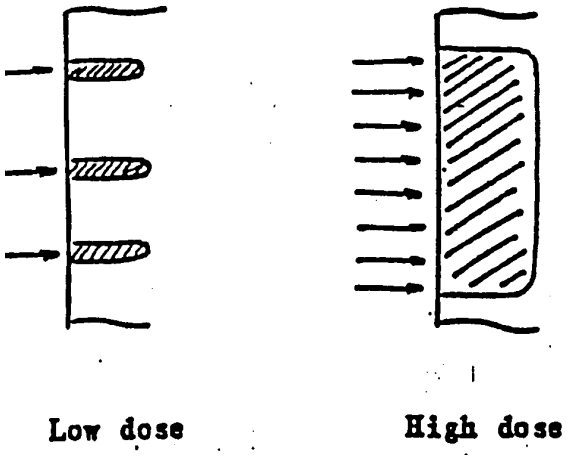


Figure 11.10: Schematic diagram showing the range of lattice damage for low dose and high dose implants.

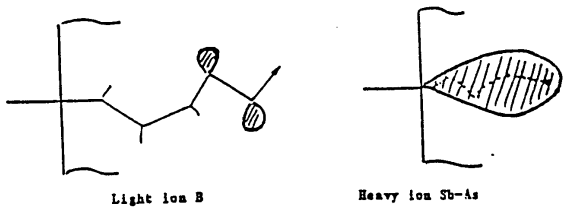


Figure 11.11: Schematic diagram of the damage pattern for light ions and heavy ions in the same target (silicon).

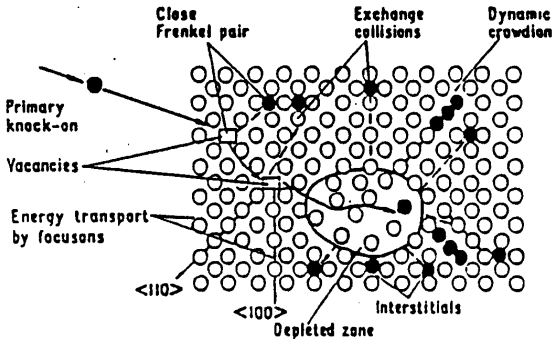
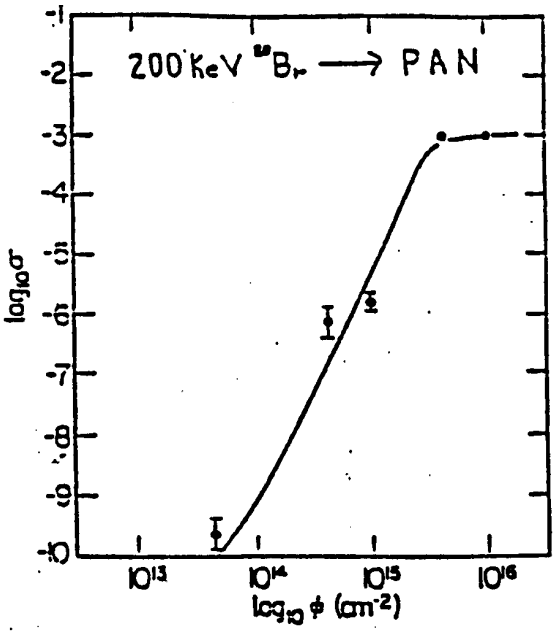


Figure 11.12: Example of typical defects induced by ion implantation.

Figure 11.13: Implantation induced conductivity of a normally insulating polymer.



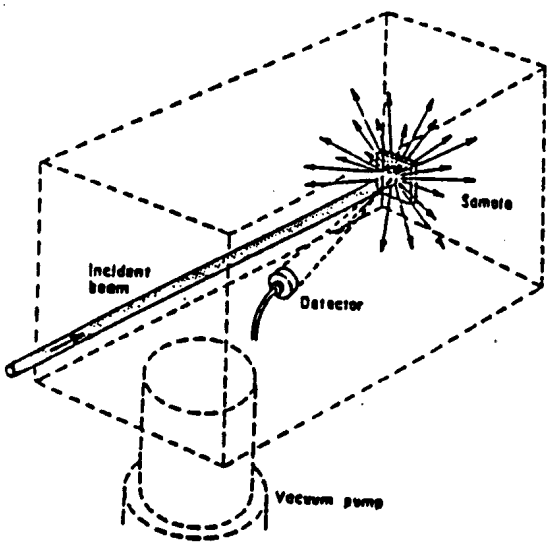


Figure 11.14: In the RBS experiment, the scattering chamber where the analysis/experiment is actually performed contains the essential elements: the sample, the beam, the detector, and the vacuum pump.

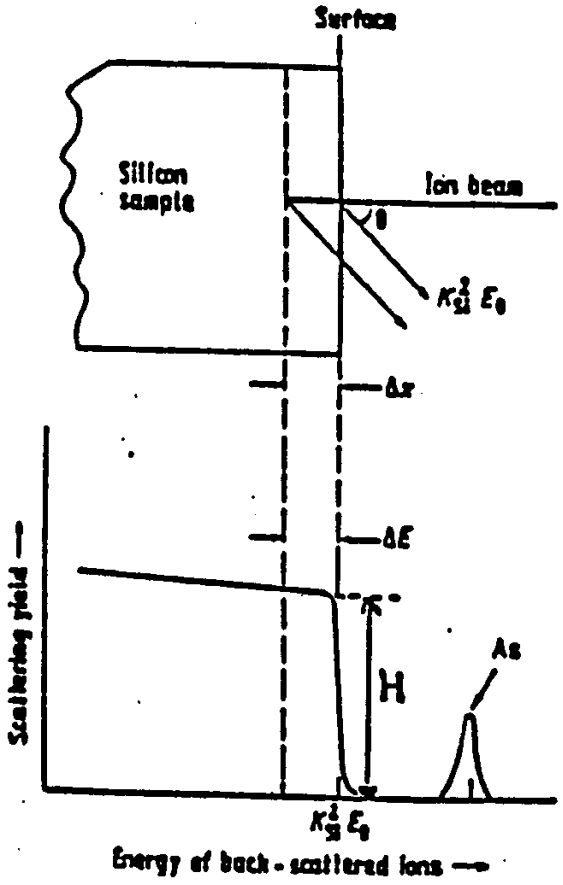


Figure 11.15: Schematic diagram showing the energy distribution of ions back-scattered from a Si sample (not aligned) which was implanted with As atoms.

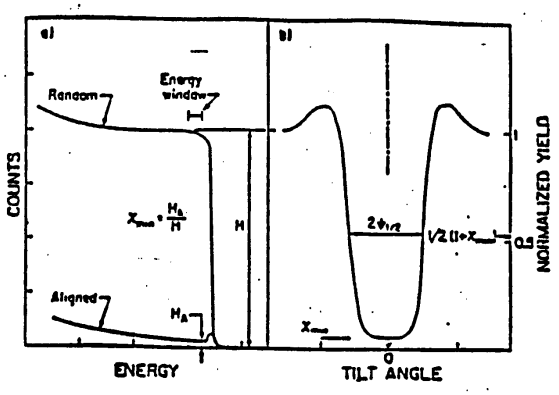


Figure 11.16: Schematic backscattering spectrum and angular yield profile.

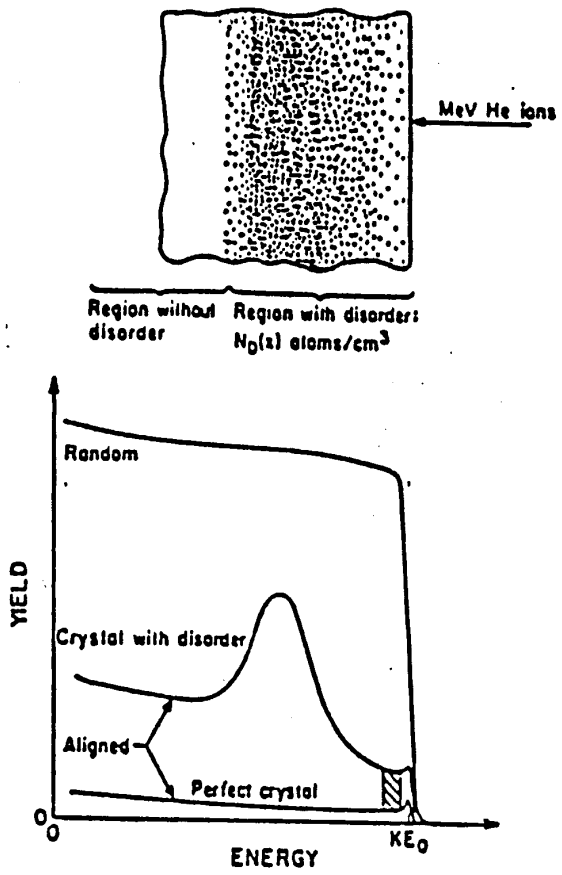


Figure 11.17: Schematic random and aligned spectra for MeV ^{4}He ions incident on a crystal containing disorder. The aligned spectrum for a perfect crystal without disorder is shown for comparison. The difference (shaded portion) in the aligned spectra between disordered and perfect crystals can be used to determine the concentration $N_D(0)$ of displaced atoms at the surface.

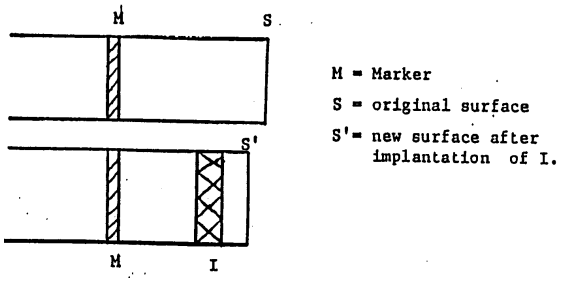


Figure 11.18: Schematic of the marker experiment which demonstrates surface recession through surface sputtering.

Appendix A

Time–Independent Perturbation Theory

References

- Davydov - *Quantum Mechanics*, Ch. 7.
- Morse and Feshbach, *Methods of Theoretical Physics*, Ch. 9.
- Shankar, *Principles of Quantum Mechanics*, Ch. 17.
- Cohen-Tannoudji, Diu and Laloë, *Quantum Mechanics*, vol. 2, Ch. 11.
- T-Y. Wu, Quantum Mechanics, Ch. 6.

A.1 Introduction

Another review topic that we discuss here is time–independent perturbation theory because of its importance in experimental solid state physics in general and transport properties in particular.

There are many mathematical problems that occur in nature that cannot be solved exactly. It also happens frequently that a *related* problem can be solved *exactly*. Perturbation theory gives us a method for relating the problem that can be solved exactly to the one that cannot. This occurrence is more general than quantum mechanics –many problems in electromagnetic theory are handled by the techniques of perturbation theory. In this course however, we will think mostly about quantum mechanical systems, as occur typically in solid state physics.

Suppose that the Hamiltonian for our system can be written as

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}' \tag{A.1}$$

where $\mathcal{H}_0$ is the part that we can solve exactly and $\mathcal{H}'$ is the part that we cannot solve. Provided that $\mathcal{H}' \ll \mathcal{H}_0$ we can use perturbation theory; that is, we consider the solution of the unperturbed Hamiltonian $\mathcal{H}_0$ and then calculate the effect of the perturbation Hamiltonian $\mathcal{H}'$. For example, we can solve the hydrogen atom energy levels exactly, but when we apply an electric or a magnetic field we can no longer solve the problem exactly. For

this reason, we treat the effect of the external fields as a perturbation, provided that the energy associated with these fields is small:

$$\mathcal{H} = \frac{p^2}{2m} - \frac{e^2}{r} - e\vec{r}\cdot\vec{E} = \mathcal{H}_0 + \mathcal{H}' \tag{A.2}$$

where

$$\mathcal{H}_0 = \frac{p^2}{2m} - \frac{e^2}{r} \tag{A.3}$$

and

$$\mathcal{H}' = -e\vec{r}\cdot\vec{E}. \tag{A.4}$$

As another illustration of an application of perturbation theory, consider a weak periodic potential in a solid. We can calculate the free electron energy levels (empty lattice) exactly. We would like to relate the weak potential situation to the empty lattice problem, and this can be done by considering the weak periodic potential as a perturbation.

A.1.1 Non-degenerate Perturbation Theory

In non-degenerate perturbation theory we want to solve Schrödinger's equation

$$\mathcal{H}\psi_n = E_n\psi_n \tag{A.5}$$

where

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}' \tag{A.6}$$

and

$$\mathcal{H}' \ll \mathcal{H}_0. \tag{A.7}$$

It is then assumed that the solutions to the unperturbed problem

$$\mathcal{H}_0\psi_n^0 = E_n^0\psi_n^0 \tag{A.8}$$

are known, in which we have labeled the unperturbed energy by E_n^0 and the unperturbed wave function by ψ_n^0. By *non-degenerate* we mean that there is only one eigenfunction ψ_n^0 associated with each eigenvalue E_n^0.

The wave functions ψ_n^0 form a complete orthonormal set

$$\int \psi_n^{*0}\psi_m^0 d^3r = \langle\psi_n^0|\psi_m^0\rangle = \delta_{nm}. \tag{A.9}$$

Since $\mathcal{H}'$ is small, the wave functions for the total problem ψ_n do not differ greatly from the wave functions ψ_n^0 for the unperturbed problem. So we expand $\psi_{n'}$ in terms of the complete set of ψ_n^0 functions

$$\psi_{n'} = \sum_n a_n\psi_n^0. \tag{A.10}$$

Such an expansion can always be made; that is no approximation. We then substitute the expansion of Eq. A.10 into Schrödinger's equation (Eq. A.5) to obtain

$$\mathcal{H}\psi_{n'} = \sum_n a_n(\mathcal{H}_0 + \mathcal{H}')\psi_n^0 = \sum_n a_n(E_n^0 + \mathcal{H}')\psi_n^0 = E_{n'}\sum_n a_n\psi_n^0 \tag{A.11}$$

and therefore we can write

$$\sum_n a_n(E_{n'} - E_n^0)\psi_n^0 = \sum_n a_n \mathcal{H}'\psi_n^0. \tag{A.12}$$

If we are looking for the perturbation to the level m, then we multiply Eq. A.12 from the left by ψ_m^{0*} and integrate over all space. On the left hand side of Eq. A.12 we get $\langle\psi_m^0|\psi_n^0\rangle = \delta_{mn}$ while on the right hand side we have the matrix element of the perturbation Hamiltonian taken between the unperturbed states:

$$a_m(E_{n'} - E_m^0) = \sum_n a_n\langle\psi_m^0|\mathcal{H}'|\psi_n^0\rangle \equiv \sum_n a_n\mathcal{H}'_{mn} \tag{A.13}$$

where we have written the indicated matrix element as $\mathcal{H}'_{mn}$. Equation A.13 is an iterative equation on the a_n coefficients where each a_m coefficient is related to a complete set of a_n coefficients by the relation

$$a_m = \frac{1}{E_{n'} - E_m^0}\sum_n a_n\langle\psi_m^0|\mathcal{H}'|\psi_n^0\rangle = \frac{1}{E_{n'} - E_m^0}\sum_n a_n\mathcal{H}'_{mn} \tag{A.14}$$

in which the summation includes the $n = n'$ and m terms. We can rewrite Eq. A.14 to involve terms in the sum $n \neq m$

$$a_m(E_{n'} - E_m^0) = a_m\mathcal{H}'_{mm} + \sum_{n\neq m} a_n\mathcal{H}'_{mn} \tag{A.15}$$

so that the coefficient a_m is related to all the other a_n coefficients by:

$$a_m = \frac{1}{E_{n'} - E_m^0 - \mathcal{H}'_{mm}}\sum_{n\neq m} a_n\mathcal{H}'_{mn} \tag{A.16}$$

where n' is an index denoting the energy of the state we are seeking. The equation (A.16) written as

$$a_m(E_{n'} - E_m^0 - \mathcal{H}'_{mm}) = \sum_{n\neq m} a_n\mathcal{H}'_{mn} \tag{A.17}$$

is an identity in the a_n coefficients. If the perturbation is small then $E_{n'}$ is very close to E_m^0 and the first order corrections are found by setting the coefficient on the right hand side equal to zero and $n' = m$. The next order of approximation is found by substituting for of a_n on the right hand side of Eq. A.17 and substituting for a_n the expression

$$a_n = \frac{1}{E_{n'} - E_n^0 - \mathcal{H}'_{nn}}\sum_{n''\neq n} a_{n''}\mathcal{H}'_{nn''} \tag{A.18}$$

which is obtained from Eq. A.16 by the transcription $m \to n$ and $n \to n''$. In the above, the energy level $E_{n'} = E_m$ is the level for which we are calculating the perturbation. We now look for the a_m term in the sum $\sum_{n''\neq n} a_{n''}\mathcal{H}'_{nn''}$ of Eq. A.18 and bring it to the left hand side of Eq. A.17. If we are satisfied with our solutions, we end the perturbation calculation at this point. If we are not satisfied, we substitute for the $a_{n''}$ coefficients in Eq. A.18 using the same basic equation as Eq. A.18 to obtain a triple sum. We then select out the a_m term, bring it to the left hand side of Eq. A.17, etc. This procedure gives us an easy recipe to find the energy in Eq. A.11 to any order of perturbation theory. We now write these iterations down more explicitly for first and second order perturbation theory.

1^{st} Order Perturbation Theory

In this case, no iterations of Eq. A.17 are needed and the sum $\sum_{n\neq m} a_n \mathcal{H}'_{mn}$ on the right hand side of Eq. A.17 is neglected, for the reason that if the perturbation is small, $\psi_{n'} \sim \psi^0_n$. Hence only a_m in Eq. A.10 contributes significantly. We merely write $E_{n'} = E_m$ to obtain:

$$a_m(E_m - E^0_m - \mathcal{H}'_{mm}) = 0. \tag{A.19}$$

Since the a_m coefficients are arbitrary coefficients, this relation must hold for all a_m so that

$$(E_m - E^0_m - \mathcal{H}'_{mm}) = 0 \tag{A.20}$$

or

$$E_m = E^0_m + \mathcal{H}'_{mm}. \tag{A.21}$$

We write Eq. A.21 even more explicitly so that the energy for state m for the perturbed problem E_m is related to the unperturbed energy E^0_m by

$$E_m = E^0_m + \langle \psi^0_m | \mathcal{H}' | \psi^0_m \rangle \tag{A.22}$$

where the indicated diagonal matrix element of $\mathcal{H}'$ can be integrated as the average of the perturbation in the state ψ^0_m. The wave functions to lowest order are not changed

$$\psi_m = \psi^0_m. \tag{A.23}$$

2^{nd} order perturbation theory

If we carry out the perturbation theory to the next order of approximation, one further iteration of Eq. A.17 is required:

$$a_m(E_m - E^0_m - \mathcal{H}'_{mm}) = \sum_{n\neq m} \frac{1}{E_m - E^0_n - \mathcal{H}'_{nn}} \sum_{n''\neq n} a_{n''} \mathcal{H}'_{nn''} \mathcal{H}'_{mn} \tag{A.24}$$

in which we have substituted for the a_n coefficient in Eq. A.17 using the iteration relation given by Eq. A.18. We now pick out the term on the right hand side of Eq. A.24 for which $n'' = m$ and bring that term to the left hand side of Eq. A.24. If no further iteration is to be done, we throw away what is left on the right hand side of Eq. A.24 and get an expression for the arbitrary a_m coefficients

$$a_m \left[(E_m - E^0_m - \mathcal{H}'_{mm}) - \sum_{n\neq m} \frac{\mathcal{H}'_{nm}\mathcal{H}'_{mn}}{E_m - E^0_n - \mathcal{H}'_{nn}} \right] = 0. \tag{A.25}$$

Since a_m is arbitrary, the term in square brackets in Eq. A.25 vanishes and the second order correction to the energy results:

$$E_m = E^0_m + \mathcal{H}'_{mm} + \sum_{n\neq m} \frac{|\mathcal{H}'_{mn}|^2}{E_m - E^0_n - \mathcal{H}'_{nn}} \tag{A.26}$$

in which the sum on states $n \neq m$ represents the 2nd order correction.

To this order in perturbation theory we must also consider corrections to the wave function

$$\psi_m = \sum_n a_n \psi_n^0 = \psi_m^0 + \sum_{n \neq m} a_n \psi_n^0 \tag{A.27}$$

in which ψ_m^0 is the large term and the correction terms appear as a sum over all the other states $n \neq m$. In handling the correction term, we look for the a_n coefficients, which from Eq. A.18 are given by

$$a_n = \frac{1}{E_n' - E_n^0 - \mathcal{H}_{nn}'} \sum_{n'' \neq n} a_{n''} \mathcal{H}_{nn''}'. \tag{A.28}$$

If we only wish to include the lowest order correction terms, we will take only the most important term, i.e., $n'' = m$, and we will also use the relation $a_m = 1$ in this order of approximation. Again using the identification $n' = m$, we obtain

$$a_n = \frac{\mathcal{H}_{nm}'}{E_m - E_n^0 - \mathcal{H}_{nn}'} \tag{A.29}$$

and

$$\psi_m = \psi_m^0 + \sum_{n \neq m} \frac{\mathcal{H}_{nm}' \psi_n^0}{E_m - E_n^0 - \mathcal{H}_{nn}'}. \tag{A.30}$$

For homework, you should do the next iteration to get 3rd order perturbation theory, in order to see if you really have mastered the technique (this will be an optional homework problem).

Now look at the results for the energy E_m (Eq. A.26) and the wave function ψ_m (Eq. A.30) for the 2nd order perturbation theory and observe that these solutions are implicit solutions. That is, the correction terms are themselves dependent on E_m. To obtain an explicit solution, we can do one of two things at this point.

1. We can ignore the fact that the energies differ from their unperturbed values in calculating the correction terms. This is known as Raleigh-Schrödinger perturbation theory. This is the usual perturbation theory given in Quantum Mechanics texts and for homework you may review the proof as given in these texts.

2. We can take account of the fact that E_m differs from E_m^0 by calculating the correction terms by an iteration procedure; the first time around, you substitute for E_m the value that comes out of 1st order perturbation theory. We then calculate the second order correction to get E_m. We next take this E_m value to compute the new second order correction term etc. until a convergent value for E_m is reached. This iterative procedure is what is used in *Brillouin–Wigner* perturbation theory and is a better approximation than Raleigh-Schrödinger perturbation theory to both the wave function and the energy eigenvalue for the same order in perturbation theory.

The Brillouin–Wigner method is often used for practical problems in solids. For example, if you have a 2-level system, the Brillouin–Wigner perturbation theory to second order gives an exact result, whereas Rayleigh–Schrödinger perturbation theory must be carried out to infinite order.

Let us summarize these ideas. If you have to compute only a small correction by perturbation theory, then it is advantageous to use Rayleigh-Schrödinger perturbation theory

because it is much easier to use, since no iteration is needed. If one wants to do a more convergent perturbation theory (i.e., obtain a better answer to the same order in perturbation theory), then it is advantageous to use Brillouin–Wigner perturbation theory. There are other types of perturbation theory that are even more convergent and harder to use than Brillouin–Wigner perturbation theory (see Morse and Feshbach vol. 2). But these two types are the most important methods used in solid state physics today.

For your convenience we summarize here the results of the second–order non–degenerate Rayleigh-Schrödinger perturbation theory:

$$E_m = E_m^0 + \mathcal{H}'_{mm} + \sum_n{}' \frac{|\mathcal{H}'_{nm}|^2}{E_m^0 - E_n^0} + ... \tag{A.31}$$

$$\psi_m = \psi_m^0 + \sum_n{}' \frac{\mathcal{H}'_{nm}\psi_n^0}{E_m^0 - E_n^0} + \tag{A.32}$$

where the sums in Eqs. A.31 and A.32 denoted by primes exclude the $m = n$ term. Thus, Brillouin–Wigner perturbation theory (Eqs.A.26 and A.30) contains contributions in second order which occur in higher order in the Rayleigh-Schrödinger form. In practice, Brillouin–Wigner perturbation theory is useful when the perturbation term is too large to be handled conveniently by Rayleigh–Schrödinger perturbation theory but still small enough for perturbation theory to work insofar as the perturbation expansion forms a convergent series.

A.1.2 Degenerate Perturbation Theory

It often happens that a number of quantum mechanical levels have the same or nearly the same energy. If they have exactly the same energy, we know that we can make any linear combination of these states that we like and get a new eigenstate also with the same energy. In the case of degenerate states, we have to do perturbation theory a little differently, as described in the following section.

Suppose that we have an f-fold degeneracy (or near-degeneracy) of energy levels

$$\underbrace{\psi_1^0, \psi_2^0, ...\psi_f^0}_{\text{states with the same or nearly the same energy}} \qquad \underbrace{\psi_{f+1}^0, \psi_{f+2}^0, ...}_{\text{states with quite different energies}}$$

We will call the set of states with the same (or approximately the same) energy a "nearly degenerate set" (NDS). In the case of degenerate sets, the iterative Eq. A.17 still holds. The only difference is that for the degenerate case we solve for the perturbed energies by a different technique, as described below.

Starting with Eq. A.17, we now bring to the left-hand side of the iterative equation all terms involving the f energy levels that are in the NDS. If we wish to calculate an energy within the NDS in the presence of a perturbation, we consider all the a_n's within the NDS as large, and those outside the set as small. To first order in perturbation theory, we ignore the coupling to terms outside the NDS and we get f linear homogeneous equations in the a_n's where $n = 1, 2, ...f$. We thus obtain the following equations from Eq. A.17:

$$\begin{array}{ccccc}
a_1(E_1^0 + \mathcal{H}'_{11} - E) & +a_2\mathcal{H}'_{12} & +... & +a_f\mathcal{H}'_{1f} & = 0 \\
a_1\mathcal{H}'_{21} & +a_2(E_2^0 + \mathcal{H}'_{22} - E) & +... & +a_f\mathcal{H}'_{2f} & = 0 \\
\vdots & \vdots & \ddots & \vdots & \\
a_1\mathcal{H}'_{f1} & +a_2\mathcal{H}'_{f2} & +\ldots & +a_f(E_f^0 + \mathcal{H}'_{ff} - E) & = 0.
\end{array} \tag{A.33}$$

In order to have a solution of these f linear equations, we demand that the coefficient determinant vanish:

$$\begin{vmatrix} (E_1^0 + \mathcal{H}'_{11} - E) & \mathcal{H}'_{12} & \mathcal{H}'_{13} & \dots & \mathcal{H}'_{1f} \\ \mathcal{H}'_{21} & (E_2^0 + \mathcal{H}'_{22} - E) & \mathcal{H}'_{23} & \dots & \mathcal{H}'_{2f} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{H}'_{f1} & \mathcal{H}'_{f2} & \dots & \dots & (E_f^0 + \mathcal{H}'_{ff} - E) \end{vmatrix} = 0 \tag{A.34}$$

The f eigenvalues that we are looking for are the eigenvalues of the matrix in Eq. A.34 and the set of orthogonal states are the corresponding eigenvectors. Remember that the matrix elements $\mathcal{H}'_{ij}$ that occur in the above determinant are taken between the unperturbed states in the NDS.

The generalization to second order degenerate perturbation theory is immediate. In this case, Eqs. A.33 and A.34 have additional terms. For example, the first relation in Eq. A.33 would then become

$$a_1(E_1^0 + \mathcal{H}'_{11} - E) + a_2\mathcal{H}'_{12} + a_3\mathcal{H}'_{13} + \ldots + a_f\mathcal{H}'_{1f} = -\sum_{n \neq NDS} a_n\mathcal{H}'_{1n} \tag{A.35}$$

and for the a_n in the sum in Eq. A.35, which are now small (because they are outside the NDS), we would use our iterative form

$$a_n = \frac{1}{E - E_n^0 - \mathcal{H}'_{nn}} \sum_{m \neq n} a_m \mathcal{H}'_{nm}. \tag{A.36}$$

But we must only consider the terms in the above sum which are large; these terms are all in the NDS. This argument shows that every term on the left side of Eq. A.35 will have a correction term. For example the correction term to a general coefficient a_i will look as follows:

$$a_i\mathcal{H}'_{1i} + a_i \sum_{n \neq NDS} \frac{\mathcal{H}'_{1n}\mathcal{H}'_{ni}}{E - E_n^0 - \mathcal{H}'_{nn}} \tag{A.37}$$

where the first term is the original term from 1st order degenerate perturbation theory and the term from states outside the NDS gives the 2nd order correction terms. So, if we are doing higher order degenerate perturbation theory, we write for each entry in the secular equation the appropriate correction terms (Eq. A.37) that are obtained from these iterations. For example, in 2nd order degenerate perturbation theory, the (1,1) entry to the matrix in Eq. A.34 would be

$$E_1^0 + \mathcal{H}'_{11} + \sum_{n \neq NDS} \frac{|\mathcal{H}'_{1n}|^2}{E - E_n^0 - \mathcal{H}'_{nn}} - E. \tag{A.38}$$

As a further illustration let us write down the (1,2) entry:

$$\mathcal{H}'_{12} + \sum_{n \neq NDS} \frac{\mathcal{H}'_{1n}\mathcal{H}'_{n2}}{E - E_n^0 - \mathcal{H}'_{nn}}. \tag{A.39}$$

Again we have an implicit dependence of the 2nd order term in Eqs. A.38 and A.39 on the energy eigenvalue that we are looking for. To do 2nd order degenerate perturbation we again

have two options. If we take the energy E in Eqs. A.38 and A.39 as the unperturbed energy in computing the correction terms, we have 2nd order degenerate Rayleigh-Schrödinger perturbation theory. On the other hand, if we iterate to get the best correction term, then we call it Brillouin–Wigner perturbation theory.

How do we know in an actual problem when to use degenerate 1st or degenerate 2nd order perturbation theory? If the matrix elements $\mathcal{H}'_{ij}$ coupling members of the NDS vanish, then we must go to 2nd order. Generally speaking, the first order terms will be much larger than the 2nd order terms, provided that there is no symmetry reason for the first order terms to vanish.

Let us explain this further. By the matrix element $\mathcal{H}'_{12}$ we mean $(\psi_1^0|\mathcal{H}'|\psi_2^0)$. Suppose the perturbation Hamiltonian $\mathcal{H}'$ under consideration is due to an electric field $\vec{E}$

$$\mathcal{H}' = -e\vec{r}\cdot\vec{E} \tag{A.40}$$

where $e\vec{r}$ is the dipole moment of our system. If now we consider the effect of inversion on $\mathcal{H}'$, we see that $\vec{r}$ changes sign under inversion $(x, y, z) \rightarrow -(x, y, z)$, i.e., $\vec{r}$ is an odd function. Suppose that we are considering the energy levels of the hydrogen atom in the presence of an electric field. We have s states (even), p states (odd), d states (even), etc. The electric dipole moment will only couple an even state to an odd state because of the oddness of the dipole moment under inversion. Hence there is no effect in 1^{st} order non–degenerate perturbation theory. For the $n = 1$ level, there is an effect due to the electric field in second order so that the correction to the energy level goes as the square of the electric field, i.e., $|\vec{E}|^2$. For the n =2 levels, we treat them in degenerate perturbation theory because the $2s$ and $2p$ states are degenerate in the simple treatment of the hydrogen atom. Here, first order terms only appear in entries coupling s and p states. To get corrections which split the p levels among themselves, we must go to 2nd order degenerate perturbation theory.

Appendix B

Harmonic Oscillators, Phonons, and Electron-Phonon Interaction

B.1 Harmonic Oscillators

In this section we review the solution of the harmonic oscillator problem in quantum mechanics using raising and lowering operators. This is aimed at providing a quick review as background for the lecture on phonon scattering processes and other topics in this course.

The Hamiltonian for the harmonic oscillator in one-dimension is written as:

$$\mathcal{H} = \frac{p^2}{2m} + \frac{1}{2}\kappa x^2. \tag{B.1}$$

We know classically that the frequency of oscillation is given by $\omega = \sqrt{\kappa/m}$ so that

$$\mathcal{H} = \frac{p^2}{2m} + \frac{1}{2}m\omega^2 x^2 \tag{B.2}$$

Define the lowering and raising operators a and $a^\dagger$ respectively by

$$a = \frac{p - im\omega x}{\sqrt{2\hbar m\omega}} \tag{B.3}$$

$$a^\dagger = \frac{p + im\omega x}{\sqrt{2\hbar m\omega}} \tag{B.4}$$

Since $[p, x] = \hbar/i$, then $[a, a^\dagger]$ =1 so that

$$\mathcal{H} = \frac{1}{2m}\Big[(p + i\omega mx)(p - i\omega mx) + m\hbar\omega\Big] \tag{B.5}$$

$$= \hbar\omega[a^\dagger a + 1/2]. \tag{B.6}$$

Let $N = a^\dagger a$ denote the number operator and its eigenstates $|n\rangle$ so that $N|n\rangle = n|n\rangle$ where n is any real number. However

$$\langle n|N|n\rangle = \langle n|a^\dagger a|n\rangle = \langle y|y\rangle = n \geq 0 \tag{B.7}$$

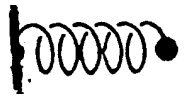


Figure B.1: Simple harmonic oscillator with single spring.

where $|y\rangle = a|n\rangle$ and the absolute value square of the eigenvector cannot be negative. Hence n is a positive number or zero.

$$Na|n\rangle = a^\dagger aa|n\rangle = (aa^\dagger - 1)a|n\rangle = (n-1)a|n\rangle \tag{B.8}$$

Hence $a|n\rangle = c|n-1\rangle$ and $\langle n|a^\dagger a|n\rangle = |c|^2$. However from Eq. B.7 $\langle n|a^\dagger a|n\rangle = n$ so that $c = \sqrt{n}$ and $a|n\rangle = \sqrt{n}|n-1\rangle$. Since the operator a lowers the quantum number of the state by unity, a is called the annihilation operator. Therefore n also has to be an integer, so that the null state is eventually reached by applying operator a for a sufficient number of times.

$$Na^\dagger|n\rangle = a^\dagger aa^\dagger|n\rangle = a^\dagger(1 + a^\dagger a)|n\rangle = (n+1)a^\dagger|n\rangle \tag{B.9}$$

Hence $a^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle$ so that $a^\dagger$ is called a raising operator or a creation operator. Finally,

$$\mathcal{H}|n\rangle = \hbar\omega[N + 1/2]|n\rangle = \hbar\omega(n + 1/2)|n\rangle \tag{B.10}$$

so the eigenvalues become

$$E = \hbar\omega(n + 1/2), \qquad n = 0, 1, 2, \ldots. \tag{B.11}$$

B.2 Phonons

In this section we relate the lattice vibrations to harmonic oscillators and identify the quanta of the lattice vibrations with phonons. Consider the 1-D model of atoms connected by springs (see Fig. B.1). The Hamiltonian for this case is written as:

$$\mathcal{H} = \sum_{s=1}^{\mathcal{N}} \left(\frac{p_s^2}{2m} + \frac{1}{2}\kappa(x_{s+1} - x_s)^2 \right) \tag{B.12}$$

This equation doesn't look like a set of independent harmonic oscillators since x_s and x_{s+1} are coupled. Let

$$\begin{aligned} x_s &= 1/\sqrt{\mathcal{N}} \textstyle\sum_k Q_k e^{iksa} \\ p_s &= 1/\sqrt{\mathcal{N}} \textstyle\sum_k P_k e^{iksa}. \end{aligned} \tag{B.13}$$

These Q_k, P_k's are called phonon coordinates. It can be verified that the commutation relation for momentum and coordinate implies a commutation relation between P_k and $Q_{k'}$

$$[p_s, x_{s'}] = \frac{\hbar}{i}\delta_{ss'} \Longrightarrow [P_k, Q_{k'}] = \frac{\hbar}{i}\delta_{kk'}. \tag{B.14}$$

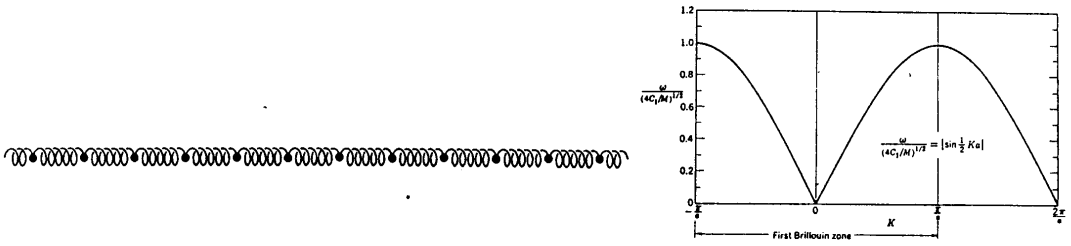


Figure B.2: Schematic for a one dimensional phonon model and the corresponding dispersion relation.

The Hamiltonian in phonon coordinates is:

$$\mathcal{H} = \sum_k \left(\frac{1}{2m} P_k^\dagger P_k + \frac{1}{2} m\omega_k^2 Q_k^\dagger Q_k \right) \tag{B.15}$$

with the dispersion relation given by

$$\omega_k = \sqrt{2\kappa}(1 - \cos ka) \tag{B.16}$$

This is all in Kittel ISSP, pp. 611-613. (see Fig. B.2) Again let

$$a_k = \frac{iP_k^\dagger + m\omega_k Q_k}{\sqrt{2\hbar m\omega_k}}, \tag{B.17}$$

$$a_k^\dagger = \frac{-iP_k + m\omega_k Q_k^\dagger}{\sqrt{2\hbar m\omega_k}} \tag{B.18}$$

so that the Hamiltonian is written as:

$$\mathcal{H} = \sum_k \hbar\omega_k (a_k^\dagger a_k + 1/2) \Rightarrow E = \sum_k (n_k + 1/2)\hbar\omega_k \tag{B.19}$$

The quantum of energy $\hbar\omega_k$ is called a phonon. The state vector of a system of phonons is written as $|n_1, n_2, \ldots, n_k, \ldots\rangle$, upon which the raising and lowering operator can act:

$$a_k|n_1, n_2, \ldots, n_k, \ldots\rangle = \sqrt{n_k}\, |n_1, n_2, \ldots, n_k - 1, \ldots\rangle \tag{B.20}$$

$$a_k^\dagger|n_1, n_2, \ldots, n_k, \ldots\rangle = \sqrt{n_k + 1}\, |n_1, n_2, \ldots, n_k + 1, \ldots\rangle \tag{B.21}$$

From Eq. B.21 it follows that the probability of annihilating a phonon of mode k is the absolute value squared of the diagonal matrix element or n_k.

B.3 Electron-Phonon Interaction

The basic Hamiltonian for the electron-lattice system is

$$\mathcal{H} = \sum_k \frac{p_k^2}{2m} + \frac{1}{2}{\sum_{kk'}}' \frac{e^2}{|\vec{r}_k - \vec{r}_{k'}|} + \sum_i \frac{P_i^2}{2M} + \frac{1}{2}{\sum_{ii'}}' V_{\text{ion}}(\vec{R}_i - \vec{R}_{i'}) + \sum_{k,i} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i) \tag{B.22}$$

where the first two terms constitute $\mathcal{H}_{\rm electron}$, the third and fourth terms are denoted by $\mathcal{H}_{\rm ion}$ and the last term is $\mathcal{H}_{\rm electron-ion}$. The electron-ion interaction term can be separated into two parts: the interaction of electrons with ions in their equilibrium positions, and an additional term due to lattice vibrations:

$$\begin{aligned} \mathcal{H}_{\rm el-ion} &= \mathcal{H}^0_{\rm el-ion} + \mathcal{H}_{\rm el-ph} && \text{(B.23)} \\ \sum_{k,i} V_{\rm el-ion}(\vec{r}_k - \vec{R}_i) &= \sum_{k,i} V_{\rm el-ion}[\vec{r}_k - (\vec{R}^0_i + \vec{s}_i)] && \text{(B.24)} \end{aligned}$$

where $\vec{R}^0_i$ is the equilibrium lattice site position and $\vec{s}_i$ is the displacement of the atoms from their equilibrium positions in a lattice vibration so that

$$\mathcal{H}^0_{\rm el-ion} = \sum_{k,i} V_{\rm el-ion}(\vec{r}_k - \vec{R}^0_i) \tag{B.25}$$

and

$$\mathcal{H}_{\rm el-ph} = -\sum_{k,i} \vec{s}_i \cdot \vec{\nabla} V_{\rm el-ion}(\vec{r}_k - \vec{R}^0_i). \tag{B.26}$$

In solving the Hamiltonian we use an adiabatic approximation, which solves the electronic part of the Hamiltonian by

$$(\mathcal{H}_{\rm electron} + \mathcal{H}^0_{\rm el-ion})\psi = E_{\rm el}\psi \tag{B.27}$$

and seeks a solution of the total problem as

$$\Psi = \psi(\vec{r}_1, \vec{r}_2, \cdots \vec{R}_1, \vec{R}_2, \cdots)\varphi(\vec{R}_1, \vec{R}_2, \cdots) \tag{B.28}$$

such that $\mathcal{H}\Psi = E\Psi$. Here Ψ is the wave function for the electron-lattice system. Plugging this into the Eq. B.22, we find

$$E\Psi = \mathcal{H}\Psi = \psi(\mathcal{H}_{\rm ion} + E_{\rm el})\varphi - \sum_i \frac{\hbar^2}{2M_i}\left(\varphi\nabla^2_i\psi + 2\vec{\nabla}_i\varphi \cdot \vec{\nabla}_i\psi\right) \tag{B.29}$$

Neglecting the last term, which is small, we have

$$\mathcal{H}_{\rm ion}\varphi = (E - E_{\rm el})\varphi \tag{B.30}$$

Hence we have decoupled the electron-lattice system.

$$(\mathcal{H}_{\rm electron} + \mathcal{H}^0_{\rm el-ion})\psi = E_{\rm el}\psi \tag{B.31}$$

which gives us the energy band structure and ψ satisfies Bloch's theorem while φ is the wave function for the ions

$$\mathcal{H}_{\rm ion}\varphi = E_{\rm ion}\varphi \tag{B.32}$$

which gives us phonon spectra and harmonic oscillator like wave functions, as we have already seen in §B.2.

The discussion has thus far left out the electron-phonon interaction $\mathcal{H}_{\rm el-ph}$

$$\mathcal{H}_{\rm el-ph} = -\sum_{k,i} \vec{s}_i \cdot \vec{\nabla} V_{\rm el-ion}(\vec{r}_k - \vec{R}^0_i) \tag{B.33}$$

which is then treated as a perturbation. Since the displacement vector can be written in terms of the normal coordinates $Q_{\vec{q},j}$

$$\vec{s}_i = \frac{1}{\sqrt{\mathcal{N}M}} \sum_{\vec{q},j} Q_{\vec{q},j} e^{i\vec{q}\cdot\vec{R}_i^0}\ \hat{e}_j \tag{B.34}$$

where j denotes the polarization index, $\mathcal{N}$ is the total number of ions and M is the ion mass. Hence

$$\mathcal{H}_{\text{el-ph}} = -\sum_{k,i} \frac{1}{\sqrt{\mathcal{N}M}} \sum_{\vec{q},j} Q_{\vec{q},j} e^{i\vec{q}\cdot\vec{R}_i^0}\ \hat{e}_j \cdot \vec{\nabla} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i^0) \tag{B.35}$$

where the normal coordinate can be expressed in terms of the lowering and raising operators

$$Q_{\vec{q},j} = \left(\frac{\hbar}{2\omega_{\vec{q},j}}\right)^{\frac{1}{2}} (a_{\vec{q},j} + a^{\dagger}_{-\vec{q},j}). \tag{B.36}$$

Writing out the time dependence explicitly,

$$a_{\vec{q},j}(t) = a_{\vec{q},j} e^{-i\omega_{\vec{q},j}t} \tag{B.37}$$

$$a^{\dagger}_{\vec{q},j}(t) = a^{\dagger}_{\vec{q},j} e^{i\omega_{\vec{q},j}t} \tag{B.38}$$

we obtain

$$\begin{aligned} \mathcal{H}_{\text{el-ph}} &= -\sum_{\vec{q},j} \left(\frac{\hbar}{2\mathcal{N}M\omega_{\vec{q},j}}\right)^{\frac{1}{2}} (a_{\vec{q},j} e^{-i\omega_{\vec{q},j}t} + a^{\dagger}_{\vec{q},j} e^{i\omega_{\vec{q},j}t}) \\ &\times \sum_{k,i} (e^{i\vec{q}\cdot\vec{R}_i^0} + e^{-i\vec{q}\cdot\vec{R}_i^0}) \hat{e}_j \vec{\nabla} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i^0) \qquad \text{(B.39)} \\ &= -\sum_{\vec{q},j} \left(\frac{\hbar}{2\mathcal{N}M\omega_{\vec{q},j}}\right)^{\frac{1}{2}} \left(a_{\vec{q},j} \sum_{k,i} \hat{e}_j e^{i(\vec{q}\vec{R}_i^0 - \omega_{\vec{q},j}t)} \cdot \vec{\nabla} V_{\text{el-ion}}(\vec{r}_k - \vec{R}_i^0) \right. \\ &+ \left. c.c. \right) \qquad \text{(B.40)} \end{aligned}$$

If we are only interested in the interaction between one electron and a phonon on a particular branch, say the longitudinal acoustic (LA) branch, then we drop the summation over j and k

$$\mathcal{H}_{\text{el-ph}} = -\left(\frac{\hbar}{2\mathcal{N}M\omega_{\vec{q}}}\right)^{\frac{1}{2}} \left(a_{\vec{q}} \sum_i \hat{e} e^{i(\vec{q}\cdot\vec{R}_i^0 - \omega_{\vec{q}}t)} \cdot \vec{\nabla} V_{\text{el-ion}}(\vec{r} - \vec{R}_i^0) + c.c \right) \tag{B.41}$$

where the first term in the bracket corresponds to the phonon absorption and the c.c. term corresponds to the phonon emission.

With $\mathcal{H}_{\text{el-ph}}$ at hand, we can solve transport problems (e.g., τ due to phonon scattering) and optical problems (e.g., indirect transitions) exactly since all of these problems involve the matrix element $\langle f|\mathcal{H}_{\text{el-ph}}|i\rangle$ of $\mathcal{H}_{\text{el-ph}}$ linking states $|i\rangle$ and $|j\rangle$.

Appendix C

Artificial Atoms

PHYSICS TODAY JANUARY 1993
Marc A. Kastner
Marc Kastner is the Donner Professor of Science in the department of physics at the Massachusetts Institute of Technology, in Cambridge.

The charge and energy of a sufficiently small particle of metal or semiconductor are quantized just like those of an atom. The current through such a quantum dot or one-electron transistor reveals atom-like features in a spectacular way.

The wizardry of modern semiconductor technology makes it possible to fabricate particles of metal or "pools" of electrons in a semiconductor that are only a few hundred angstroms in size. Electrons in these structures can display astounding behavior. Such structures, coupled to electrical leads through tunnel junctions, have been given various names: single electron transistors, quantum dots, zero-dimensional electron gases and Coulomb islands. In my own mind, however, I regard all of these as artificial atoms-atoms whose effective nuclear charge is controlled by metallic electrodes. Like natural atoms, these small electronic systems contain a discrete number of electrons and have a discrete spectrum of energy levels. Artificial atoms, however, have a unique and spectacular property: The current through such an atom or the capacitance between its leads can vary by many orders of magnitude when its charge is changed by a single electron. Why this is so, and how we can use this property to measure the level spectrum of an artificial atom, is the subject of this article.

To understand artificial atoms it is helpful to know how to make them. One way to confine electrons in a small region is by employing material boundaries by surrounding a metal particle with insulator, for example. Alternatively, one can use electric fields to confine electrons to a small region within a semiconductor. Either method requires fabricating very small structures. This is accomplished by the techniques of electron and x-ray lithography. Instead of explaining in detail how artificial atoms are actually fabricated, I will describe the various types of atoms schematically.

Figures C.1a and C.1b show two kinds of what is sometimes called, for reasons that will soon become clear, a single-electron transistor. In the first type (figure C.1a), which I call the all-metal artificial atom,[1] electrons are confined to a metal particle with typical dimensions of a few thousand angstroms or less. The particle is separated from the leads by thin insulators, through which electrons must tunnel to get from one side to the other. The leads are labeled "source" and "drain" because the electrons enter through the former

and leave through the latter the same way the leads are labeled for conventional field effect transistors, such as those in the memory of your personal computer. The entire structure sits near a large, well-insulated metal electrode, called the gate.

Figure C.1b shows a structures[2] that is conceptually similar to the all-metal atom but in which the confinement is accomplished with electric fields in gallium arsenide. Like the all-metal atom, it has a metal gate on the bottom with an insulator above it; in this type of atom the insulator is AlGaAs. When a positive voltage V_g is applied to the gate, electrons accumulate in the layer of GaAs above the AlGaAs. Because of the strong electric field at the AlGaAs-GaAs interface, the electrons' energy for motion perpendicular to the interface is quantized, and at low temperatures the electrons move only in the two dimensions parallel to the -interface. The special feature that makes this an artificial atom is the pair of electrodes on the top surface of the GaAs. When a negative voltage is applied between these and the source or drain, the electrons are repelled and cannot accumulate underneath them. Consequently the electrons are confined in a narrow channel between the two electrodes. Constrictions sticking but into the channel repel the electrons and create potential barriers at either end of the channel. A plot of a potential similar to the one seen by the electrons is shown in the inset in figure C.1. For an electron to travel from the source to the drain it must tunnel through the barriers. The "pool" of electrons that accumulates between the two constrictions plays the same role that the small particle plays in the all-metal atom, and the potential barriers from the constrictions play the role of the thin insulators. Because one can control the height of these barriers by varying the voltage on the electrodes, I call this type of artificial atom the controlled-barrier atom. Controlled- barrier atoms in which the heights of the two potential barriers can be varied independently have also been fabricated.[2] (The constrictions in these devices are similar to those used for measurements of quantized conductance in narrow channels as reported in PHYSICS TODAY, November 1988, page 21.) In addition, there are structures that behave like controlled-barrier atoms but in which the barriers are caused by charged impurities or grain boundaries.[2,4]

Figure C.1c shows another, much simpler type of artificial atom. The electrons in a layer of GaAs are sandwiched between two layers of insulating AlGaAs. One or both of these insulators acts as a tunnel barrier. If both barriers are thin, electrons can tunnel through them, and the structure is analogous to the single-electron transistor without the gate. Such structures, usually called quantum dots, have been studied extensively.[5,6] To create the structure, one starts with two-dimensional layers like those in figure C.1b. The cylinder can be made by etching away unwanted regions of the layer structure, or a metal electrode on the surface, like those in figure C.1b, can be used to repel electrons everywhere except in a small circular section of GaAs. Although a gate electrode can be added to this kind of structure, most of the experiments have bee done without one, so I call this the two-probe atom.

C.1 Charge quantization

One way to learn about natural atoms is to measure the energy required to add or remove electrons. This usually done by photoelectron spectroscopy. For example the minimum photon energy needed to remove an electron is the ionization potential, and the maximum energy (photons emitted when an atom captures an electron is the electron affinity. To learn about artificial atoms we also measure the energy needed to add or subtract electron.

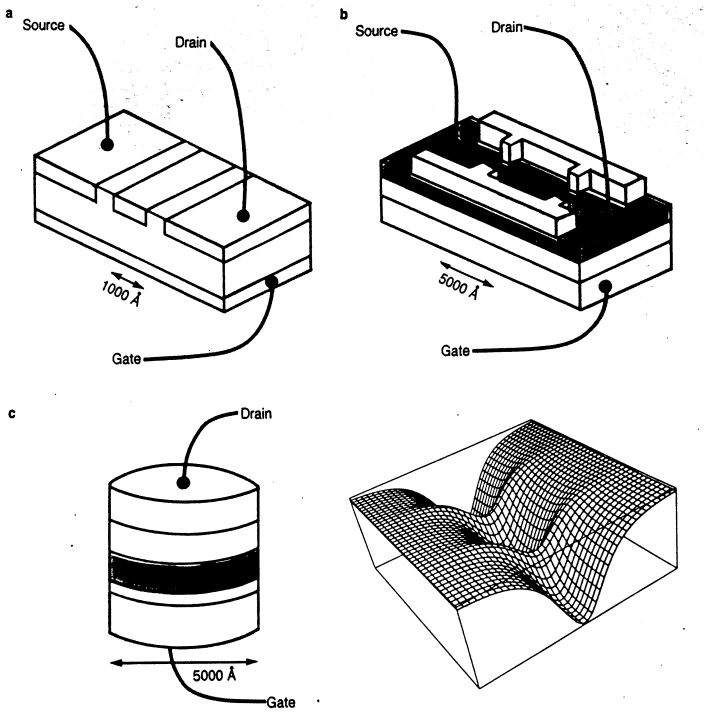


Figure C.1: The many forms of artificial atoms include the all-metal atom (a), the controlled-barrier atom (b) and the two-probe atom, or "quantum dot" (c). Areas shown in blue are metallic, white areas are insulating, and red areas are semiconducting. The dimensions indicated are approximate. The inset shows a potential similar to the one in the controlled-barrier atom, plotted as a function of position at the semiconductor-insulator interface. The electrons must tunnel through potential barriers caused by the two constrictions. For capacitance measurements with a two-probe atom, only the source barrier is made thin enough for tunneling, but for current measurements both source and drain barriers are thin.

However, we do it by measuring the current through the artificial atom.

Figure C.2 shows the current through a controller barrier atom[7] as a function of the voltage V_g between the gate and the atom. One obtains this plot by applying very small voltage between the source and drain, just large enough to measure the tunneling conductance between them. The results are astounding. The conductance(displays sharp resonances that are almost periodic in V_g By calculating the capacitance between the artificial atom and the gate we can show[2,8] that the period is the voltage necessary to add one electron to the confined pool of electrons. That is why we sometimes call the controller barrier atom a single-electron transistor: Whereas the transistors in your personal computer turn on only on(when many electrons are added to them, the artificial atom turns on and off again every time a single electron added to it.

A simple theory, the Coulomb blockade model, explains the periodic conductance resonances.[9] (See PHYSICS TODAY, May 1988, page 19.) This model is quantitatively correct for the all-metal atom and qualitatively correct for the controlled-barrier atom.[10] To understand the model, think about how an electron in the all-metal atom tunnels from one lead onto the metal particle and then onto the other lead. Suppose the particle is neutral to begin with. To add a charge Q to the particle requires energy $Q^2/2C$, where C is the total capacitance between the particle and the rest of the system; since you cannot add less than one electron the flow of current requires a Coulomb energy $e^2/2C$. This energy barrier is called the Coulomb blockade. A fancier way to say this is that charge quantization leads to an energy gap in the spectrum of states for tunneling: For an electron to tunnel onto the particle, its energy must exceed the Fermi energy of the contact by $e^2/2C$, and for a hole to tunnel, its energy must be below the Fermi energy by the same amount. Consequently the energy gap has width e^2/C. If the temperature is low enough that $kT < e^2/2C$, neither electrons nor holes can flow from one lead to the other.

The gap in the tunneling spectrum is the difference between the "ionization potential" and the "electron affinity" of the artificial atom. For a hydrogen atom the ionization potential is 13.6 eV, but the electron affinity, the binding energy of H^-, is only 0.75 eV. This large difference arises from the strong repulsive interaction between the two electrons bound to the same proton. Just as for natural atoms like hydrogen, the difference between the ionization potential and electron affinity for artificial atoms arises from the electron-electron interactions; the difference, however, is much smaller for artificial atoms because they are much bigger than natural ones.

By changing the gate voltage V_g one can alter the energy required to add charge to the particle. V_g is applied between the gate and the source, but if the drain-source voltage is very small, the source, drain and particle will all be at almost the same potential. With V_g applied, the electrostatic energy of a charge Q on the particle is

$$E = QV_g + Q^2/2C \tag{C.1}$$

For negative charge Q, the first term is the attractive interaction between Q and the positively charged gate electrode, and the second term is the repulsive interaction among the bits of charge on the particle. Equation C.1 shows that the energy as a function of Q is a parabola with its minimum at $Q = -CV_g$. For simplicity I have assumed that the gate is the only electrode that contributes to C; in reality, there are other contributions.[7]

By varying V_g we can choose any value of Q_0, the charge that would minimize the energy in equation C.1 if charge were not quantized. However, because the real charge

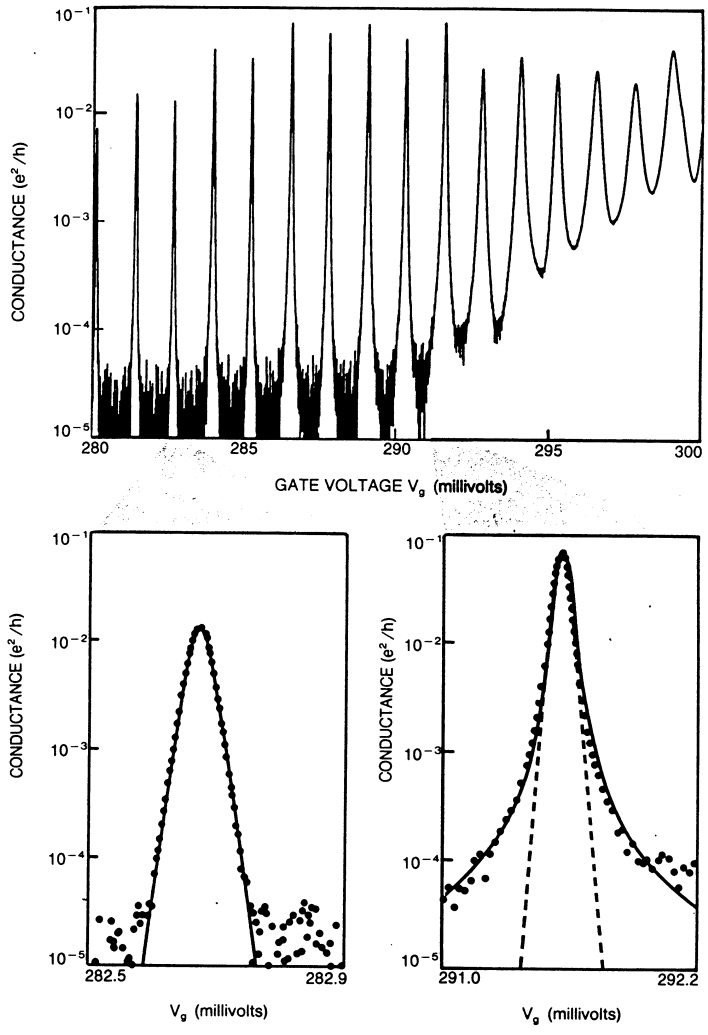


Figure C.2: Conductance of a controlled-barrier atom as a function of the voltage V_g on the gate at a temperature of 60 mK. At low V_g (solid blue curve) the shape of the resonance is given by the thermal distribution of electrons in the source that are tunneling onto the atom, but at high V_g a thermally broadened Lorentzian (red curve) is a better description than the thermal distribution alone (dashed blue curve). (Adapted from ref. 7.)

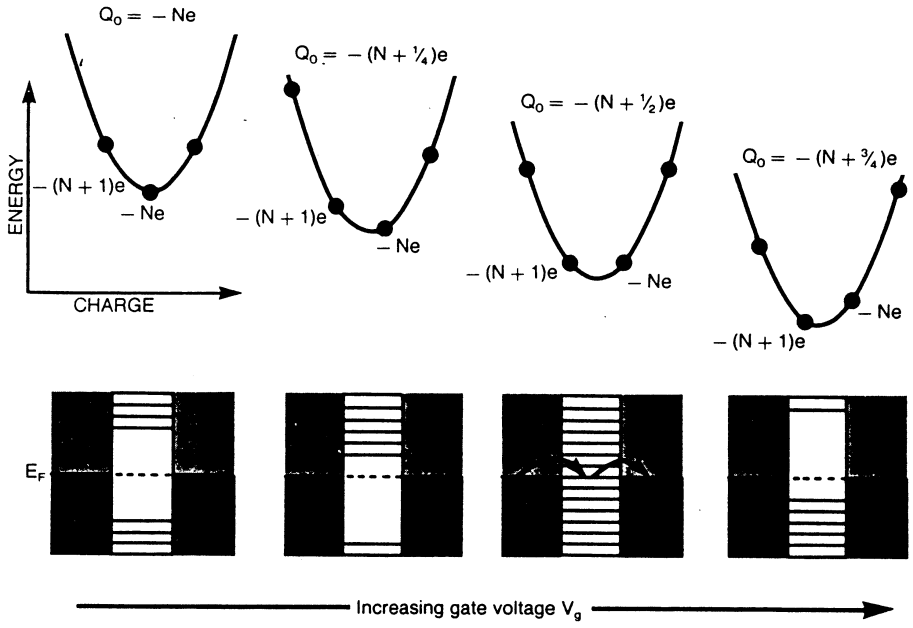


Figure C.3: Total energy (top) and tunneling energies (bottom) for an artificial atom. As voltage is increased the charge Q_0 for which the energy is minimized changes from $-Ne$ to $-(N + 1/4)e$. Only the points corresponding to discrete numbers of electrons on the atom are allowed (dots on upper curves). Lines in the lower diagram indicate energies needed for electrons or holes to tunnel onto the atom. When $Q_0 = -(N + 1/2)e$ the gap in tunneling energies vanishes and current can flow.

is quantized, only discrete values of the energy E are possible. (See figure C.3.) When $Q_0 = -Ne$, an integral number N of electrons minimizes E, and the Coulomb interaction results in the same energy difference $e^2/2C$ for increasing or decreasing N by 1. For all other values of Q_0 except $Q_0 = -(N + 1/2)e$ there is a smaller, but nonzero, energy for either adding or subtracting an electron. Under such circumstances no current can flow at low temperature. However, if $Q_0 = -(N + 1/2)e$ the state with $Q = -Ne$ and that with $Q = (N + 1)e$ are degenerate, and the charge fluctuates between the two values even at zero temperature. Consequently the energy gap in the tunneling spectrum disappears, and current can flow. The peaks in conductance are therefore periodic, occurring whenever $CV_g = Q_0 = -(N + 1/2)e$, spaced in gate voltage by e/C.

As shown in figure C.3, there is a gap in the tunneling spectrum for all values of V_g except the charge-degeneracy points. The more closely spaced discrete levels shown outside this gap are due to excited states of the electrons present on the artificial atom and will be discussed more in the next section. As V_g is increased continuously, the gap is pulled down relative to the Fermi energy until a charge degeneracy point is reached. On moving through this point there is a discontinuous change in the tunneling spectrum: The gap collapses and then reappears shifted up by e^2/C. Simultaneously the charge on the artificial atom increases by 1 and the process starts over again. A charge-degeneracy point and a conductance peak are reached every time the voltage is increased by e/C, the amount necessary to add one electron to the artificial atom. Increasing the gate voltage of an artificial atom is therefore analogous to moving through the periodic table for natural atoms by increasing the nuclear charge.

The quantization of charge on a natural atom is something we take for granted. However, if atoms were larger, the energy needed to add or remove electrons would be smaller, and the number of electrons on them would fluctuate except at very low temperature. The quantization of charge is just one of the properties that artificial atoms have in common with natural ones.

C.2 Energy quantization

The Coulomb blockade model accounts for charge quantization but ignores the quantization of energy resulting from the small size of the artificial atom. This confinement of the electrons makes the energy spacing of levels in the atom relatively large at low energies. If one thinks of the atom as a box, at the lowest energies the level spacings are of the order $\hbar^2/ma^2$, where a is the size of the box. At higher energies the spacings decrease for a three-dimensional atom because of the large number of standing electron waves possible for a given energy. If there are many electrons in the atom, they fill up many levels, and the level spacing at the Fermi energy becomes small. The all-metal atom has so many electrons (about 10^7) that the level spectrum is effectively continuous. Because of this, many experts do not regard such devices as "atoms," but I think it is helpful to think of them as being atoms in the limit in which the number of electrons is large. In the controlled-barrier atom, however, there are only about 30–60 electrons, similar to the number in natural atoms like krypton through xenon. Two-probe atoms sometimes have only one or two electrons. (There are actually many more electrons that are tightly bound to the ion cores of the semiconductor, but those are unimportant because they cannot move.) For most cases, therefore, the spectrum of energies for adding an extra electron to the atom is discrete, just

as it is for natural atoms. That is why a discrete set of levels is shown in figure C.3.

One can measure the energy level spectrum directly by observing the tunneling current at fixed V_g as a function of the voltage $V_{\rm ds}$ between drain and source. Suppose we adjust V_g so that, for example, $Q_0 = -(N + 1/4)e$ and then begin to increase $V_{\rm ds}$. The Fermi level in the source rises in proportion to $V_{\rm ds}$ relative to the drain, so it also rises relative to the energy levels of the artificial atom. (See the inset to figure C.4a.) Current begins to flow when the Fermi energy of the source is raised just above the first quantized energy level of the atom. As the Fermi energy is raised further, higher energy levels in the atom fall below it, and more current flows because there are additional channels for electrons to use for tunneling onto the artificial atom. We measure an energy level by measuring the voltage at which the current increases or, equivalently, the voltage at which there is a peak in the derivative of the current, $dI/dV_{\rm ds}$. (We need to correct for the increase in the energy of the atom with $V_{\rm ds}$, but this is a small effect.) Many beautiful tunneling spectra of this kind have been measured[5] for two-terminal atoms. Figure C.4a shows one for a controlled barrier atom.[7]

Increasing the gate voltage lowers all the energy levels in the atom by eV_g, so that the entire tunneling spectrum shifts with V_g, as sketched in figure C.3. One can observe this effect by plotting the values of $V_{\rm ds}$ at which peaks appear in $dI/dV_{\rm ds}$. (See figure C.4b.) As V_g increases you can see the gap in the tunneling spectrum shift lower and then disappear at the charge-degeneracy point, just as the Coulomb blockade model predicts. You can also see the discrete energy levels of the artificial atom. For the range of V_s shown in figure C.4 the voltage is only large enough to add or remove one electron from the atom; the discrete levels above the gap are the excited states of the atom with one extra electron, and those below the gap are the excited states of the atom with one electron missing (one hole). At still higher voltages (not shown in figure C.4) one observes levels for two extra electrons or holes and so forth. The charge-degeneracy points are the values of V_g for which one of the energy levels of the artificial atom is degenerate with the Fermi energy in the leads when $V_{\rm ds} = 0$, because only then can the charge of the atom fluctuate.

In a natural atom one has little control over the spectrum of energies for adding or removing electrons. There the electrons interact with the fixed potential of the nucleus and with each other, and these two kinds of interaction determine the spectrum. In an artificial atom, however, one can change this spectrum completely by altering the atom's geometry and composition. For the all-metal atom, which has a high density of electrons, the energy spacing between the discrete levels is so small that it can be ignored. The high density of electrons also results in a short screening length for external electric fields, so electrons added to the atom reside on its surface. Because of this, the electron-electron interaction is always e^2/C (where C is the classical geometrical capacitance), independent of the number of electrons added. This is exactly the case for which the Coulomb blockade model was invented, and it works well: The conductance peaks are perfectly periodic in the gate voltage. The difference between the "ionization potential" and the "electron affinity" is e^2/C, independent of the number of electrons on the atom.

In the controlled-barrier atom, as you can see from figure C.4, the level spacing is one or two tenths of the energy gap. The conductance peaks are not perfectly periodic in gate voltage, and the difference between ionization potential and electron affinity has a quantum mechanical contribution. I will discuss this contribution a little later in more detail.

In the two-probe atom the electron-electron interaction can be made very small, so that

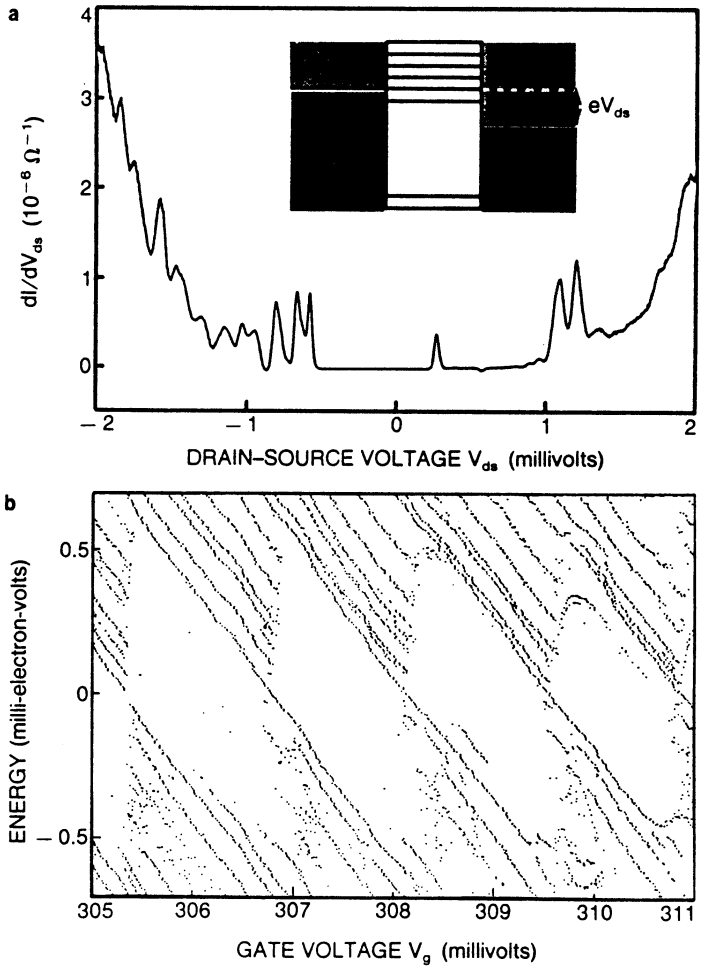


Figure C.4: Discrete energy levels of an artificial atom can be detected by varying the drain-source voltage. When a large enough $V_{\rm ds}$ is applied, electrons overcome the energy gap and tunnel from the source to the artificial atom. (See inset of a.) a: Every time a new discrete state is accessible the tunneling current increases, giving a peak in $dI/dV_{\rm ds}$. The Coulomb blockade gap is the region between about –0.5 mV and + 0.3 mV where there are no peaks. b: Plotting the positions of these peaks at various gate voltages gives the level spectrum. Note how the levels and the gap move downward as V_g increases, just as sketched in the lower part of figure C.3. (Adapted from ref. 7.)

one can in principle reach the limit opposite to that of the all-metal atom. One can find the energy levels of a two-probe atom by measuring the capacitance between its two leads as a function of the voltage between them.[6] When no tunneling occurs, this capacitance is the series combination of the source-atom and atom-drain capacitances. For capacitance measurements, two-probe atoms are made with the insulating layer between the drain and atom so thick that current cannot flow under any circumstances. Whenever the Fermi level in the source lines up with one of the energy levels of the atom, however, electrons can tunnel freely back and forth between the atom and the source. This causes the total capacitance to increase, because the source-atom capacitor is effectively shorted by the tunneling current. The amazing thing about this experiment is that a peak occurs in the capacitance every time a single electron is added to the atom. (See figure C.5a.) The voltages at which the peaks occur give the energies for adding electrons to the atom, just as the voltages for peaks in $dI/dV_{\rm ds}$ do for the controlled-barrier atom or for a two probe atom in which both the source-atom barrier and the atom-drain barrier are thin enough for tunneling. The first peak in figure C.5a corresponds to the one-electron artificial atom.

Figure C.5b shows how the energies for adding electrons to a two-probe atom vary with a magnetic field perpendicular to the GaAs layer. In an all-metal atom the levels would be equally spaced, by e^2/C, and would be independent of magnetic field because the electron-electron interaction completely determines the energy. By contrast, the levels of the two-probe atom are irregularly spaced and depend on the magnetic field in a systematic way. For the two-probe atom the fixed potential determines the energies at zero field. The level spacings are irregular because the potential is not highly symmetric and varies at random inside the atom because of charged impurities in the GaAs and AlGaAs. It is clear that the electron-electron interactions that are the source of the Coulomb blockade are not always so important in the two-probe atom as in the all-metal and controlled-barrier atoms. Their relative importance depends in detail on the geometry.[5]

C.3 Artificial atoms in a magnetic field

Level spectra for natural atoms can be calculated theoretically with great accuracy, and it would be nice to be able to do the same for artificial atoms. No one has yet calculated. an entire spectrum, like that in figure C.4a. However, for a simple geometry we can now predict the charge-degeneracy points, the values of V_g corresponding to conductance peaks like those in figure C.2. From the earlier discussion it should be clear that in such a calculation one must take into account the electron's interactions with both the fixed potential and the other electrons.

The simplest way to do this is with an extension of the Coulomb blockade model.[11–13] It is assumed, as before, that the contribution to the gap in the tunneling spectrum from the Coulomb interaction is e^2/C no matter how many electrons are added to the atom. To account for the discrete levels one pretends that once on the atom, each electron interacts independently with the fixed potential. All one has to do is solve for the energy levels of a single electron in the fixed potential that creates the artificial atom and then fill those levels in accordance with the Pauli exclusion principle. Because the electron-electron interaction is assumed always to be e^2/C, this is called the constant-interaction model.

Now think about what happens when one adds electrons to a controlled-barrier atom by increasing the gate voltage while keeping $V_{\rm ds}$ just large enough so one can measure the

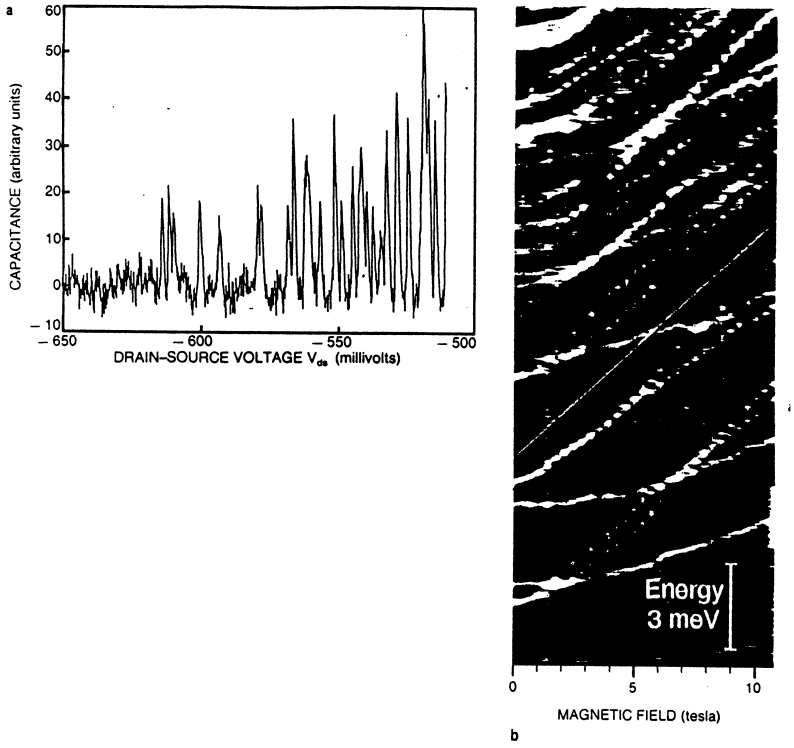


Figure C.5: Capacitance of a two-probe atom that has only one barrier thin enough to allow tunneling. a: The capacitance has a peak every time a single electron is added to the atom. The positions of the peaks give the energy spectrum of the atom. b: Peaks in capacitance plotted versus applied magnetic field. The green line indicates the rate of change of the energy expected when the magnetic field dominates. (Adapted from ref. 6.)

conductance. When there are $N-1$ electrons on the atom the $N-1$ lowest energy levels are filled. The next conductance peak occurs when the gate voltage pulls the energy of the atom down enough that the Fermi level in the source and drain becomes degenerate with the Nth level. Only when an energy level is degenerate with the Fermi energy can current flow; this is the condition for a conductance peak. When V_g is increased further and the next conductance peak is reached, there are N electrons on the atom, and the Fermi level is degenerate with the $(N+1)$-th level. Therefore to get from one peak to the next the Fermi energy must be raised by $e^2/C + (E_{N+1} - E_N)$, where E_N, is the energy of the Nth level of the atom. If the energy levels are closely spaced the Coulomb blockade result is recovered, but in general the level spacing contributes to the energy between successive conductance peaks.

It turns out that we can test the results of this kind of calculation best if a magnetic field is applied perpendicular to the GaAs layer. For free electrons in two dimensions, applying the magnetic field results in the spectrum of Landau levels with energies $(n+1/2)\hbar\omega_c$ where the cyclotron frequency is $\omega_c = eB/m^*c$, and m^* is the effective mass of the electrons. In the controlled barrier atom and the two-probe atom, we expect levels that behave like Landau levels at high fields, with energies that increase linearly in B. This behavior occurs because when the field is large enough the cyclotron radius is much smaller than the size of the electrostatic potential well that confines the electrons, and the electrons act as if they were free. Levels shifting proportionally to B, as expected, are seen experimentally. (See figure C.5b.)

To calculate the level spectrum we need to model the fixed potential, the analog of the potential from the nucleus of a natural atom. The simplest choice is a harmonic oscillator potential, and this turns out to be a good approximation for the controlled-barrier atom. Figure C.6a shows the calculated level spectrum as a function of magnetic field for non-interacting electrons in a two dimensional harmonic oscillator potential. At low fields the energy levels dance around wildly with magnetic field. This occurs because some states have large angular momentum and the resulting magnetic moment causes their energies to shift up or down strongly with magnetic field. As the field is increased, however, things settle down. For most of the field range shown there are four families of levels, two moving up, the other two down. At the highest fields there are only two families, corresponding to the two possible spin states of the electron.

Suppose we measure, in an experiment like the one whose results are shown in figure C.2, the gate voltage at which a specific peak occurs as a function of magnetic field. This value of V_g is the voltage at which the Nth energy level is degenerate with the Fermi energy in the source and drain. A shift in the energy of the level will cause a shift in the peak position. The blue line in figure C.6a is the calculated energy of the 39th level (chosen fairly arbitrarily for illustration purposes), so it gives the prediction of the constant-interaction model for the position of the 39th conductance peak. As the magnetic field increases, levels moving up in energy cross those moving down, but the number of electrons is fixed, so electrons jump from upward-moving filled levels to downward-moving empty ones. The peak always follows the 39th level, so it moves up and down in gate voltage.

Figure C.6b shows a measurement[14] of V_g for one conductance maximum, like one of those in figure C.2, as a function of B. The behavior is qualitatively similar to that predicted by the constant-interaction model: The peak mover up and down with increasing B, and the frequency of level crossings changes at the field where only the last two families of levels

remain. However, at high B the frequency is predicted to be much lower than what is observed experimentally. While the constant-interaction model is in qualitative agreement with experiment, it is not quantitatively correct.

To anyone who has studied atomic physics, the constant-interaction model seems quite crude. Even the simplest models used to calculate energies of many electron atoms determine the charge density and potential self- consistently. One begins by calculating the charge density that would result from noninteracting electrons in the fixed potential, and then one calculates the effective potential an electron sees because of the fixed potential and the potential resulting from this charge density. Then one calculates the charge density again. One does this repeatedly until the charge density and potential are self-consistent. The constant-interaction model fails because it is not self- consistent. Figure C.6c shows the results of a self-consistent calculation for the controlled-barrier atom.[14] It is in good agreement with experiment-much better agreement than the constant-interaction model gives.

C.4 Conductance line shapes

In atomic physics, the next step after predicting energy levels is to explore how an atom interacts with the electromagnetic field, because the absorption and emission of photons teaches us the most about atoms. For artificial atoms, absorption and emission of electrons plays this role, so we had better understand how this process works. Think about what happens when the gate voltage in the controlled-barrier atom is set at a conductance peak, and an electron is tunneling back and forth between the atom and the leads. Since the electron spends only a finite time τ on the atom, the uncertainty principle tells us that the energy level of the electron has a width $\hbar/\tau$. Furthermore, since the probability of finding the electron on the atom decays as $e^{t/\tau}$, the level will have a Lorentzian line shape.

This line shape can be measured from the transmission probability spectrum $T(E)$ of electrons with energy E incident on the artificial atom from the source. The spectrum is given by

$$T(E) = \frac{\Gamma^2}{\Gamma^2 + (E - E_N)^2} \tag{C.2}$$

where Γ is approximately $\hbar/\tau$ and E_N is the energy of the Nth level. The probability that electrons are transmitted from the source to the drain is approximately,proportional[15] to the conductance G. In fact, $G \simeq (e^2/h)T$, where e^2/h is the quantum of conductance. It is easy to show that one must have $G < e^2/h$ for each of the barriers separately to observe conductance resonances. (An equivalent argument is used to show that electrons in a disordered conductor are localized for $G < e^2/h$. See, for example, the article by Boris L. Al'tshuler and Patrick A. Lee in PHYSICS TODAY, December 1988, page 36.) This condition is equivalent to requiring that the separation of the levels is greater than their width Γ.

Like any spectroscopy, our electron spectroscopy of artificial atoms has a finite resolution. The resolution is determined by the energy spread of the electrons in the source, which are trying to tunnel into the artificial atom. These electrons are distributed according to the Fermi-Dirac function,

$$f(E) = \frac{1}{\exp[(E - E_F)/kT] + 1} \tag{C.3}$$

where E_F is the Fermi energy. The tunneling current is given by

$$I = \int \frac{e}{h} T(E)[f(E) - f(E - eV_{\rm ds})]dE. \tag{C.4}$$

Equation C.4 says that the net current is proportional to the probability $f(E)T(E)$ that there is an electron in the source with energy E and that the electron can tunnel between the source and drain minus the equivalent probability for electrons going from drain to source. The best resolution is achieved by making $V_{\rm ds} \ll kT$. Then $[f(E) - f(E - eV_{\rm ds})] \simeq eV_{\rm ds}(df/dE)$, and I is proportional to $V_{\rm ds}$, so the conductance is $I/V_{\rm ds}$.

Figure C.2 shows that equations C.2–C.4 describe the experiments well: At low V_g, where Γ is much less than kT, the shape of the conductance resonance is given by the resolution function df/dE. But at higher V_g one sees the Lorentzian tails of the natural line shape quite clearly. The width Γ depends exponentially on the height and width of the potential barrier, as is usual for tunneling. The height of the tunnel barrier decreases with V_g, which is why the peaks become broader with increasing V_g. Just as we have control over the level spacing in artificial atoms, we also can control the coupling to the leads and therefore the level widths. It is clear why the present generation of artificial atoms show unusual behavior only at low temperatures: When kT becomes comparable to the energy separation between resonances, the peaks overlap and the features disappear.

C.5 Applications

The behavior of artificial atoms is so unusual that it is natural to ask whether they will be useful for applications to electronics. Some clever things can be done., Because of the electron-electron interaction, only one electron at a time can pass through the atom. With devices like the "turnstile" device[16,17] shown on the cover of this issue the two tunnel barriers can be raised and lowered independently. Suppose the two barriers are raised and lowered sequentially at a radio or microwave frequency ν. Then, with a small source-drain voltage applied, an electron will tunnel onto the atom when the source-atom barrier is low and off it when the atom-drain barrier is low. One electron will pass in each time interval ν^{-1}, producing a current eV. Other applications, such as sensitive electrometers, can be imagined.[9,18] However, the most interesting applications may involve devices in which several artificial atoms are coupled together to form artificial molecules[16,17,19] or in which many are coupled to form artificial solids. Because the coupling between the artificial atoms can be controlled, new physics as well as new applications may emerge. The age of artificial atoms has only just begun.

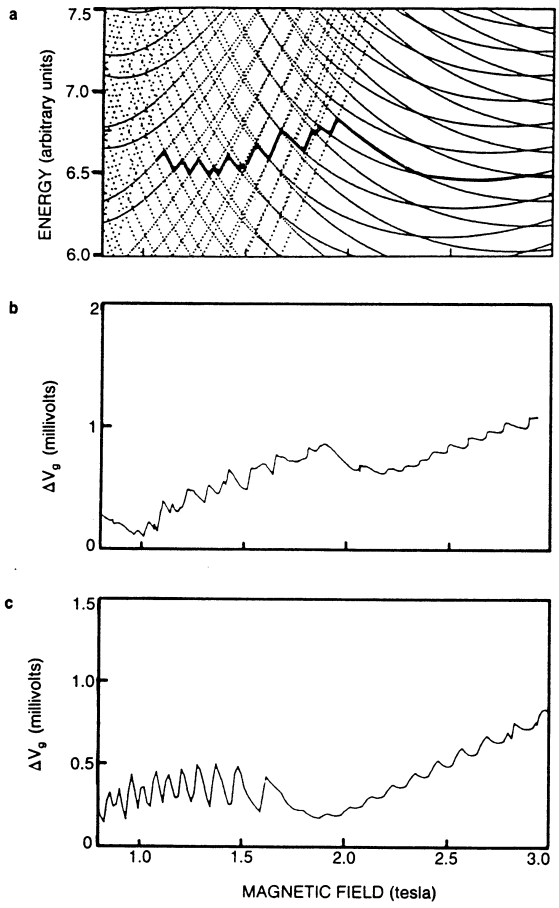


Figure C.6: Effect of magnetic field on energy level spectrum and conductance peaks. a: Calculated level spectrum for noninteracting electrons in a harmonic oscillator electrostatic potential as a function of magnetic field. The blue line is the prediction that the constant interaction model gives for the gate voltage for the 39th conductance peak. b: Measured position of a conductance peak in a controlled-barrier atom as a function of field. c: Position of the 39th conductance peak versus field, calculated self-consistently. The scale in c does not match that in b because parameters in the calculation were not precisely matched to the experimental conditions. (Adapted from Ref. 14.)

References

1. T. A. Fulton, G. J. Dolan, Phys. Rev. Lett. 59, 109 (1987).
2. U. Meirav, M. A. Kastner, S. J. Wind, Phys. Rev. Lett. 65, 771 (1990). M. A. Kastner, Rev. Mod. Phys. 64, 849 (1992).
3. L. P. Kouwenhoven, N. C. van der Vaart, A. T. Johnson, W. Kool, C. J. P. M. Harmans, J. G. Williamson, A. A. M. Staring, C. T. Foxon, Z. Phys. B 85, 367 (1991), and refs. therein.
4. V. Chandrasekhar, Z. Ovadyahu, R. A. Webb, Phys. Rev. Lett. 67, 2862 (1991). R. J. Brown, M. Pepper, H. Ahmed, D. G. Hasko, D. A. Ritchie, J. E. F. Frost, D. C. Peacock, G. A. C. Jones, J. Phys.: Condensed Matter 2, 2105 (1990).
5. B. Su, V. J. Goldman, J. E. Cunningham, Science 255, 313 (1992). M. A. Reed, J. N. Randall, R. J. Aggarwal, R. J. Matyi, T. M. Moore, A. E. Wetsel, Phys. Rev. Lett. 60, 535 (1988). M. Tewordt, L. Martin-Moreno, J. T. Nicholls, M. Pepper, M. J. Kelly, V. J. Law, D. A. Ritchie, J. E. F. Frost, G. A. C. Jones, Phys. Rev. B 45, 14407 (1992).
6. R. C. Ashoori, H. L. Stormer, J. S. Weiner, L. N. Pfeiffer, S. J. Pearton, K. Baldwin, K. W. West, Phys. Rev. Lett. 68, 3088 (1992).
7. E. B. Foxman, P. L. McEuen, U. Meirav, N. S. Wingreen, Y. Meir, P. A. Belk, N. R. Belk, M. A. Kastner, S. J. Wind, "The Effects of Quantum Levels on Transport Through a Coulomb Island," MIT preprint (July 1992). See also A. T. Johnson, L. P. Kouwenhoven, W. de Jong, N.C. van der Vaart, C. J. P. M. Harmans, C. T. Faxon, Phys. Rev. Lett. 69, 1592 (1992).
8. A. Kumar, Surf. Sci. 263, 335 (1992). A. Kumar, S. E. Laux, F. Stern, Appl. Phys. Lett. 54, 1270 (1989).
9. D. V. Averin, K. K. Likharev, in Mesmcopic Phenomena in Solids, B. L. Al'tshuler, P. A. Lee, R. A. Webb, eds., Elsevier, Amsterdam (1991), p. 173.
10. H. van Houton, C. W. J. Beenakker, Phys. Rev. Lett. 63,1893 (1989).
11. D. V. Averin, A. N. Korotkov, Zh. Eksp. Teor. Fiz. 97, 1661 (1990) [Sov. Phys. JETP 70, 937 (1990)].
12. Y. Meir, N. S. Wingreen, P. A. Lee, Phys. Rev. Lett. 66, 3048 (1991).
13. C. J. Beenakker, Phys. Rev. B 44, 1646 (1991).
14 P.L. McEuen, E. B. Foxman, J. Kinaret, U. Meirav, M. A. Kastner, N. S. Wingreen, S. J. Wind, Phys. Rev. B 45, 11419 (1992).
15. R. Landauer, IBM J. Res. Dev. 1, 223 (1957).
16. L. P. Kouwenhoven, A. T. Johnson, N. C. van der Vaart, W. Kool, C. J. P. M. Harmans, C. T. Foxon, Phys. Rev. Lett. 67, 1626 (1991).
17. L. J. Geerligs, V. F. Anderegg, P A. M. Holweg, J. E. Mooij, H. Pothier, D. Esteve, C. Urbina, M. H. Devoret, Phys. Rev. Lett. 64, 2691 (1990).
18. H. Grabert, M. H. Devoret, eds., Single Charge Tunneling, Plenum, New York (1992).
19. R.J. Haug, J.M. Hong, K.Y.Lee, Surf. Sci. 263, 415 (1991).