Physics of Semiconductors (1)

Shingo Katsumoto
Institute for Solid State Physics, University of Tokyo

April 12, 2021

In the first half of this fiscal year (FY2021), I am assigned to the lecture on "Physics of Semiconductors." It's been a long time since I gave a lecture in one semester last time (8 years), and several new themes, which I want to introduce, have appeared. I am not very good at giving lectures like a machine-gun and want to take a course with a comparatively small amount of content. Though, maybe it's not enough for motivated students who want to learn a lot. Therefore, the notes follow the lecture and cover the advanced content that would open the eyes for more expansive fields. I would also like to introduce references for those who want to expand their studies.

Chapter 1 General Properties of Semiconductors

1 What characterizes semiconductors?

Semiconductors refer to a form of solid that is usually classified by electrical conduction. Metal have large densities of states around the Fermi levels (that is, the Fermi surfaces exist) and are good conductors while insulators have their Fermi levels deep inside the wide energy gaps and interrupt electric currents. Semiconductors stand somewhere between them. They usually have comparatively narrow energy band-gaps, low but finite electric conductance at high temperatures, become insulating with lowering the temperautres.

However, such a viewpoint is not always usuful nowadays. For example, a fine insulator with a large bandgap of 5.5 eV at room temperature, such as diamond, is also called a semiconductor, and devices are being made from it while materials with zero bandgap, such as graphene, are also important targets of researches in the semicondutor field. Rather, it seems to me that the property of "structure sensitive" fits better into the recent usage of the term "semiconductors." This is a long-used expression, which indicates the property that the electric conduction is sensitive to the ultra-small amount of impurities. Although, after the appearances of heterojunctions, MOS structures, superlattices, nanostructures, etc., I think the same expression is applicable to the sensitiveness of the transport properties on such real space structures.

In most cases, the object of such structure sensitive transportation is the electric charge, but recently the spin current in which the magnetic moment is transported by spin has also become an important research object. Research on spintronic devices is also active, and there is a possibility that some will eventually become practical.

For the spin current, which is the flow of magnetic momentum, the non-magnetic metals whose spins are canceled by the time-reversal symmetry, are like an empty space. It can be seen as a system similar to a semiconductor, which is a charge neutral space due to the charge cancellation of the nuclei and electrons. In fact, the inside of the metal is almost equipotential under normal experimental conditions, but a spin Hall spin current may exist. In spintronics, these systems are also "structurally sensitive" and look like semiconductors from the eye of semiconductor researchers. However, this is rather a unique view of myself, and usually, semiconductors are defined as those that are structurally sensitive in electric conduction.

2 Crystal Structures

2.1 Lattice

A solid classified into crystal commonly has a spatially periodic structure of **basis**, which is also a certain structure of atoms. We represent such a state of matter as a **lattice**. "Spatially periodic structure" can be

represented as follows. An arbitrary point in a crystal with spatial coorinate $\boldsymbol{r}$ has an infinite number of equivallent points $\boldsymbol{r}'$, which is represented, in the case of three-dimensional lattice, with three constant vectors $\boldsymbol{a}_i$ $(i = 1, 2, 3)$ and three integers l_i $(i = 1, 2, 3)$ as

$$\boldsymbol{r}' = \boldsymbol{r} + \sum_{i=1,2,3} l_i \boldsymbol{a}_i = \boldsymbol{r} + \boldsymbol{R}. \tag{1.1}$$

Then the unit of the period is a certain set of atoms around $\boldsymbol{r}$. We take an arbitrary point in the unit. Such points form the lattice. We call such points as the lattice point.

The basis is the unit of the periodicity in the crystal and should be taken as to have the minimum number of atoms. $\boldsymbol{a}_i$ in eq.(1.1) are called **primitive vectors**, while $\boldsymbol{R}$ is called a **lattice vector**. The parallelpiped with $\boldsymbol{a}_i$ as the edges contains a single basis is called **primitive cell**, with which we can fill up the space without gap. Primitive vectors often can be taken in multiple ways and usually taken as to make the symmetry of the lattice highest. A primitive cell is defined as a polygon which contains single basis and fills up the entire space without gap. Then there are infinite ways to define a primitive cell other than the above mentioned paralledpiped. When a block with multiple primitive cells is taken as the unit of period and the periodic structure has a higher symmetry, the block is more convenient for the unit. We then consider a **unit cell**, which may consist of single or multiple primitive cells.

As mentioned above, a crystal is composed of a unit structure and a lattice. The example of diamond structure, which often appears in group-IV semiconductors, is illustrated in Fig.1.1. In Fig.1.1(a) the atomic positions are indicated by middle-sized spheres, of which colors (black and white) indicate two different atomic sites in the crystal. The basis is composed of a black and a white atoms and a primitive cell can be taken as to contain these two sites. A point in the primitive cell, *e.g.* the position of black atom, can be taken as the lattice point. The consequent lattice is, as shown in (b), **face centered cubic** (fcc). Let $\boldsymbol{e}_{x,y,z}$ be the unit vectors of the Cartesian coodinate system, then the primitive vectors can be taken as

$$\boldsymbol{a}_1 = \frac{a_0}{2}(\boldsymbol{e}_x + \boldsymbol{e}_y), \quad \boldsymbol{a}_2 = \frac{a_0}{2}(\boldsymbol{e}_y + \boldsymbol{e}_z), \quad \boldsymbol{a}_3 = \frac{a_0}{2}(\boldsymbol{e}_z + \boldsymbol{e}_x). \tag{1.2}$$

The primitive cell of the parallelpiped spanned by the vectors in eq.(1.2) is drawn with solid lines in Fig.1.1(a). On the other hand, the cubic drawn in the figure is often taken as a unit cell.

The lattices are classified by seven **crystal system** and additonal lattice point (no point, face-centered, body-centered, base-centered) into 14 species of **Baravais lattice**.

2.2 Bravais lattice

The number of crystal structures is huge, maybe infinite, if we count, *e.g.* differences in molecular structures of organic crystals. On the other hand, the number of independent lattice structures is as small as 14 as shown in Fig.1.2. These 14 lattices are called three dimensional **Bravais lattice**.

The definition of the Brave lattice classification is based on the discussion of spatial symmetry. The spatial symmetry of a manifold is defined by whether the manifold is invariant for the symmetry operations, such as rotation, reflection, translation, etc. For detailed discussion see, *e.g.* Ref.[1, 2]. Here we briefly summarize how we reach the 14 Bravais lattice.

We first classify the lattices with the relative lengths of primitive translatinal vectors a_1, a_2, a_3 and the angles defined by two edges θ_{12}, θ_{23}, θ_{31} (see the right-down inset of Fig.1.2). And for the rotational symmetry

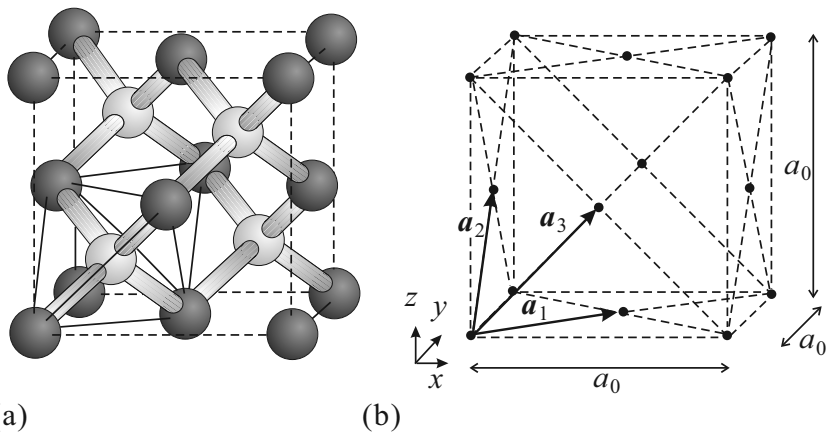


Figure 1.1: (a) Diamond structure. The circles corespond to atomic positions, while the thin cylinders correspond to covalent bonds. There is a single atom species though the two positions identified with colors, are different. The sold lines indicate the primitive cell. (b) Face-centered cubic lattice of the diamond crystal of (a). $\boldsymbol{a}_{1-3}$ are the primitive vectors.

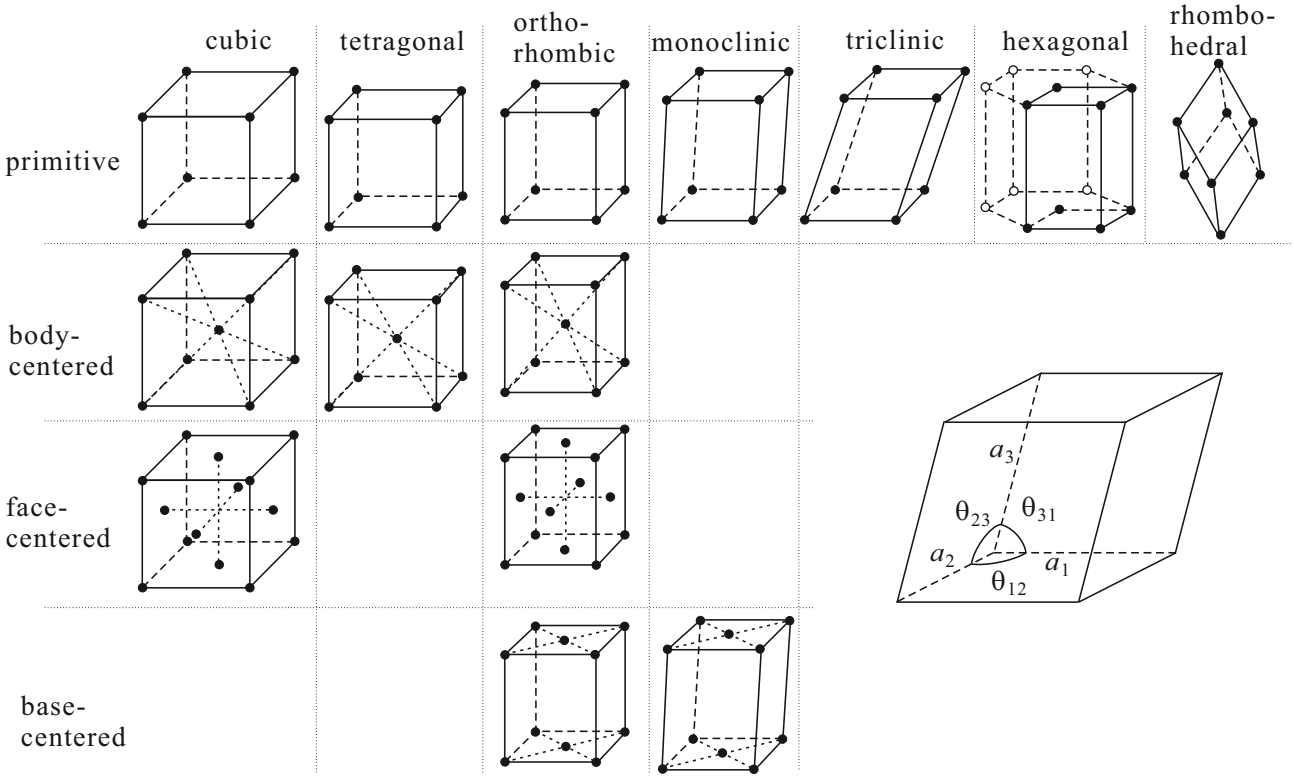


Figure 1.2: Bravais lattices of three-dimension. The parameters for the classification are illustrated in the right-down space.

around a primitive translational vector, the angle $\pi/2$ is a spacial value and whether the angles θ_{ij} are $\pi/2$ or not is the other condition. These conditions leads us to the classification in Tab.1. This classification is called **crystal systems**. In three dimension, possible crystal systems are seven species in in Tab.1 and on the first column in Fig.1.2.

In the classification of crystal system, the focus is on the symmetry in the positions of neighboring lattice points. There are, however, some cases, in which we need to consider the relation between the next neighboring lattice points. For example, we take two (simple) cubic lattices and put them so as to lattice points of one of them are placed to the center of cubic in the other lattice. In this case if we look at the neighboring relation, the cubic symmetry seems to be lost but the second next ones are originally in cubic symmetry and the lattice is still classified into cubic crystal system but contains an additonal lattice point at the center of cubic. The lattice, which contains an additional lattice point at the center of simple cubic, is called **body-centered cubic** (bcc).

In this way, the additional lattice points to the simple crystal system is another condition for the classification. The positions of such additional points are face-centerd, body-centered and base-centered. As a consequence, we get 14 Bravais lattice shown in Fig.1.2.

Because we have ambiguity in taking "lattice", the classification by Bravais lattice also has ambiguities. For a simplest example, in an fcc lattice crystal, if we take the unit as a single face-centered cubic, then the lattice

	θ_{12}	θ_{23}	θ_{31}	$a_1,\ a_2,\ a_3$
cubic	$\pi/2$	$\pi/2$	$\pi/2$	$a_1 = a_2 = a_3$
tetragonal	$\pi/2$	$\pi/2$		$a_1 = a_2 \neq a_3$
orthorhombic	$\pi/2$			$a_1 \neq a_2 \neq a_3$
monoclinic	$\pi/2$	$\pi/2$		
triclinic	$\pi/2$			
hexagonal	$\pi/2$	$2\pi/3$		$a_1 = a_2$
rhombohedral(trigonal)	θ_0	θ_0	$\theta_0 \neq \pi/2$	$a_1 = a_2 = a_3$

Table 1: Conditions for Bravais lattice classification in three dimension. The difinitions of the parameters are shown in the right-down panel in Fig.1.2.

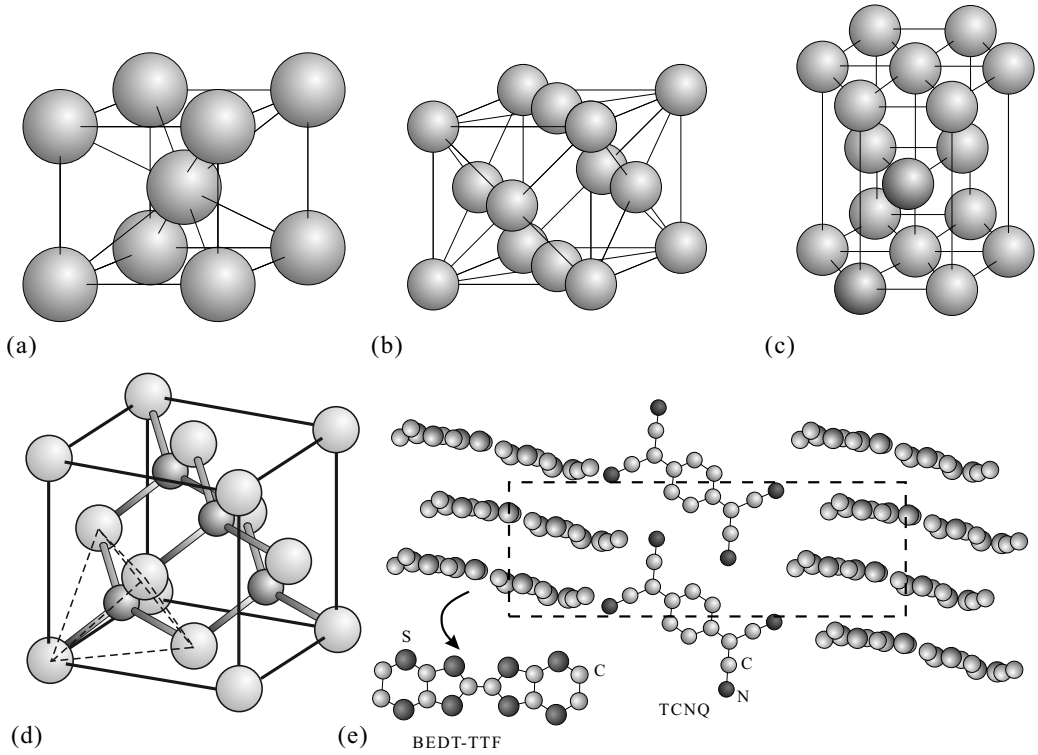


Figure 1.3: Examples of crystal structures. The centers of spheres indicate the positions of nuclei. (a) bcc-structure (Li, Na, Ba etc.), (b) fcc-structure (Al, Ni, Ag, Au etc.), (c) hcp-structure (Mg, Co, Cd etc.). These three often appear in crystals of metals. We get (a) and (b) with simply putting atoms at the lattice points of bcc and fcc lattices in Fig.1.2. (c) is classified into simple hexagonal lattice with taking two atomic positions which have a little darker color, as the basis. (d) is zinc-blende, which often appears in compound semiconductors. (e) is the structure of BEDT-TTF-TCNQ, which is one of organic semiconductors. It is difficult to see the structure in the BEDT-TTF molecults in the main panel. The inset at the left-down shows the molecule structure in the view from the direction vertical to the molecular plane.

is simple cubic. Also, the rhombohedral lattice can viewed as a composite of three regular hexagonal prisms with 120° rotations to each other.

In Bravais lattices, the abbrebiations fcc for face-centered cubic and bcc for body-centered cubic are frequently used. These lattices often appear in metal crystals and the primitive cells often consist of single atoms. Then for the crystal structures, fcc and bcc are also often used. The hexagonal close-packed structure shown in Fig.1.3(c) is also used as a crystal structure that appears approximately well in metal crystals, and the abbreviation hcp is used. The abbreviations for crystal structure are bcc, fcc, and hcp, but there is no hcp "lattice" in the sense defined here. That is, in Fig.1.3, the three atomic positions in the middle of the structure are not equivalent to the peripheral atomic positions, and one atom cannot be taken as a unit structure [1]. The basic structure can be taken as a combination of one atomic position in the upper and lower surfaces of Fig.1.3(b) and one atomic position in the middle, and the Brave lattice is a hexagonal lattice.

Bravais lattice is a classification that focuses on the symmetry of the lattice and is important in the discussion of symmetry, but the fact that the symmetry of the lattice and the symmetry of the crystal are not the same means. It is clear from the fact that the unit cell is regarded as a lattice "point" in the lattice and the details in the unit cell are discarded. Let us take the diamond structure in Fig.1.1 again as an example. The position of the lower left apex of the regular tetrahedron, which is a part of the ptimitive cell, is the basic cell position, and the spatial arrangement is fcc in Fig.1.2. The gray colored parallelpiped is a primitive cell containing two atoms. In Fig.1.1, the difference of the two atomic positions is indicated by shades of colors. In the diamond structure, the atom species is the same for the two. In the case these are alternatively occupied by different species of atoms, *e.g.* Ga and As, the crystal structure is caled zinc-blende(Fig.1.3(d)). That is, zinc-blende structure also belongs to fcc Bravais lattice. On the other hand, two atoms in the primitive cell are of the same

[1] On the web, many "hexagonal close-packed lattices" are searched, but in these explanations, the term "lattice structure" was used for "crystal structure", and this combination was created. Also, "closest packing" mathematically means packing the spheres most densely. Crystals that have a mathematically perfect hcp structure are not known in real atoms because of their anisotropy.

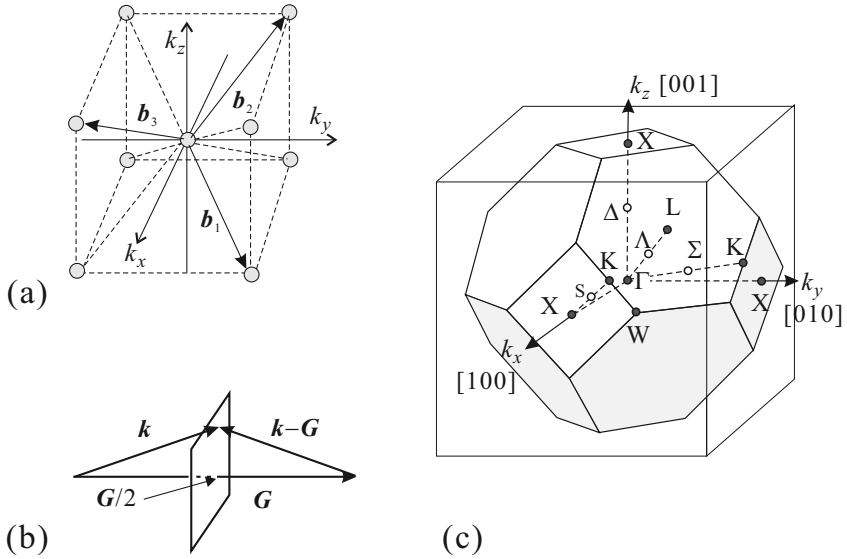


Figure 1.4: (a) Reciprocal lattice of fcc-lattice. The structure is classified to bcc. (b) A way to "cut" the reciprocal space to obtain Brillouin zone. Namely, cut at $\boldsymbol{G}/2$ with the plane perpendicular to $\boldsymbol{G}$, where $\boldsymbol{G}$ is a reciprocal lattice vector. (c) The first Brillouin zone obtained with the cuttings described in (b). The points indicated as Γ, X, L, $\cdots$ are the points with high symmetries.

species in the diamond structure, though of the different species in the zinc-blende structure. The former is symmetric for the inversion operation at the midpoint along the axis connecting the two positions while the latter is asymmetric.

Another expample is in Fig.1.3(e), which shows an organic molecular crystal called $(BEDT\text{-}TTF)_2TCNQ$. The atomic positions take a complecated form though the basis is single molecule and the lattice is triclinic. It is easy to understand the basis has strong anisotropy due to the atomic structure of the molecule and the symmetry of the crystal and that of the lattice is different. The symmetries of crystals are classified by the symmetry operations to 230 space groups.

2.3 Reciprocal lattice, Brillouin zone

Because the lattices of crystals have discrete translational symmetry, a potential $U(\boldsymbol{r})$($\boldsymbol{r}$ is spatial coordinate) caused by the lattice can be expanded in the Fourier seriese as

$$U(\boldsymbol{r}) = \sum_{\boldsymbol{G}} U_{\boldsymbol{G}} e^{i\boldsymbol{G}\boldsymbol{r}}. \tag{1.3}$$

From the periodicity in (1.1), $U(\boldsymbol{r}+\boldsymbol{R}) = U(\boldsymbol{r})$. Then we obtain the condition for $\boldsymbol{G}$ as

$$\boldsymbol{G}\cdot\boldsymbol{R} = 2\pi n \quad (n : \text{integer}), \quad \because e^{i\boldsymbol{G}\cdot\boldsymbol{R}} = 1. \tag{1.4}$$

The vectors $\boldsymbol{G}$ which fulfill the condition (1.4) are called **reciprocal lattice vector**. Just like the real-space lattice, if we define **primitive reciprocal lattice vectors** as

$$\boldsymbol{a}_i \cdot \boldsymbol{b}_j = 2\pi\delta_{ij} \quad (i,j = 1,2,3), \tag{1.5}$$

then we can write down $\boldsymbol{b}_j$ $(j = 1,2,3)$ with $|A| \equiv \boldsymbol{a}_1 \cdot (\boldsymbol{a}_2 \times \boldsymbol{a}_3)$ as

$$\boldsymbol{b}_1 = \frac{2\pi\boldsymbol{a}_2 \times \boldsymbol{a}_3}{|A|}, \quad \boldsymbol{b}_2 = \frac{2\pi\boldsymbol{a}_3 \times \boldsymbol{a}_1}{|A|}, \quad \boldsymbol{b}_3 = \frac{2\pi\boldsymbol{a}_1 \times \boldsymbol{a}_2}{|A|}. \tag{1.6}$$

A reciprocal lattice vector can be represented as $\boldsymbol{G} = \sum_{i=1,2,3} h_i \boldsymbol{b}_i$ $(h_i : \text{integer})$. Generally a function with the periodicity of a lattice can be Fourier expanded with the reciprocal lattice. It is legitimate to say a lattice and the corresponding reciprocal lattice are in the relation of Fourier transformation [2] .

As we considered primitive cells in spatial lattices, we can define the units of periodic repetition in reciprocal spaces. That is the **Brillouin zone**. A general way to obtain Brillouin zones is described in Fig.1.4(b). Let us see how to obtain the first Brillouin zone around the origin. The procesure is simply to cut the reciprocal space with planes containing the points $\boldsymbol{G}/2$ and perpendicular to $\boldsymbol{G}$, where $\boldsymbol{G}$ are the reciprocal lattice vectors starting from the origin. The minimum space (polyhedron) around the origin surrounded by such surfaces is

[2] As an analytic expression, it is enough to remember that the Fourier transform of a regular series of δ-functions is, again, a regular series of δ-functions. This is a very general principle. An example is the optical frequency comb.

the first Brillouin zone. This way of "cutting at $\boldsymbol{G}/2$" will have meaning in considering the band structure, which will be discussed in the next chapter.

The example of fcc-lattice is shown in Fig.1.4. First, the primitive reciprocal lattice vectors are obtained from eq.(1.6). Then we find the reciprocal lattice for the fcc-lattice is bcc-lattice as shown in Fig.1.4(a). Next we apply the method in (b). The reciprocal lattice vectors pointing the nearest neighbor reciprocal lattice points are the primitive reciprocal lattice vectors $\pm\boldsymbol{b}_1,\ \pm\boldsymbol{b}_2,\ \pm\boldsymbol{b}_3$. There are equivalent eight planes which cut the vectors vertically at the midpoints. The polyhedron covered with these planes is a regular octahedron. However, the planes which cut the vectors pointing the next nearest neighbor reciprocal lattice points at the midpoints, also cut the octahedron around the vertices. The procesure thus results in the first Brillouin zone shown in Fig.1.4(c), where $\Gamma,\ \mathrm{X},\ \mathrm{L},\ \cdots$ indicate the points with high symmetry. The points are often used in the display of band structure.

2.4 Crystals often used as semiconductors

II	III	IV	V	VI
${}_4$Be $2s^2$	${}_5$B $2s^2 2p$	${}_6$C $2s^2 2p^2$	${}_7$N $2s^2 2p^3$	${}_8$O $2s^2 2p^4$
${}_{12}$Mg $3s^2$	${}_{13}$Al $3s^2 3p$	${}_{14}$Si $3s^2 3p^2$	${}_{15}$P $3s^2 3p^3$	${}_{16}$S $3s^2 3p^4$
${}_{30}$Zn $3d^{10}$ $4s^2$	${}_{31}$Ga $3d^{10}$ $4s^2 4p$	${}_{32}$Ge $3d^{10}$ $4s^2 4p^2$	${}_{33}$As $3d^{10}$ $4s^2 4p^3$	${}_{34}$Se $3d^{10}$ $4s^2 4p^4$
${}_{48}$Cd $4d^{10}$ $5s^2$	${}_{49}$In $4d^{10}$ $5s^2 5p$	${}_{50}$Sn $4d^{10}$ $5s^2 5p^2$	${}_{51}$Sb $4d^{10}$ $5s^2 5p^3$	${}_{52}$Te $4d^{10}$ $5s^2 5p^4$

Which materials shold be called "semiconductors" is a difficult problem, and some scholars propose the classification of "every material that is not metal". In fact, diamond, which was a typical insulator a while ago, has recently completely established itself as a semiconductor. Here, let's have a quick look at the simple and clear "crystals" of the spatial periodic structure, and those that are often used as semiconductors in the industry.

As specific examples, we take materials consist of comparatively small numbers of elements from Group II to Group VI in the periodic table. In the right table, we show the part of the periodic table under consideration with the electronic orbital occupation. We can guess from the table that the semiconductors composed of these elements takes similar lattice structures. Here we mainly introduce crystal structures.

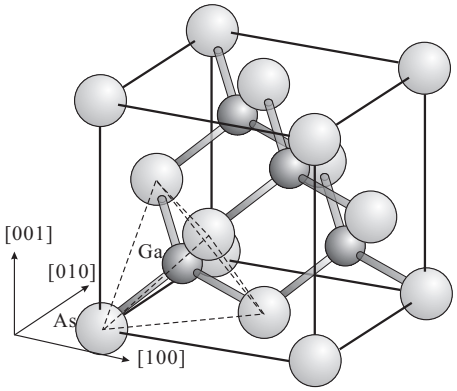
2.4.1 Group IV semiconductors

Elementary semiconductors of C, Si, Ge take **diamond structure** (Bravais lattice is fcc). The bonds in these crystals are dominated by covalent binding of sp^3 hybrid orbitals. Silicon (Si) is of course the king of semiconductors in the industry. Tin (Sn) are metals in many phases but the form called α-Sn (gray tin) is a semiconductor with diamond crystal structure.

Creation of low dimensional electron system is a big charm of semiconductor physics, and it is also very important in the semiconductor industry. In the case of Si, a metal-oxide-semiconductor (MOS) structure has long been used to create two-dimensional electron systems. Since the oxide layer generally takes an amorphous structure, the interfacial scattering probability of two-dimensional electrons is high, and it is difficult to obtain an electron system with high mobility. On the other hand, a two-dimensional electron system with high mobility has been realized by a method of forming a bf heterojunction using mixed crystals of Si-Ge.

semiconductor	lattice constant Å	energy gap (RT eV)	electron mass m_0	hole mass
C	3.56683	5.47	0.25	0.2
Ge	5.64613	0.66	1.64, 0.082	0.04, 0.28
Si	5.43102	1.12	0.98, 0.19	0.16, 0.49

2.4.2 III-V compound semiconductors

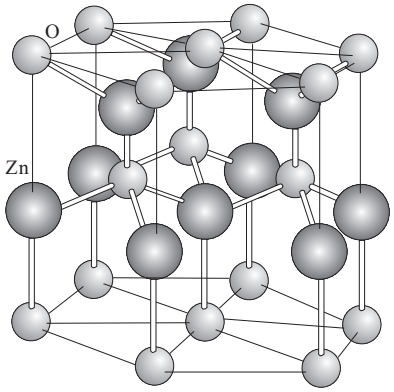


Semiconductors made by combining group III elements and group V elements on a one-to-one basis, group III and group IV atoms occupy the lattice points of the diamond structure alternately, that is, **zinc blende** structure. Al, Ga, In are often used as group III, and As, P, Sb, etc. are often used as group V. Many kinds of compound semiconductors are formed by these combinations, and more kinds of semiconductors can be synthesized by further mixing different kinds of elements to form bf mixed crystal. Strictly speaking, these are no longer crystals because they have lost the spatially regular structure, but most of the concepts in the crystals work well by considering some blunting due to lattice disorder.

In reality, these compound semiconductors and mixed crystal semiconductors are often synthesized by bf epitaxial growth, and for this purpose, the crystal form and lattice constant should be similar between the heterogeneous semiconductors to be joined. This will be described later.

Many of III-V semiconductors have a direct gap at the origin of the reciprocal lattice space, Γ point, and are therefore often used for optical devices. In addition, there are many combinations that can form high-quality heterojunctions by epitaxial growth, and they are frequently used for devices for high-speed operation.

2.4.3 III-N compound semiconductors



In nitride semiconductors, whose applications have expanded rapidly for blue light emitting diodes, high-frequency, high-power devices, GaN, InN, and AlN are currently the main research targets. They take hexagonal Wurtzite crystal forms. They are usually grown by use of epitaxial growth and annealed at high temperature to improve their quality. The figure in the right shows the crystal structure of Wurzeit (for the case of ZnO).

2.4.4 II-VI compound semiconductors

Group II-VI semiconductors take various crystal forms such as zinc-blende, wurtzite, and chalcopyrite. There are various compounds such as ZnO and CdTe, and before GaN became the leading player in blue-color optical devices, the II-VI system was mainly studied as a candidate. ZnO is still considered to be a material that threatens the GaN system if the device characteristics and manufacturing method are improved because the material is easily available. The ZnO system tends to have a small structure such as nanotubes, and it is not easy to form it into a thin film device. On the other hand, its application as a nanostructure device is also attracting attentions. The Hg system is said to have a "negative bandgap" and has become well known for its use in constructing topological materials.

2.5 Organic semiconductor materials

Organic thin films as semiconductors are attracting attention because they are lightweight, flexible, and inexpensive. Most organic solids are molecular solids in which intermolecular bonds are formed by van der Waals forces. The qualities of organic semiconductors have been imporoved and the various concepts of semiconductor physics are now applicable to organic ones. However, it is often more realistic to consider the electronic state in the molecule and the solid state as an aggregate of them separately, reflecting the fact that they are molecular solids. Especially in the case of macromolecules, on one hand Bloch electrons and bands in the molecule are good approximation, on the other hand, the electrical conduction of the whole solid should be analyzed with the theory for amorphous solids developed in the 1970s and 1980s.

3 Crystal growth

In order to utilise the structural sensitivity of semiconductors as functions to explore condensed matter physics, to setup them as laboratories of quantum and many-body effects and to use them as devices, we need to obtain,

as the starting point, obtain crystals with very low concentrations of defects and impurities. For that, the original materials with ultra high purity, higher than those in reagents by orders, should be prepared with cheap cost, huge amount, in very short time, and with very low environmental load. The crystal growth is hence a high field in the semiconductor industrial science. The physics, the main issue of this lecture, is not directly connected to that field but I would like to introduce some in a very short time.

Crystal growth methods of inorganic semiconductors can be classified to one for three-dimensional bulk growth and another for two-dimensional growth on wafers of crystals cut from three dimensional ingot. The latter is called epitaxial growth.

3.1 Growth of bulk crystals

Mining and refinement of source materials are very important processes before the crystal growth, and we need to choose the best degree of material quality and refinement method considering the cost and the final product. In the case of crystalline silicon, it is said that astonishing purity of 11N(99.999999999%) is required for substrates of MOS-LSI [3], which is called "semiconductor grade".

On the other hand, a solar cell device generally has an area 10 orders of magnitude wider than that of MOS-LSI, and the tolerance for leakage current per area also differs by a few orders of magnitude. Hence for them, the purity of 6N∼7N is enough under reduction of impurities that form non-radiative recombination centers or pn characteristics degrading deep levels. Such wafers are called in "solar grade".

In the latter, usually low quality Si called "metal grade" is used as a starting material. There have been long term seekings for purification method with low power consumption and some new progress has been made though the world market is now dominated by companies which provides cheap wafers produced with traditional method in 2013. Such situation is largely affected by international affairs or economic atmosphere. I am sorry but must say that "basic researches" are affected by such political situation in reality.

Bulk single crystals of inorganic semiconductors are usually obtained from gradual solidification by cooling from high temperature melts. This is comparatively easy for single element semiconductor Si or Ge. In the growths of compound semiconductors, mixed melts of multiple elements should be prepared and the difference in melting point, vapour pressure and mutual solubility are the possible problems.

3.1.1 Czochralski process

In **Czochralski** (CZ) process, as illustrated in Fig.1.5 a thin seed crystal is put down to the surface of a melt from source materials, and a thick cylindrical crystal is pulled up. The grown crystal is formed in a cylindrical shape because the seed is rotated during the pulling up growth process. This is a representative method to

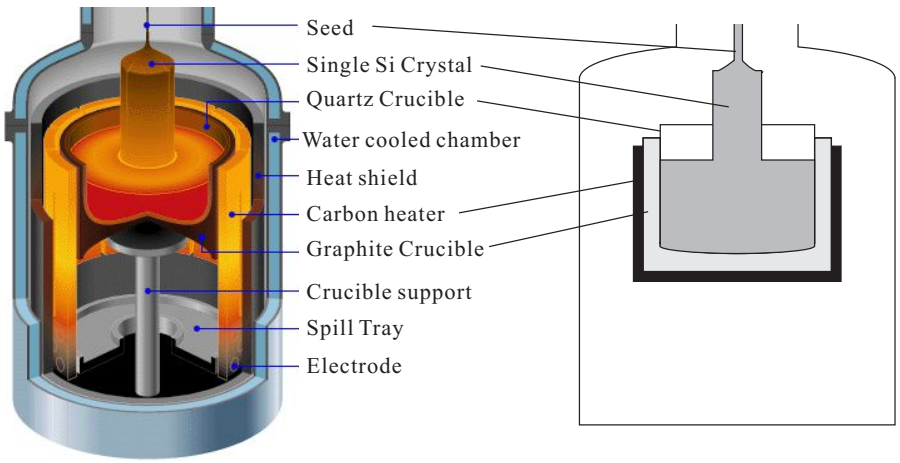


Figure 1.5: Schematic drawing of Czochralski process. Left: Three-dimensional schematic illustration. Right: Cross-sectional illustration.
From http://people.seas.harvard.edu/ jones/es154/lectures/lecture_2/materials/materials.html

[3]Here they are using a special definition of "purity". I have experienced that such an ultra-pure Si ingot contains a considerable amount of oxygen measured from low temperature magnetic susceptibility measurement with a SQUID magnetometer. 11N is hence the value on the ignorance of these impurities. Oxygen has little effect on logic LSIs but is a problem in the application for power devices.

obtain a dislocation free crystal of Si. The thin disk form popular for LSI wafer appears after slicing the cylindrical columnar shape.

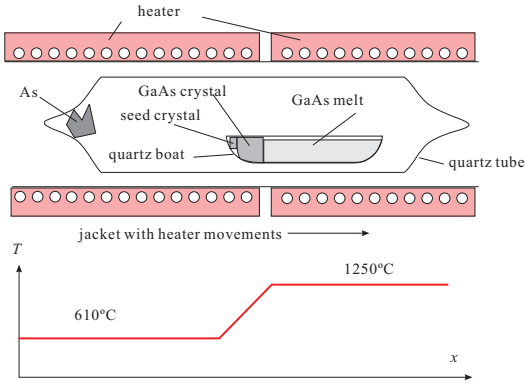


Figure 1.6: Schematic illustration of a boat method (horizontal Bridgeman process).

Such a primitive CZ method cannot generally be applied to compound semiconductors due to large difference in the evaporation pressure. Actually CZ method is often adopted in growth of III-V semiconductors GaAs, InP, GaP, etc. but not in the primitive form because the group V materials have much higher vapour pressures than those of the group III, resulting in the rapid escape of group V materials from the melts. Instead, **Liquid Encapsulated Czochralski** (LEC) process, in which the melt of the sources is encapsulated with an inert liquid like B_2O_3.

3.1.2 Boat method

Another popular method for bulk-growth of compound semiconductors is the one called "boat method". The boat method is further classified into horizontal Bridgeman (HB) method and temperature gradient freeze method. In the former, a furnace with two temperature regions is moved along a boat, in which the source materials are melt, and from one end the melt is freezed into a single crystal.

Figure 1.6 shows a schematic illustration of HB method for the case of GaAs. Initially a metallic solid As is set to one end of a quartz tube, which also contains a quartz made boat. In the beginning Ga melt and a seed crystal are in the boat. The side of the As metal is heated to 610 °C while the other side to 1250 °C. As sublimates severely above 600 °C and gets into Ga melt forming GaAs melt. At 1250 °C, GaAs is in melting phase and at 610 °C in solid phase. As the furnace moves to the right in the figure, a GaAs single crystal is solidified from the end of the seed crystal to the right.

3.1.3 Zone melting method

As mentioned in the footnote in pageE1-8, "ultrahigh purity" Si actually often contains high concentration of oxygen, which mainly comes from the crucibles during the growth. For power MOS FET or other devices in which such oxygen causes troubles, single crystals are grown by **floating zone** melting (FZ) method. In the initial stage, a rod of polycrystal with a high purity is prepared in standing manner and a seed crystal is put on top of the rod. At first a zone of the polycrystal rod from the top is melted e.g., by concentration of infrared beam with confocal method or by rf loss heating. The melt in contact with the seed crystal changes into single crystal and the melted zone slowly goes down to form a single crystal rod. During the process the melt does not touch any other materials and the high purity of polycrystal is kept. On the other hand, such big radiuses of grown rods as those in CZ method cannot be obtained.

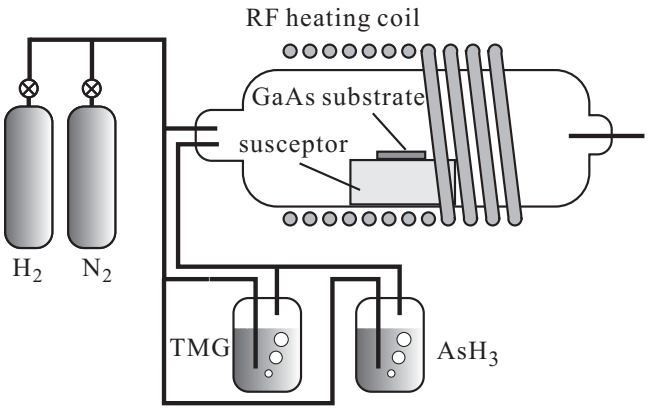


Figure 1.7: Extremely simplified schematic illustration of MOVPE (MOCVD) apparatus of GaAs deposition. The "suscepter" absorbs the power of RF and gets heat.

3.2 Epitaxial growth of thin films

Epitaxial growth, in which thin crystal films are grown with deposition of materials onto crystal substrates, is classified into **liquid phase epitaxy** (LPE), **vapor phase epitaxy** (VPE), and epitaxy in vacuum or in very low pressure gas. Here I will pick up metal organic vapor phase epitaxy (MOVPE) and molecular beam epitaxy (MBE) from the number of epitaxial growth methods.

3.2.1 Metal-organic vapor phase epitaxy

Also often called as Metal organic chemical vapour deposition (MOCVD). Often used for the growth of compound semiconductors. Let us see the case of GaAs.

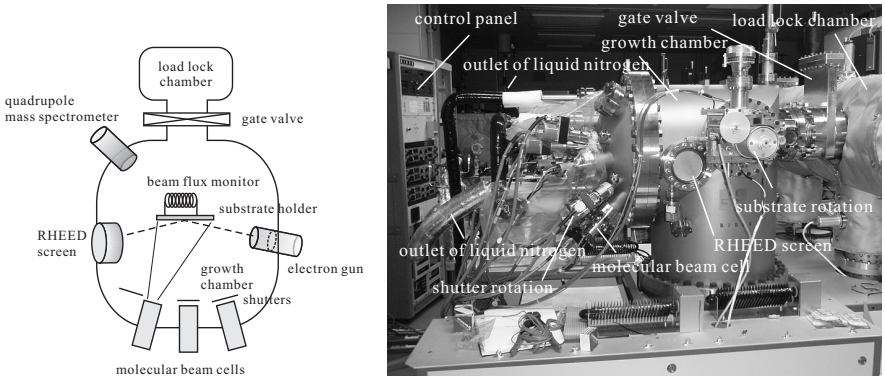
In general in epitaxial growth of a thin film crystal, component materials are carried onto the substrate with some carriers or with some other method, and react with the substrate surface to form single crystal films. Therefore the keys for the growth are the surface states of the substrate, the way of carrying the materials, the dynamics of deposited atoms, etc. In the case of MOVPE, the sources are carried by hydrogen and nitrogen gases. Ga is put on tri-methyl gallium ($(CH)_3Ga$, TMG), and As on arsine (AsH_3). They are carried onto the substrate and decomposed into atoms on the surface by heating. Then they are chemically bonded to the surface atoms to form GaAs crystal films. Omitting all the intermediate chemical reactions and the initial and the final states can be written as

$$(CH_3)_3Ga + AsH_3 \longrightarrow GaAs + 3CH_4.$$

TMG and arsine have low vapor pressures and as shown in Fig.1.7, liquids of them are bubbled with hydrogen to be vaporized. Hydrogen gas is deoxidizing atmosphere for GaAs surface. Thus flat and high quality films can be grown though the actual chemical reaction is not so simple. Doping of impurities, growth of mixed crystals are possible with preparation of materials. All of metal organic gases of group II or III, arsine or phosphine of group V are explosive and at the same time nerve gases. They are extremely dangerous and should be treated with highest care and rigid safeguards.

3.2.2 Molecular beam epitaxy

Molecular beam epitaxy (MBE) is a representative growth method of ultra-thin semiconductor films. Characteristic features are: (1) deposition in ultra-high vacuum; (2) single crystalline substrates and various methods for surface cleaning; (3) heating of substrates to activate the motions of deposited atoms; (4) stoichiometric deposition of materials is not always required; (5) *in situ* characterization of grown films in various ways is possible because the growth front is always on the surface to the vacuum.



(a) (b)

Figure 1.8: (a) Schematic illustration of an MBE machine. (b) Photograph of a real machine (RIBER S32).

Figure1.8(a) is a schematic illustration of an MBE machine, (b) shows a photograph of a real machine. In order for keeping ultra-high vacuum in the growth chamber, (a) "pre-evacuation chamber(s)" is used for loading and unloading of substrates. Molecular beam cells (Knudsen cells, K-cells; Langmuir cells, L-cells), which have sources of evaporation in them, are kept at intermediate temperatures for them to avoid adsorption of gas molecules. While the growth, the shrouds covering substrate, heating system and molecular beam cells are cooled down with continuous flow of liquid nitrogen to adsorb outgassed molecules. When evaporations are going on, the source molecules are inside the chamber and the total gas pressure increases, which makes the quality of vacuum obscure. To monitor the quality, we need a partial pressure gauge for gas species. That requirement, not 100 % is fulfilled by a mass spectrometer, with which partial pressure can be measured as a function of the ratio of the molecular mass to the charge.

Substrates for growths are introduced into the preparation (pre-evacuation) chamber after surface cleaning with chemical etching and protection of the cleaned surface with oxidation. The oxide film at the surface is evaporated simply by heating the substrate in ultra-high vacuum.

To confirm the evaporation and to see the growth mode during the growth, we need some in-situ monitor of the surface state. For that purpose a conventional method is refractive high energy electron diffraction, RHEED. In the RHEED technique, as illustrated in Fig.1.8(a), electron beam with 15∼30keV acceleration is injected onto the surface with very shallow angle and the diffraction pattern of reflected beam is imaged on the illumination screen. The image reflects the atomic structure of the surface, that is, it is the pattern of reciprocal

lattice. Because the incident beam goes onto the surface with very shallow angle, when the surface is a mirror, the diffraction is close to two-dimensional, that is, the diffraction pattern is a set of vertical reciprocal lattice "rods".The image on the screen is a slice of the reciprocal rods with a plane almost parallel to the rods. Actual diffraction patterns of rods have some widths due to various reasons and in such a two-dimensional growth, images like upper-left of Fig.1.9 are obtained.

The image in Fig.1.9 has a strong diffraction spot at upper-center. This is due to the simple mirror-like reflection from the surface (mirror spot) and the more flat is the surface, the higher the intensity is. After opening the shutters molecular beams reach the substrate and the growth starts. Molecules or atoms migrate on the surface of the substrate with thermal activation after adsorbation and hit the lattice points at last, forming strong bonding to substrate crystals. This is one of the possible mechanisms for crystal growth and such a "state of growth" is called "growht mode". The growth mode mentioned above is called layer-by-layer mode.

In the initial stage of layer-by-layer mode, one atomic monolayer growth contains a cycle from a flat surface through a rough surface to a new flat surface. Such a single cycle causes one period in intensity oscillation of mirror spot. The oscillation hence makes it possible for us to monitor the growth of each atomic layer. The oscillation damps in proceeding of the growth due to some incoherency though in many cases growth interruption recovers the flatness for lowering the surface energy of roughness. These properties opens up a way to "flat surface growth" of "migration enhanced epitaxy", in which the intensity of mirror spot is monitored and the shutters are controlled to keep the highest intensity in the oscillation.

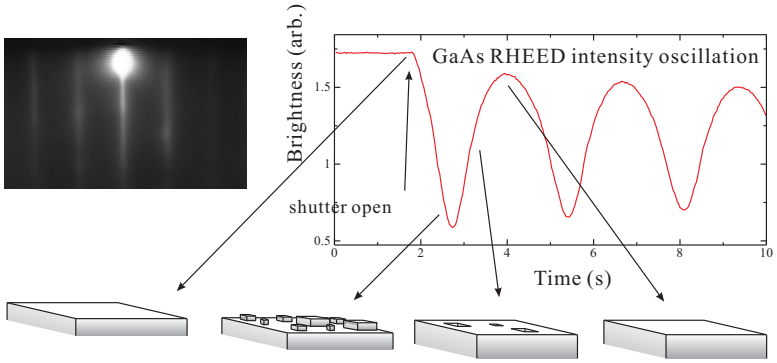


Figure 1.9: Upper-left: RHEED image for two-dimensional MBE growth. The intense spot at upper-center is the mirror spot. The right shows the oscillation of the intensity at the mirror spot as the film grows. The cartoons illustrate the surface states at indicated points in the intensity oscillation of the mirror spot.

With increasing substrate temperatures, generally the dominant growth mechanism changes into "step flow mode", in which migrating atoms on the surface attach to the edges of surface steps causing widening of terraces, that is, flow of steps. In this mode no oscillation occurs.

References

[1] N. W. Ashcroft and N. D. Mermin, "Solid State Physics" Chapter 4 (Brooks/Cole Publishing, 1976).

[2] D F Johnston, Rep. Prog. Phys., **23**, 66 (1960).

[3] D. K. Ferry, "Semiconductors: Bonds and bands" (IOP Publishing, 2013).

Chapter 2 Band structure, effective mass approximation

In solid-state physics, the term "band structure" refers to the energy dispersion relations of the crystal eigenstates in the reciprocal lattice space introduced in the previous chapter. The theme of this chapter is introduction of the concept and how to calculate it theoretically. In addition, we will introduce the effective mass approximation, which is indispensable for handling band electrons in a simple and clear view.

1 Band electrons

In free space, the kinetic energy of an electron takes a continuous value from zero. On the other hand, the kinetic energy takes a discrete value in the bound state in the localized potential of the nucleus. There are two views on the energy eigenstates in the periodic potential. One is the perturbation to the state in free space, which creates a section (energy gap) where the eigenvalues do not exist, and the energy eigenvalues are cut to bands. The other is that the discrete level due to the localized potential spreads in a band due to the tunnel between the adjacent sites. The former is called **nealy free electron approximation** (NFEA), and the latter is called **tight-binding approximation** (TBA).

1.1 Bloch theorem

It is needless to proove the Bloch theorem, which is very basics of the solid state physics. For the reference, the theorem is described as follows. Energy eigenstates in a periodic potential are expressed in the real space ($\boldsymbol{r}$) expression as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = u_{n\boldsymbol{k}}(\boldsymbol{r}) \exp(i\boldsymbol{k}\cdot\boldsymbol{r}), \tag{1.1}$$

where n is the band index, $u_{n\boldsymbol{k}}$ is a function with the lattice periodicity, *i.e.*

$$\forall \boldsymbol{R} \in \{(\text{lattice vector})\},\ u_{n\boldsymbol{k}}(\boldsymbol{r}) = u_{n\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{R}). \tag{1.2}$$

Here $\boldsymbol{k}$ is the wavenumber.

1.2 Nearly free electron approximation (NFEA)

We write the equation for the eigenstates in a lattice potential as

$$\mathscr{H}\psi(\boldsymbol{r}) = \left[-\frac{\hbar^2}{2m_0}\nabla^2 + V(\boldsymbol{r})\right]\psi(\boldsymbol{r}) = E\psi(\boldsymbol{r}), \tag{1.3}$$

where $V(\boldsymbol{r})$ is the lattice potential.

From the periodicity of $V(\boldsymbol{r})$, $u_{\boldsymbol{k}}(\boldsymbol{r})$, they can be Fourier expanded as

$$V(\boldsymbol{r}) = \sum_{\boldsymbol{G}} V_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}, \quad u_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}, \tag{1.4}$$

where $\boldsymbol{G}$ are the reciprocal lattice vectors. With substituting (1.1) and (1.4) into the Schrödinger equation (1.3), we obtain

$$\sum_{\boldsymbol{G}}\left[\left\{\frac{\hbar^2}{2m_0}(\boldsymbol{k}+\boldsymbol{G})^2 - E)\right\}C_{\boldsymbol{G}} + \sum_{\boldsymbol{G}'} V_{\boldsymbol{G}-\boldsymbol{G}'}C_{\boldsymbol{G}'}\right] e^{i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}} = 0.$$

Because each term in the sum of $\boldsymbol{G}$ should be zero, the following simultaneous equations for $\{C_{\boldsymbol{G}}\}$ are obtained.

$$\sum_{\boldsymbol{G}'}\left[\left\{\frac{\hbar^2}{2m_0}(\boldsymbol{k}+\boldsymbol{G})^2 - E\right\}\delta_{\boldsymbol{G}\boldsymbol{G}'} + V_{\boldsymbol{G}-\boldsymbol{G}'}\right]C_{\boldsymbol{G}'} = 0. \tag{1.5}$$

The condition for eq.(1.5) to have non-trivial solutions is

$$\left|\left[\left\{\frac{\hbar^2}{2m_0}(\boldsymbol{k}+\boldsymbol{G})^2 - E\right\}\delta_{\boldsymbol{G}\boldsymbol{G}'} + V_{\boldsymbol{G}-\boldsymbol{G}'}\right]_{\boldsymbol{G}\boldsymbol{G}'}\right| = 0. \tag{1.6}$$

In NFEA, we consider the pertubation $\delta V_{\boldsymbol{G}-\boldsymbol{G}'}$ to ($V(\boldsymbol{r}) = 0$)

$$\psi(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}}, \quad C_0 = 1,\ C_{\boldsymbol{G}} = 0\ (\boldsymbol{G} \neq 0),\ E = \frac{\hbar^2\boldsymbol{k}^2}{2m_0}. \tag{1.7}$$

As a result of pertubation, $\delta C_{\boldsymbol{G}}$ is caused. In (1.5), the terms $\delta V \delta C$, $\delta E \delta C$ are in the higher order to be ignored. Then

$$\frac{\hbar^2}{2m_0}[(\boldsymbol{k}+\boldsymbol{G})^2 - \boldsymbol{k}^2]\delta C_{\boldsymbol{G}} + V_{\boldsymbol{G}} = 0 \quad \therefore \delta C_{\boldsymbol{G}} = \frac{2m_0}{\hbar^2}\frac{-V_{\boldsymbol{G}}}{(\boldsymbol{k}+\boldsymbol{G})^2 - \boldsymbol{k}^2}.$$

However the approximation collapses at

$$(\boldsymbol{k}+\boldsymbol{G})^2 - \boldsymbol{k}^2 = 0. \tag{1.8}$$

Therefore around the point (1.8), we approximate that only C_0 and $C_{\boldsymbol{G}}$ are non-zero. The we can write down (1.6) as

$$\begin{vmatrix} \frac{\hbar^2}{2m_0}\boldsymbol{k}^2 - E & V_{-\boldsymbol{G}} \\ V_{\boldsymbol{G}} & \frac{\hbar^2}{2m_0}(\boldsymbol{k}+\boldsymbol{G})^2 - E \end{vmatrix} = 0, \tag{1.9}$$

which gives the energy eigenstates as

$$E = \frac{1}{2}[E^{(0)}(\boldsymbol{k}) + E^{(0)}(\boldsymbol{k}+\boldsymbol{G})] \pm \frac{1}{2}\sqrt{[E^{(0)}(\boldsymbol{k}) - E^{(0)}(\boldsymbol{k}+\boldsymbol{G})]^2 + 4|V_{\boldsymbol{G}}|^2}, \tag{1.10}$$

where $E^{(0)}(\boldsymbol{k}) \equiv \hbar^2\boldsymbol{k}^2/2m_0$. The result indicates the appearance of the energy separation of $\pm V_{\boldsymbol{G}}$(**bandgap** or **forbidden band**). For a system with the lattice constant a, the condition (1.8) is $2a\cos\theta = n\lambda$(n is an integer, λ is the wavelength of electron). This is nothing but the Bragg condition for diffraction of waves. Thus the result can be interpreted as the electron wave get a Bragg reflection from the lattice and the interference between the waves creates a standing wave, which results in the bandgap.

1.3 Reduced zone expression

A Bloch function can be written as fowllows with $\boldsymbol{G}$ a reciprocal lattice vector as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = u_{n\boldsymbol{k}}(\boldsymbol{r})e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = u_{n\boldsymbol{k}}(\boldsymbol{r})e^{-i\boldsymbol{G}\cdot\boldsymbol{r}}e^{i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}}.$$

Function $v(\boldsymbol{r}) \equiv u_{n\boldsymbol{k}}(\boldsymbol{r})e^{-i\boldsymbol{G}\cdot\boldsymbol{r}}$ also has the periodicity $v(\boldsymbol{r}) = v(\boldsymbol{r}+\boldsymbol{R})$, where $\boldsymbol{R}$ is lattice vectors. $\psi_{n\boldsymbol{k}}$ can thus be expressed as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \xi_{n'\boldsymbol{k}+\boldsymbol{G}}(\boldsymbol{r}) \tag{1.11}$$

with another Bloch function $\xi_{n'\boldsymbol{k}}$. Namely, the expression in (1.1) has the arbitrariness of reciprocal lattice vectors. In other words, when a wavefunction has a spatial modulation of lattice period, there is ambiguity whether the modulation is included in the lattice periodic part $u(\boldsymbol{r})$ or in the plane wave part $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$.

On the other hand, the system represented by Schrödinger equation (1.3)has the time-reversal symmetry and $E(\boldsymbol{k}) = E(-\boldsymbol{k})$. The above two relations on $E(\boldsymbol{k})$ leads to $E(\boldsymbol{G}+\boldsymbol{k}) = E(\boldsymbol{G}-\boldsymbol{k})$. That is, $E(\boldsymbol{k})$ is symmetric to the zone boudaries.

The arbitrarity in (1.11) leads to the arbitrarity in the representation of $E(\boldsymbol{k})$. As Fig.1.2(a), in **extended zone representation**, $E(\boldsymbol{k})$ is represented as a single-valued function of $\boldsymbol{k}$ while as in Fig.1.2(b), in **reduced zone representation**, representation of $E(\boldsymbol{k})$ is folded into the first Brillouin zone.

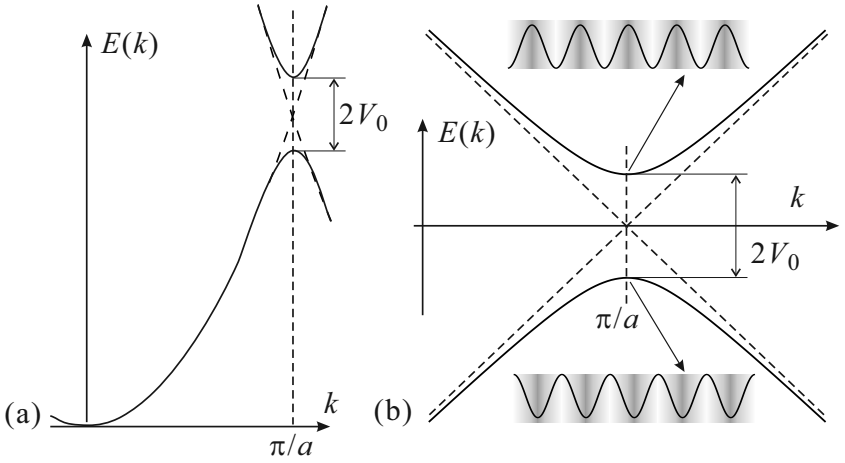


Figure 1.1: (a) In NFEA, a bandgap of (1.10) appears at $k = G/2$. (b) Blowup of the region around the bandgap in figure (a). The cartoons explain why the standing waves get the energy gap

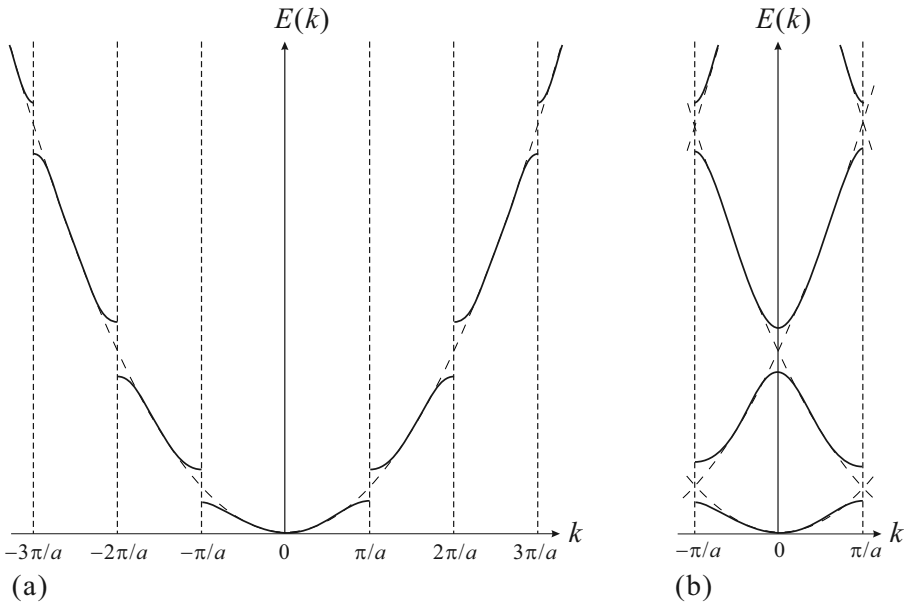


Figure 1.2: Two ways of expression for the energy bands of NFEA. (a) Extended zone expression. (b) Reduced zone expression.

distance :	reciprocal lattice	number of points
0 :	(0,0,0)	1
$\sqrt{3}$:	(1,1,1), (1,1,-1), (1,-1,1), $\cdots$	8
2 :	(2,0,0), (0,2,0), (0,0,2), (-2,0,0),$\cdots$	6
$\sqrt{8}$:	(2,2,0), (2,0,2), (0,2,2), (-2,2,0), $\cdots$	12
$\sqrt{11}$:	(3,1,1), (1,3,1), (1,1,3), (-3,1,1), $\cdots$	24
.....		

Table 2: Classification of reciprocal lattice points with the distance from Γ-point (unit $G_0 \equiv 2\pi/a$).

In Fig.1.2(b), the second and the third bands is obtained from cutting for the first Brillouin zone $[-\pi/2, \pi/2]$ of the overlap of the extended zone representations from the two neighboring reciprocal lattice points $k = \pm 2\pi/a$. This is a natural consequence that the reduced zone representation is possible from the arbitrariness of the reciprocal lattice vectors as in (1.11).

1.4 Empty lattice approximation

In NFEA, in the limit of $V_0 \to 0$, this is nothing but a free space and the energygap disappears, the dispersion relation is simply paraboric. However, the free space is not a system in which the spatial periodicity of the lattice is lost, and the continuous translational symmetry also includes the periodic translational symmetry of the lattice. Rather, it can be considered that the lattice of the empty primitive cell remains. In the **empty lattice approximation**, we hence consider the reduced zone representation of the parabolic energy dispersion. In Bloch function representation, the plane wave function $e^{i\boldsymbol{k}'\cdot\boldsymbol{r}}$ is separated into the lattice periodic part $u_{n\boldsymbol{k}}(\boldsymbol{r})$ and the crystal wavenumber part $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ and apply reduced zone representaiton.

Let's take an example with a three-dimensional crystal. Consider the reciprocal lattice and Brillouin zone in Fig.1.4 in the case of fcc. First, to obtain the reduced zone representation, since the principle of reduced zone representation is indefiniteness of the reciprocal lattice vector as in (1.11), consider the reciprocal lattice of fcc and the bcc lattice of Fig.1.4 (a), and draw parabola with the origin at each reciprocal lattice point. Then we cut the diagrams with the first Brillouin zone shown in Fig.1.4(c). The reciprocal lattice points we need to consider in this drawing are summarized in Tab.2. The farer from the origin, the higher the energy branch of the parabora from the reciprocal lattice point.

The problem with a three-dimensional band structure expression is how to display it. It is not possible to draw multiple parabolas in a three-dimensional space. Usually we only draw the energy dispersion on some representative lines in the reciprocal space. Fig.1.3 shows a way of drawing often used to display the band structure. The energy dispersions on the lines which connect points with high symmetries are drawn. As in

the figure the line goes along L→ Γ →X→K→ Γ. (a) shows the empty lattice approximation while (b) shows the realistic band dispersion in Si calculated with empirical pseudo-potential approximation (we will see in the next week). There is no bandgap in the empty lattice approximation naturally. On the other hand, we see clear resemblance between them. When we go into realistic calculations with gaps, the diagram is usuful to seew which branch corresponds to which reciprocal point. Furthermore, when a level repulsion causes energygap, we need to consider symmetry of the lattice and the empty lattice approximation is also usuful for seeing that.

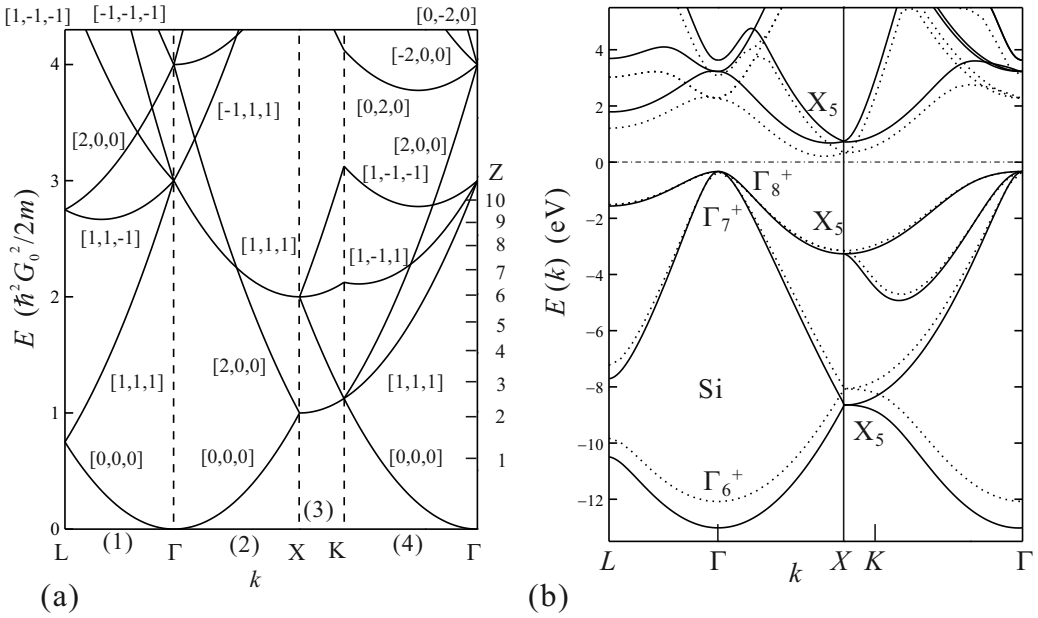


Figure 1.3: (a) Empty lattice approximation of fcc-lattice. Three numbers in [· · ·] indicate corresponding origins of parabolas. (b) Realistic band structure of Si calculated by empirical pseudo-potential method.

1.5 Tight binding approximation

In the next week we begin with tight-binding approximation.

2.1.5 Tight-binding approximation

In the nearly free electron approximation (NFEA), the lattice potential causes interference of electron waves in free space and opens “slits", namely energy gaps in the original continuous energy spectrum. On the other hand in the view of tight-binding approximation (TBA), the energy bands with finite widths are formed from originally discrete energy levels in spatially localized potential at each site by electron tunneling between the different sites.

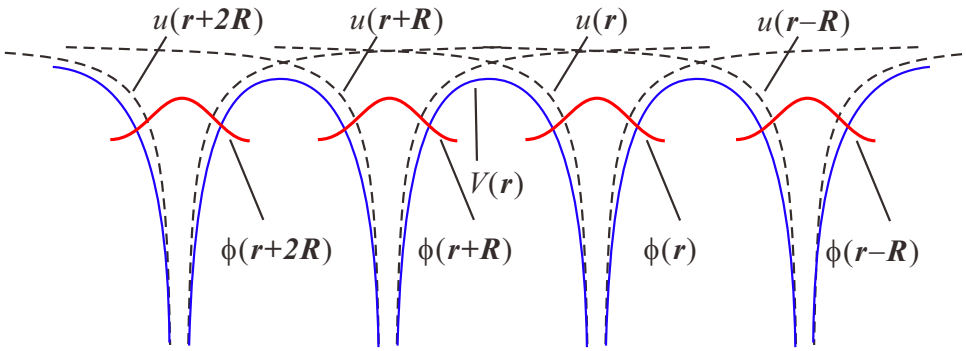


Fig. 2.4 Schematic diagram of TBA. The periodic potential (blue lines) is the sum of atomic potentials represented by broken lines. In TBA, the linear combination of localized orbitals drawn by red lines should be in the form of Bloch function.

The concept is illustrated in Fig. 2.4. The Hamiltonian for single-site is written as

$$\mathscr{H}_a = \hat{T} + u, \tag{2.12}$$

where $\hat{T}$, u are the kinetic energy and the site-localized potential respectively. Let $\boldsymbol{R}_i$ be the potision of site i, then the real-space representation of $\mathscr{H}_a$ can be written as $\mathscr{H}_a(\boldsymbol{R}_i) = \hat{T} + u(\boldsymbol{r} - \boldsymbol{R}_i)$. If we write the normalized orthogonal eigenstates of $\mathscr{H}_a(0)$ as $\{\phi_n\}$, then

$$\mathscr{H}_a(\boldsymbol{R}_i)\phi_n(\boldsymbol{r} - \boldsymbol{R}_i) = \epsilon_n\phi_n(\boldsymbol{r} - \boldsymbol{R}_i), \tag{2.13}$$

where n is the level index of the discrete states. We write the periodic potential obtained by overlapping of u as $V(\boldsymbol{r})$, the total Hamiltonian for such system illustrated in Fig. 2.4 is

$$\mathscr{H} = \hat{T} + V(\boldsymbol{r}). \tag{2.14}$$

Let us consider the solution of the eigenequation of the above Hamiltonian. For simplicity we ignore the direct overlapp integral of wavefuntions ϕ_n between the neighboring sites($\langle\phi_n(\boldsymbol{r} - \boldsymbol{R}_i)|\phi_n(\boldsymbol{r} - \boldsymbol{R}_j)\rangle = \delta_{ij}$). We also ignore the matrix elements of $\mathscr{H}$ between the states of different n. Then the total wavefunction can be written in the linear combination of ϕ_n Because all of the lattice sites are equivalent, each coefficient of the linear combination should be in the form of $c/\sqrt{N}$ where $|c| = 1$, c is constant for $\boldsymbol{r}$. Furthermore, the linear combination should be written in the form of the Bloch function. From the above requirements the eigenstate should be written in the following form.

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{N}}\sum_i e^{i\boldsymbol{k}\cdot\boldsymbol{R}_i}\phi_n(\boldsymbol{r} - \boldsymbol{R}_i) = \frac{e^{i\boldsymbol{k}\cdot\boldsymbol{r}}}{\sqrt{N}}\left[\sum_i e^{-i\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{R}_i)}\phi_n(\boldsymbol{r} - \boldsymbol{R}_i)\right]. \tag{2.15}$$

Since the last part of $[\cdots]$ has the periodicity of lattice, the form of (2.15) is in the Bloch form.

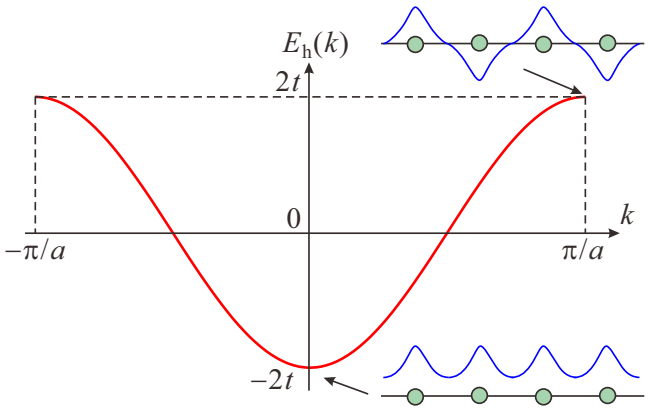


Fig. 2.5 Band dispertion relation in one-dimensional tight-binding approximation. In the model the hopping integral is finite only for nearest neighbor and t, in which situation cosine-shaped band structure is obtained. The insets illustrate the standing wave amplitude, phase at the band bottom and the top.

The expectation value of $\mathscr{H}$ (energy expectation value) with $\psi_{n\boldsymbol{k}}$ is

$$\begin{aligned}
\langle\psi_{n\boldsymbol{k}}|\mathscr{H}|\psi_{n\boldsymbol{k}}\rangle &= N^{-1}\sum_{i,j} e^{i\boldsymbol{k}\cdot(\boldsymbol{R}_j-\boldsymbol{R}_i)}\langle\phi_n(\boldsymbol{r}-\boldsymbol{R}_i)|[\hat{T}_r+V(\boldsymbol{r})]|\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\rangle \\
&= N^{-1}\sum_{i,j} e^{i\boldsymbol{k}\cdot(\boldsymbol{R}_j-\boldsymbol{R}_i)}\langle\phi_n(\boldsymbol{r}-\boldsymbol{R}_i)|[\hat{T}_r+u(\boldsymbol{r}-\boldsymbol{R}_i)+V(\boldsymbol{r})-u(\boldsymbol{r}-\boldsymbol{R}_i)]|\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\rangle \\
&= \epsilon_n + N^{-1}\sum_{i,j} e^{ik(\boldsymbol{R}_j-\boldsymbol{R}_i)}\langle\phi_n(\boldsymbol{r}-\boldsymbol{R}_i)|[V(\boldsymbol{r})-u(\boldsymbol{r}-\boldsymbol{R}_i))]|\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\rangle \\
&= \epsilon_n + \sum_{j} e^{i\boldsymbol{k}\cdot\boldsymbol{R}_j}\langle\phi_n(\boldsymbol{r})|[V(\boldsymbol{r})-u(\boldsymbol{r}))]|\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\rangle. \qquad (2.16)
\end{aligned}$$

From the second line to the third line, we have used the fact that $\hat{T}_r + u(\boldsymbol{r}-\boldsymbol{R}_i)$ is the local Hamiltonian of the site i and ignored the overlapp integral. Then to the fourth line, because of the infinite integration over $\boldsymbol{r}$ is taken, we have shhifted the origin to each $\boldsymbol{R}_i$ for the integration, which results in simply N-times of the result and the normalization is considered. So if we write the difference between the lattice potential $V(\boldsymbol{r})$ and the localized potential $u(\boldsymbol{r})$ as $v(\boldsymbol{r}) \equiv V(\boldsymbol{r}) - u(\boldsymbol{r})$,

$$E_n(\boldsymbol{k}) = \epsilon_n + \langle\phi_n(\boldsymbol{r})|v(\boldsymbol{r})|\phi_n(\boldsymbol{r})\rangle - \sum_{\boldsymbol{R}_j\neq 0} e^{i\boldsymbol{k}\cdot\boldsymbol{R}_j} t_n(\boldsymbol{R}_j), \qquad (2.17)$$

where the **hopping integral** $t_n(\boldsymbol{R}_j)$ is defined as

$$t_n(\boldsymbol{R}_j) \equiv -\langle\phi_n(\boldsymbol{r})|v(\boldsymbol{r})|\phi_n(\boldsymbol{r}-\boldsymbol{R}_j)\rangle. \qquad (2.18)$$

Let us restrict ourselves to one-dimensional systems. For simplicity, we assume the hopping integral has a finite value t only for the nearest neighbor sites. Then the sum in (2.17) is only for $R_j = \pm a$, where a is the lattice constant. The result gives

$$E_n(k) = \epsilon_n - \alpha_n - t(e^{ika} + e^{-ika}) = \epsilon_n - \alpha_n - 2t\cos ka. \qquad (2.19)$$

The cosine band in (2.19) is common for one-dimensional systems. $\alpha_n \equiv -\langle\phi_n(r)|v(r)|\phi_n(r)\rangle$ is the on-site energy shift due to the change from the localized potential to the crystal potential and called **crystal field** contribution.

Figure 2.5 shows the cosine band thus obtained. In the case of NFEA, unperturbed is the band bottom $k = 0$ and the perturbation is stronger with coming close to the edge of Brillouin zone. In the case of TBA, the starting point is the band center. When $t = 0$ the band is flat, and with increasing t it broadens to $\pm 2t$.

At the band bottom, as we know from (2.19), the eigenstate is in the mode of standing wave. As illustrated in one of the insets, the localized wavefunctions on the sites are synchronously summed up wih the same phase at the bottom. On the other hand, they are summed up with opposite phase for neighboring site at the band top to be standing wave with higher energy. These features are the same as those in a double well potenntial with a tunneling matrix element t between them. The bottom corresponds to the **bonding orbital** while the top to the **anti-bonding orbital**.

2.2 Band structure measurement, calculation

The importance of the energy band structure, that is, the energy dispersion relation in the first Brillouin zone, to materials science and engineering cannot be overstated. The views from the two extremes in the previous section are useful for understanding the band concept and for obtaining a general physical perspective on the phenomena that occur there. For quantitative analysis of experimental data or for selection of materials to design devices with desired characteristics, precise band parameters are required.

In the band structure, to analyze the optical response and hot electron characteristics, the dispersion in a wide range of crystal wavenumbers is important. Even in the optical response, it is necessary to know the dispersion of the band edge precisely for the light emission from the band edge, problems such as excitons, and electrical conduction to a low electric field. Let's call the former "global band structure" and the latter "band edge structure". Here are some typical experimental and theoretical methods to obtain for each. Appendix 2A introduces one of the ab-initio calculation methods that calculate without directly relying on the measured values of the experiment.

2.2.1 Measurement of global band structure: angle-resolved photoemission spectroscopy

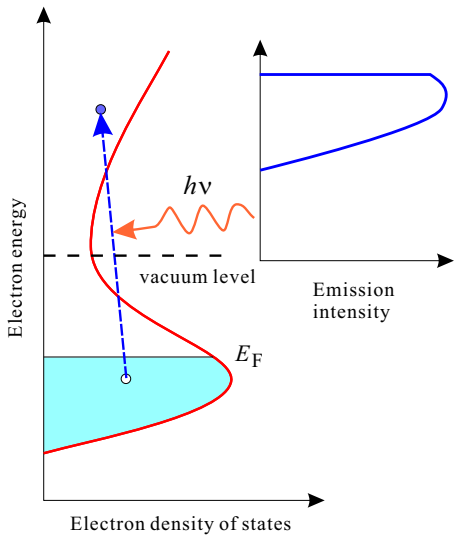
In recent years, the resolution of photoemission spectroscopy has improved remarkably, and angle-resolved photoemission spectroscopy (ARPES) is now a means for directly obtaining a global band structure.

The right figure shows the concept of photoemission spectroscopy. Let E^{ν} be the energy of photo-electron measured from the vacuume level, E_{B} the binding energy of electrons measured from the Fermi level E_{F}, $h\nu$ the energy of incoming light, ϕ the work function of the specimen. Then in the light absorption process, the energy conservation law

$$E^{\nu} = h\nu - \phi - E_{\mathrm{B}} \tag{2.20}$$

holds. In the experiments, $E_{\mathrm{kin}} = \phi + E^{\nu}$ is measured and the relation is simply written as

$$E_{\mathrm{B}} = h\nu - E_{\mathrm{kin}}, \tag{2.21}$$



which gives the binding energy. In the above ordinary photoemission spectroscopy, because the photo-electrons are collected independent of the angles, the total density of states is measured.

In the emission of photo-electron, the sum of the parallel component of crystal wavenumber $\boldsymbol{k}_{\parallel}$ and the recprocal lattice vector $\boldsymbol{G}$ is conserved. Let $\boldsymbol{k}_{\mathrm{i}}$ be the wavenumber of the initial state, $\boldsymbol{k}_{\mathrm{f}}$ the wavenumber of electrons emitted into the vacuum, then

$$(\boldsymbol{k}_{\mathrm{i}} + \boldsymbol{G})_{\parallel} = \boldsymbol{k}_{\mathrm{f}\parallel}. \tag{2.22}$$

The energy conservation reads

$$E_{\mathrm{i}}(\boldsymbol{k}_{\mathrm{i}}) + h\nu = E_{\mathrm{f}}(\boldsymbol{k}_{\mathrm{i}}) = \frac{\hbar^2 k_{\mathrm{f}}^2}{2m_0} + \phi. \tag{2.23}$$

If we know the workfunction ϕ, the dispersion relation of the final states in the crystal $E_{\mathrm{f}}(\boldsymbol{k})$, then we can obtain the dispersion relation $E_{\mathrm{i}}(\boldsymbol{k})$. Since the energy of the final state is high, the effect of lattice potential on it is small, hence

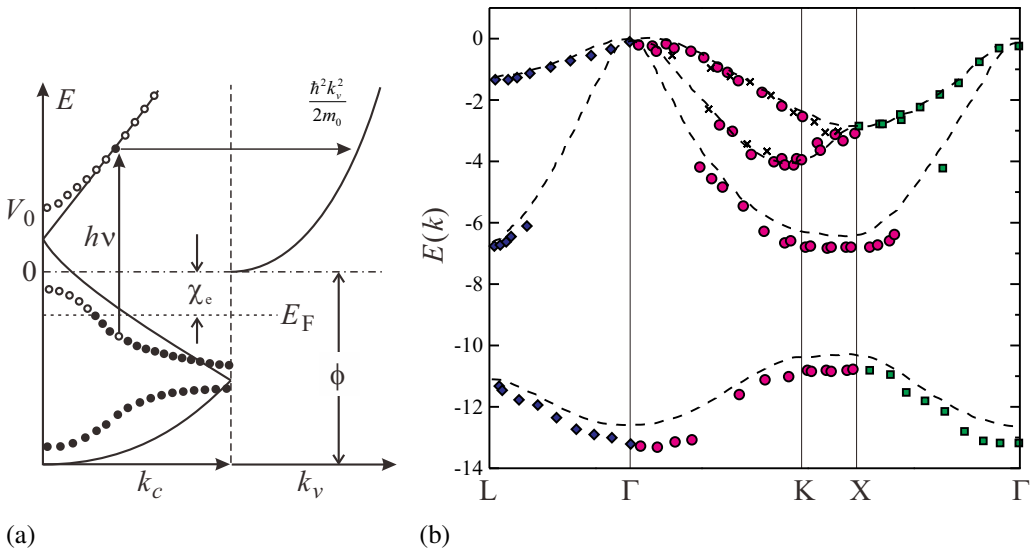


Fig. 2.6 (a) Illustration o the procesure to determine the dispersion relation in ARPES measurement. (b) Band structure obtained for GaAs with the vertical emission method.

$E_{\rm f}(\boldsymbol{k})$ is often replaced with that of free electron. Care should be taken that the wavenumber is still a crystal wavenumber and the energy conservation law (2.23) becomes

$$E_{\rm i}(\boldsymbol{k}_i) + h\nu = \frac{\hbar^2(\boldsymbol{k}_{\rm i}+\boldsymbol{G})^2}{2m_0} + V_0 = \frac{\hbar^2 k_{\rm f}^2}{2m_0} + \phi. \tag{2.24}$$

Here, still the potentail V_0 for the zero-kinetic energy “free electron in the crystal" is not known. But V_0 is also determined for various interpretations of experiments to be consistent.

As a simplest method among ARPES, a possible way to obtain the dispersion relation is to measure photo-electrons emitted veratically to the surface. In this case, as $\boldsymbol{k}_{\parallel} = 0$, the energy conservation gives

$$E_{\rm i}(\boldsymbol{k}_{\rm i}) + h\nu = \frac{\hbar^2|(\boldsymbol{k}_{\rm i}+\boldsymbol{G})_\perp|^2}{2m_0} + V_0 = \frac{\hbar^2 k_{\rm f}^2}{2m_0} + \phi, \tag{2.25}$$

and from

$$|(\boldsymbol{k}_{\rm i}+\boldsymbol{G})_\perp| = \sqrt{\frac{2m}{\hbar^2}\left(\frac{\hbar^2 k_{\rm f}^2}{2m_0} + \phi - V_0\right)}, \tag{2.26}$$

$\boldsymbol{k}_{\rm i}$ and then the dispersion can be obtained.

2.2.2 Global band structure calculation: empirical pseudo potential method

Time-independent part Schrödinger equaiton in a periodic potential $V(\boldsymbol{r}) = V(\boldsymbol{r}+\boldsymbol{R})$($\boldsymbol{R}$ is an arbitrary lattice vector) is written as

$$\mathscr{H}\psi(\boldsymbol{r}) = \left(-\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r})\right)\psi(\boldsymbol{r}) = E\psi(\boldsymbol{r}). \tag{2.27}$$

And we write a solution of the above in the Bloch form with dropping the band index as

$$\psi(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u_{\boldsymbol{k}}(\boldsymbol{r}). \tag{2.28}$$

The lattice periodic part $V(\boldsymbol{r})$, $u_{\boldsymbol{k}}(\boldsymbol{r})$ can be written in the Fourier expansion with the reciprocal lattice vector $\boldsymbol{G}$ as

$$V(\boldsymbol{r}) = \sum_{\boldsymbol{G}} V_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}, \quad u_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{G}} C_{\boldsymbol{G}} e^{i\boldsymbol{G}\cdot\boldsymbol{r}}. \tag{2.29}$$

Substituting (2.28) and (2.29) into (2.27), we obtain

$$\sum_{\boldsymbol{G}} \left[\left\{ \frac{\hbar^2}{2m}(\boldsymbol{k}+\boldsymbol{G})^2 - E) \right\} C_{\boldsymbol{G}} + \sum_{\boldsymbol{G}'} V_{\boldsymbol{G}-\boldsymbol{G}'} C_{\boldsymbol{G}'} \right] e^{i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}} = 0.$$

Because each term in the summation of $\boldsymbol{G}$ should be zero, we obtain simultaneous equations for $\{C_{\boldsymbol{G}}\}$ as

$$\sum_{\boldsymbol{G}'} \left[\left\{ \frac{\hbar^2}{2m}(\boldsymbol{k}+\boldsymbol{G})^2 - E \right\} \delta_{\boldsymbol{G}\boldsymbol{G}'} + V_{\boldsymbol{G}-\boldsymbol{G}'} \right] C_{\boldsymbol{G}'} = 0, \tag{2.30}$$

and the condition for the existence of non-trivial solution is

$$\left| \left[\left\{ \frac{\hbar^2}{2m}(\boldsymbol{k}+\boldsymbol{G})^2 - E \right\} \delta_{\boldsymbol{G}\boldsymbol{G}'} + V_{\boldsymbol{G}-\boldsymbol{G}'} \right]_{\boldsymbol{G}\boldsymbol{G}'} \right| = 0. \tag{2.31}$$

If we can perform perfect expansion of (2.30) in an actual crystal, the solution of the secular equation (2.31) gives the accurate band structure $E(\boldsymbol{k})$. Equation (2.31) tells that requirements for this calculation are the coefficient $V_{\boldsymbol{G}}$ of Fourier expansion of periodic lattice potential.

In **pseudo potential method** we calculate “effective" $V_{\boldsymbol{G}}$ under the following concepts.

(1) Structures of valence band and conduction band below and above the Fermi level determine the properties of semiconductors. The outermost electrons of consisting atoms are forming these bands. The inner electrons are strongly bound around nuclei and can be included into the periodic crystal potential. Hence we apply the above secular equation only for the outermost electrons.

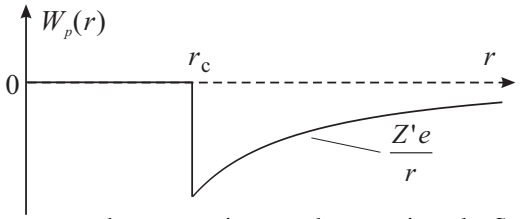
(2) (Characteristic for pseudo potential method) In the vicinity of nuclei, $V(\underline{\mathrm{r}})$ can be approximated with r being the distance from the nucleons and Z being the atomic number as

$$V(r) = Ze/r,$$

and the outermost electron wavefunction should have stronger space modulation in the amplitude. On the other hand, far from the nucleus, the inner electrons (let the number be Z_c) screens the potential and the effective atomic number decreases to $Z' = Z - Z_c$. Furthermore if we take into account the electron-electron mutual interaction in, *e.g.*, mean field approximation, the screening gives faster decay of potential than r^{-1} resulting in much weaker spatial modulation of the wavefunction. As we have seen in the tight-binding approximation (Sec.2.4), the band structure is dominated with the overlapping of wavefunctions in neighboring sites, that is, this weak potential part determines the band structure in practice.

If we perform the Fourier expansion of $V(\boldsymbol{r})$ itself, strong spatial modulation around the nuclei introduces lots of high frequency components, which are nothing to do with the band structure. Such high frequency coefficients $V_{\boldsymbol{G}}$ not only introduce useless calculations but also make it difficult to solve the secular equation eq.(2.31).

The above consideration brings about the central concept of pseudo potential method. We look for a “pseudo potential" which simplifies the wave function around the nucleus but reproduces the tailing part of the wavefunction. Then obtain $V_{\boldsymbol{G}}$ for this pseudo potential and solve the secular equation (2.31).



The simplest example can be as in the left figure

$$W_p(r) = 0 \quad (r < r_c), \quad W_p(r) = Z'e/r \quad (r \geq r_c). \tag{2.32}$$

Taking r_c to an appropriate value, we can approximately reproduce the tail of wavefunctions, keeping the eigen energies. Because the potential around the ion core is flat, a pseudo potential with small wavenumber expansion can be constituted. Summing up W_p on the positions of unit cells $\boldsymbol{R}_j$, we obtain a pseudo potential for the crystal potential as

$$V_p(\boldsymbol{r}) = \sum_{j,\alpha} W_p^{\alpha}(\boldsymbol{r} - \boldsymbol{R}_j - \boldsymbol{\tau}_{\alpha}), \tag{2.33}$$

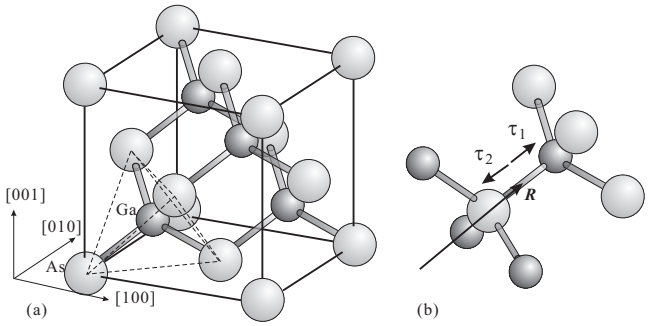


Fig. 2.7 (a) Unit cell of a zinc blende crystal (GaAs). Alternate occupation of lattice points in diamond structure with group III (Ga) atoms and group IV (As) atoms. In the tetrahedron indicated by broken lines, apices are occupied with As and the center with Ga. (b) A primitive cell contains a Ga an an As atoms. Let the edge length of the unit cell in (a) a, and take the lattice point at the center between Ga and As then their coordinates are from the lattice point $(a/8)(1,1,1)$ and $(-a/8)(1,1,1)$ respectively.

where α is the index of atomic positions in the unit cell and $\boldsymbol{\tau}_\alpha$ are relative positions of constituting atoms from a certain point in the unit cell.

Because potential (2.33) has the lattice periodicity, it can be Fourier expanded with wavenumber of reciprocal lattice points $\boldsymbol{K}$.

$$
\begin{aligned}
v_p(\boldsymbol{K}) &= \int \sum_{j,\alpha} W_p^\alpha(\boldsymbol{r}-\boldsymbol{R}_j-\boldsymbol{\tau}_\alpha) e^{-i\boldsymbol{K}\cdot\boldsymbol{r}} \frac{d\boldsymbol{r}}{V} \\
&\quad \text{Let } \boldsymbol{r}' \equiv \boldsymbol{r}-\boldsymbol{R}_j-\boldsymbol{\tau}_\alpha, \quad N : \text{number of unit cells}, \quad \Omega : \text{unit cell volume}, \quad \text{From } e^{-i\boldsymbol{K}\cdot\boldsymbol{R}_j}=1 \\
&= \frac{1}{N}\sum_j e^{-i\boldsymbol{K}\cdot\boldsymbol{R}_j} \sum_\alpha e^{-i\boldsymbol{K}\cdot\boldsymbol{\tau}_\alpha} \frac{1}{\Omega}\int_\Omega W_p^\alpha(\boldsymbol{r}') e^{-i\boldsymbol{K}\cdot\boldsymbol{r}'} d\boldsymbol{r} = \sum_\alpha e^{-i\boldsymbol{K}\cdot\boldsymbol{\tau}_\alpha} \frac{1}{\Omega}\int_\Omega W_p^\alpha(\boldsymbol{r}') e^{-i\boldsymbol{K}\cdot\boldsymbol{r}'} d\boldsymbol{r}', \\
&= \sum_\alpha e^{-i\boldsymbol{K}\cdot\boldsymbol{\tau}_\alpha} w_p^\alpha(\boldsymbol{K}).
\end{aligned} \tag{2.34}
$$

$w_p^\alpha(\boldsymbol{K})$ is the Fourier transform of (2.32), and depends on the atomic species α, that is the strength and functional form of nuclear potential, not on the crystal structure, and called **form factor**. On the other hand $e^{-i\boldsymbol{K}\cdot\boldsymbol{\tau}_\alpha}$ depends only on the crystal structure and called **structure factor**. This separation of factors make it possible to estimate band structure based on the analogies between the crystals.

In the case of zinc blende structure, from Fig. 2.7(b), we can write $\boldsymbol{\tau}_1 = -a(1/8,1/8,1/8) = -\boldsymbol{\tau}_2 \equiv \boldsymbol{\tau}$. Then (2.34) is written as

$$
\begin{aligned}
v_p(\boldsymbol{K}) &= e^{i\boldsymbol{K}\cdot\boldsymbol{\tau}_1} v_p^1(\boldsymbol{K}) + e^{-i\boldsymbol{K}\cdot\boldsymbol{\tau}_1} v_p^2(\boldsymbol{K}) = (v_p^1+v_p^2)\cos\boldsymbol{K}\cdot\boldsymbol{\tau} + (v_p^1-v_p^2)\sin\boldsymbol{K}\cdot\boldsymbol{\tau} \\
&= v_p^s(\boldsymbol{K})\cos\boldsymbol{K}\cdot\boldsymbol{\tau} + v_p^a(\boldsymbol{K})\sin\boldsymbol{K}\cdot\boldsymbol{\tau}.
\end{aligned} \tag{2.35}
$$

Here v_p^s and v_p^a are symmetric and anti-symmetric part of the form factor v_p, sin and cosine functional part is the structure factor. In the case of diamond structure (like Si or Ge), from the symmetry $v_p^1 = v_p^2$ and $v_p^a = 0$.

To obtain the value of form factor we need detailed functional form of the pseudo potential. Here comes the idea of "empirical" pseudo potential method, in which we rather *determine the form factors from experiments* as fitting parameters than to calculate them deductively from the specific form of pseudo potential.

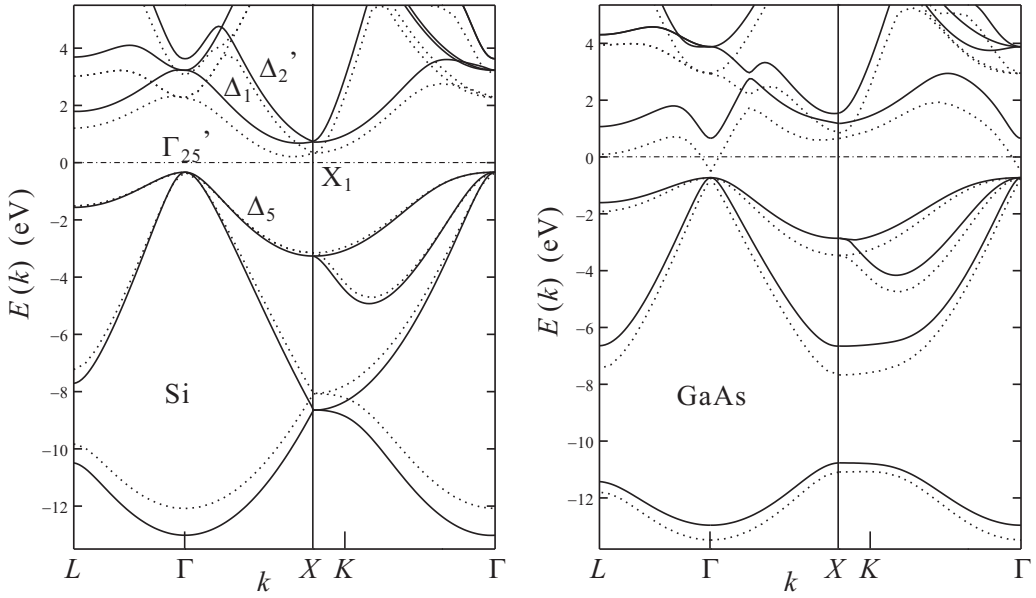
2.2.2.1 Global band structures of semiconductors with diamond and zinc blende structures

In the case of diamond structure $v_p^a(\boldsymbol{K}) = 0$, and because of the original concept of pseudo potential, we only need to calculate $v_p^s(\boldsymbol{K})$ for reciprocal lattice points with small $|\boldsymbol{K}|$. Here we restrict $|\boldsymbol{K}| \le \sqrt{11}$, then $a\boldsymbol{K}/2\pi = (000)$, (111), (200), (220), (310) and their reversed points, 51 in total as in page **??** (for a while we drop the commas between the vector components. Accordingly the size of the matrix in the left hand side in (2.30) is 51×51.

The potential for $|\boldsymbol{K}| = 0$ just shifts the energy and is set to zero. And from the above approximation, the terms for $|\boldsymbol{K}| > \sqrt{11}$ are also dropped. The atomic potential is as in (2.32), supposed to have spherical symmetry. The form factors, which are the Fourier transform fo the potential, should also be a function of the absolute value of wavenumber. We thus

	$v_p^s(111)$	$v_p^s(220)$	$v_p^s(311)$	$v_p^a(111)$	$v_p^a(200)$	$v_p^a(311)$
Si	−2.856	0.544	1.088	0	0	0
Ge	−3.128	0.136	0.816	0	0	0
GaAs	−3.128	0.136	0.816	0.952	0.68	0.136
CdTe	−2.72	0	0.544	2.04	1.224	0.544

Tab. 2.3 Empirically obtained form factors of pseudo potential from optical reflection coefficients for representative fcc semiconductors in unit of eV. The values are taken from M L. Cohen and T. K. Bergstresser, Phys. Rev. **141**, 789 (1966).



(a) Si (b) GaAs

Fig. 2.8 Band structures (solid lines)of (a) Si and (b) GaAs, obtained from the form factors in Tab.2.3. Broken lines show the results of linear muffin-tin orbital (LMTO) method, which is one of the simplest "first principles calculations". The two results are shifted to have the same value at the top of valence band. Energy zero $E(k) = 0$ is taken to the Fermi energy in the pseudo potential calculations.

only need to know the form factors for (111), (200), (220), (311) below the distance $\sqrt{11}$. Even among them for (200) the structure factor $\cos \boldsymbol{K}\cdot\boldsymbol{\tau}$ is zero and we do not need to know the value or can put it as zero. And the last three are required. Then form factors are determined to fit to experimentally measured quantities. A way for further decreasing of the number of parameters is to determine r_c in (2.32) to obtain $v_p^s(\boldsymbol{K})$ and perform the iteration to explain the experiments. Table 2.3 shows three parameters $v_p^s(\boldsymbol{K})$ chosen as to fit to the optical reflection coefficients in experiments for representative diamond and zinc blende structure semiconductors.

Procedures of pseudo potential calculation for zinc blende semiconductors are similar to the above though $v_p^a(\boldsymbol{K})$ is now finite. In the case of GaAs in Tab.2.3, since Ga and As locate in the both sides of Ge in the periodic table, the value of Ge is also adopted for $v_p^s(\boldsymbol{K})$. From (2.35), the anti-symmetric term is proportional to $\sin \boldsymbol{K}\cdot\boldsymbol{\tau}$, there is thus no contribution from (220), and those from (110), (200), (311) should be considered. Table 2.3 shows the results to reproduce optical measurements. Similarly in the case of a II-VI semiconductor CdTe, the value of Sn (gray tin) for $v_p^s(\boldsymbol{K})$ is adopted and others are obtained from experiments.

Having the values of $v_p(\boldsymbol{K})$, we substitute them into (2.31) and solve the eigenvalue problem of the 51×51 matrix and obtain $E(\boldsymbol{k})$. Global band structures thus obtained are shown in Fig. 2.8.

Because the above calculations do not take care of the spin-orbit interaction, which actually has important contributions

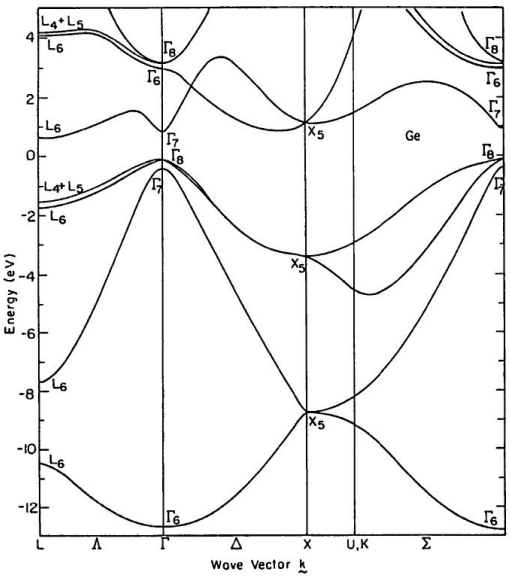


Fig. 2.9 Global band structure of Ge calculated with empirical pseudo potential method. The spin-orbit interaction is taken into account and the top of the valence band shows spin-orbit splitting. From ref.[2].

to the band structure, in particular at the top of valence band, the three bands are degenerated. In pseudo potential calculations which take into account the spin-orbit contribution, one of the three branches goes down. In the results for Ge shown in Fig. 2.9, clear spin-orbit splitting at the top of the valence band is observed.

Above obtained band structures of Si, GaAs, Ge (representative diamond, zinc blende type semiconductors) are shown in Fig. 2.8 and Fig. 2.9. The bottoms of conduction band of Si are close to X-points though a bit inside the first Brillouin zone, while it is at Γ-point in GaAs and they are at L-points in Ge. Schematic drawing of equipotential surfaces are thus shown in Fig. 2.10, which are probably familiar to the readers. From Fig. 2.8 and Fig. 2.9, we see that the expressions in Fig. 2.10 are a bit exaggerated.

As can be seen above, Si has six equivalent bottoms in the first Brillouin zone, which are called **valleys**. In metallic doped n-type samples, the number of Fermi surfaces is that of valleys, to which we should pay attention in performing, *e.g.*, some integration over the Fermi surfaces. GaAs has a single valley at $Gamma$-point and the effective mass is almost isotropic. Ge has valleys at L-points and there are 8 equivalent L-points just at the Brillouin zone boundaries. Hence each valley is divided by the neighboring zones and the effective valley number is 4.

2.2.3 Band structure at band edges: Effective mass

Whan a Bloch type electron wavefunction $\psi_{n\boldsymbol{k}}(\boldsymbol{r})$ has a dispersion relation $E_n(\boldsymbol{k})$, the group velocity is written as

$$\boldsymbol{v}_n(\boldsymbol{k}) = \hbar^{-1}\nabla_{\boldsymbol{k}} E_n(\boldsymbol{k}). \tag{2.36}$$

Hence, the acceleration is given as

$$\frac{d\boldsymbol{v}_n}{dt} = \frac{d\boldsymbol{k}}{\hbar dt}\cdot\nabla_{\boldsymbol{k}}(\nabla_{\boldsymbol{k}} E_n(\boldsymbol{k})) = \frac{\nabla_{\boldsymbol{k}}}{\hbar^2}\sum_{j=x,y,z}\frac{\partial E_n(\boldsymbol{k})}{\partial k_j}F_j. \tag{2.37}$$

Here, $\boldsymbol{F} = d\boldsymbol{p}/dt = \hbar d\boldsymbol{k}/dt$ is a vector of a "force". Now we define the **inverse effective mass tensor** $1/m^*$, which is the inverse matrix of the **effective mass tensor** with

$$\left(\frac{1}{m^*}\right)_{ij} \equiv \frac{1}{\hbar^2}\frac{\partial^2 E(\boldsymbol{k})}{\partial k_i \partial k_j}. \tag{2.38}$$

Then (2.37) can be re-written as

$$\frac{dv_i(\boldsymbol{k})}{dt} = \sum_j \left(\frac{1}{m^*}\right)_{ij} F_j, \tag{2.39}$$

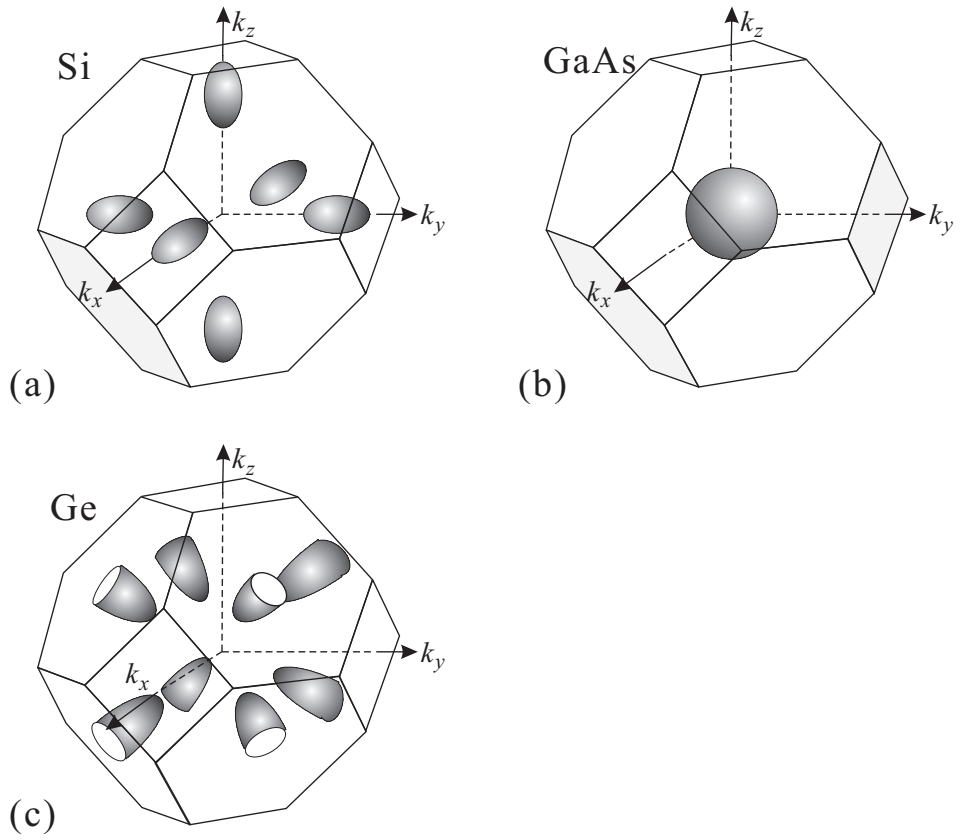


Fig. 2.10 Schematic drawings of equipotential surfaces for (a) Si, (b) GaAs, (c) Ge conduction band valleys from the information obtained in the pseudo potential calculation of Fig. 2.9. In the case of Ge, which has the valleys at L-points, because the region is limited to the first Brillouin zone, the boundaries cut the centers of the spheroidal valleys.

which is equivalent to

$$F_i = \sum_j m^*_{ij} \frac{dv_j(\boldsymbol{k})}{dt}. \tag{2.40}$$

For simplicity, we consider an energy band with an isotropic energy dispersion $E(k) = ak^2$. m^*, in general is a tensor, becomes a scalar $\hbar^2/(\partial^2 E(k)/\partial k^2) = \hbar^2/2a$. Let it be more specific. Consider the case of eq. (1.9), where a gap opens up in the nearly free electron approximation (NFEA). Around $\Delta k \sim 0$,

$$E_\pm \approx \epsilon_z \pm V_0 \left[1 + \frac{\epsilon_z}{2V_0}\left(\frac{\Delta k}{k_g}\right)^2\right], \quad k_g \equiv \frac{\sqrt{2m_0 V_0}}{\hbar}, \tag{2.41}$$

which reads to the effective mass of

$$m^* = \pm\frac{\hbar^2}{2}\frac{2V_0}{\epsilon_z}\frac{2m_0}{\hbar^2} = \pm\frac{2V_0}{\epsilon_z}m_0 = \pm\frac{\epsilon_g}{\epsilon_z}m_0. \tag{2.42}$$

Here ϵ_z is the band width, ϵ_g the band gap. In this naive approximation, the ratio between the band width and the band gap determines the effective mass, namely, the wider the energy gap in comparison with the band width, the heavier the effective mass. This is a kind of “toy model" but can predict at least some trend in the effective masses in the same type of energy bands, *i.e.*, withe the same symmetry at the same point in the reciprocal lattices. For example we can see such tendency in the effective masses at Γ point of conduction band in GaAs, InP, InAs.

2.2.4 Measurement of band-edge structure: cyclotron resonance

The information given from ARPES introduced in Sec. 2.2.1 is limited to the region below $E_{\rm F}$ as we can see from the principle. And the precision is, at present, not enough for quatitative discussion on electric conduction or optical actions. Cyclotron resonance has long been used as a means of experimentally obtaining information on the band-edge of the conduction band and valence band. This method is based on the fact that the motion of the particle with charge q and mass m in the magnetic field with flux density B projected to the plane perpendicular to the magnetic field is a circular motion with cyclotron frequency

$$\omega_{\rm c} = \frac{qB}{m}. \tag{2.43}$$

Let's consider in classical approximation. The motion of equation for the particle with charge q and effective mass tensor $\overleftrightarrow{m}$ is

$$\overleftrightarrow{m}\frac{d\boldsymbol{v}}{dt} + \frac{\overleftrightarrow{m}\boldsymbol{v}}{\tau} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{2.44}$$

The electric field of microwave oscillates as $\boldsymbol{E}e^{-i\omega t}$ and the velocity as $\boldsymbol{v}e^{-i\omega t}$ then

$$\left(-i\omega + \frac{1}{\tau}\right)\overleftrightarrow{m}\boldsymbol{v} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{2.45}$$

Once we assume the oscillatio in $\boldsymbol{v}$ as above, the resonance condition does not depend on $\boldsymbol{E}$ and we put $\boldsymbol{E} = 0$ for simplicity. For simpler expression we write $\omega' = \omega + i/\tau$, $\boldsymbol{B} = B(\alpha, \beta, \gamma)$, and put

$$\overleftrightarrow{m}^{-1} = \begin{pmatrix} m_1^{-1} & 0 & \\ 0 & m_2^{-1} & \\ 0 & 0 & m_3^{-1} \end{pmatrix}. \tag{2.46}$$

When the lattice system is cubic or rhombic this is justified. Then the equation of motion is written as

$$\begin{aligned} i\omega' m_1 v_x + q(v_y B_z - v_z B_y) &= 0, \\ i\omega' m_2 v_y + q(v_z B_x - v_x B_z) &= 0, \\ i\omega' m_3 v_z + q(v_x B_y - v_y B_x) &= 0. \end{aligned}$$

For the above to have non-trivial solution,

$$\begin{vmatrix} i\omega' m_1 & qB\gamma & -qB\beta \\ -qB\gamma & i\omega' m_2 & qB\alpha \\ qB\beta & -qB\alpha & i\omega' m_3 \end{vmatrix} = 0. \tag{2.47}$$

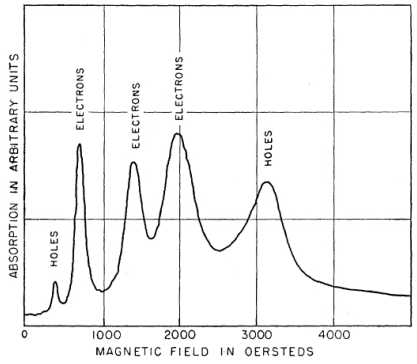
From the condition $\omega_c\tau \gg 1$,

$$\omega_c = \frac{qB}{m_c} = qB\sqrt{\frac{m_1\alpha^2 + m_2\beta^3 + m_3\gamma^2}{m_1 m_2 m_3}}. \tag{2.48}$$

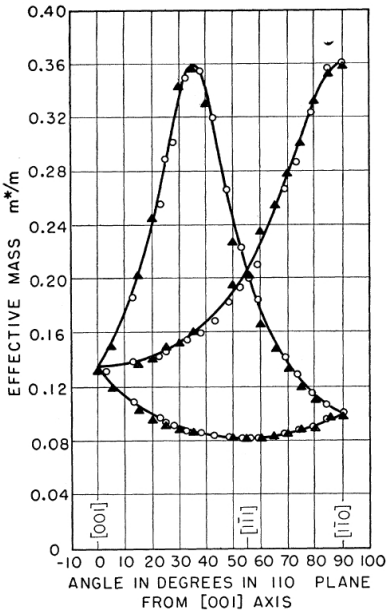
m_c is obtained from the experiment with eq.(2.43) and called **cyclotron mass**.

In experiment, the condition $\omega_c\tau \gg 1$ should hold and for that the impurities in the specimen should be very small. At low temperatures the carrier concentration discussed later is very low, and carriers are excited by light illumination. In Fig. 2.11(a), we show an example of thus obtained cyclotron resonance in the absorption of 24 GHz microwave. Several peaks are observed and assigned to electron (excitation in conduction band) and to hole (excitation in valence band). While the electrons have negative charge, the holes have positive. Hence if we use circular polarized microwave, the absorption intensity changes according to the direction of rotation. The assignment was done as above.

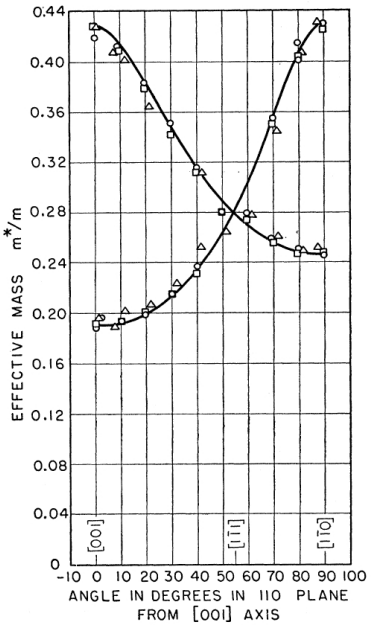
In the practical analysis, the symmetry of the crystals tells that the constant energy surfaces should be spheroids. We thus take the spatial coordinate to the main axes of a spheroid. Then the effective mass is represented by a tensor with



(a)



(b)



(c)

Fig. 2.11 (a) Example of microwave absorption signal in cyclotron resonance. (b) Magnetic field angular dependence of cyclotron mass in Ge conduction band. (c) The same for Si. From [1].

diagonal terms:m_l along the main axis and two m_t along the other two axes. Let θ be the angle between the magnetic field and the main axis, (2.48) gives

$$\left(\frac{1}{m_{\rm c}}\right)^2 = \frac{\cos^2\theta}{m_t^2} + \frac{\sin^2\theta}{m_t m_l}. \tag{2.49}$$

The magnetic field angular dependences of cyclotron mass are displayed for the conduction bands in Ge and in Si in Fig. 2.11(b) and (c) respectively. From the analysis, we can get the directions of spheroids (and the number from the crystal symmetry) and the values of m_t and m_l.

2.2.5 Band-edge structure calculation: k·p perturbation

k·p perturbation is an adequate method to obtain highly accurate band structures around band edges. Though in empirical pseudo potential method we can reproduce band structure from very few parameters with comparatively simple calculation, in k·p perturbation we need to increase the number of bands included in the calculation, which increases the dimension of matrices and large scale calculation is required.

The basics of k·p perturbation is eq.(1.4) in the first hour of this lecture. Substituting Bloch function $e^{i\boldsymbol{kr}}u_{n\boldsymbol{k}}(\boldsymbol{r})$ into the original Schrödinger equation, we obtain the equation for the lattice periodic part $u_{n\boldsymbol{k}}(\boldsymbol{r})$. In three dimensional space (1.4) can be written as

$$\left[-\frac{\hbar^2\boldsymbol{\nabla}^2}{2m_0} + V(\boldsymbol{r}) + \frac{\hbar^2\boldsymbol{k}^2}{2m_0} - i\frac{\hbar^2}{m_0}\boldsymbol{k}\cdot\boldsymbol{\nabla}\right] u_{n\boldsymbol{k}}(\boldsymbol{r}) = E_{n\boldsymbol{k}}u_{n\boldsymbol{k}}(\boldsymbol{r}). \tag{2.50}$$

Here the Bloch wavenumber $\boldsymbol{k}$ is a parameter (c-number) and not an operator.

Now we redefine the unperturbed Hamiltonian $\mathscr{H}_0$ and $\boldsymbol{k}$-dependent eigen energy as

$$\mathscr{H}_0 \equiv -\frac{\hbar^2\boldsymbol{\nabla}^2}{2m_0} + V(\boldsymbol{r}), \quad E'_n(\boldsymbol{k}) = E_{n\boldsymbol{k}} - \frac{\hbar^2\boldsymbol{k}^2}{2m_0}.$$

Then the perturbation term can be written as

$$\mathscr{H}'(\boldsymbol{k}) = -i\frac{\hbar^2}{m_0}\boldsymbol{k}\cdot\boldsymbol{\nabla} = \frac{\hbar}{m_0}\boldsymbol{k}\cdot\hat{\boldsymbol{p}}, \tag{2.51}$$

from which we can easily guess the source of the naming "k· p".

(2.51) is zero for $\boldsymbol{k}=0$. Hence we set the unperturbed state as $\boldsymbol{k}=0$. Assume that we obtain the exact eigenstates $\{u_{j0}(\boldsymbol{r})\}$ for $\boldsymbol{k}=0$, then they form a complete set and an eigenstate for finite $\boldsymbol{k}$ can be expanded as

$$u_{n\boldsymbol{k}}(\boldsymbol{r})=\sum_{j=0}^{\infty}c_{nj}(\boldsymbol{k})u_{j0}(\boldsymbol{r}).$$

These $c_{nj}(\boldsymbol{k})$ can be obtained from the perturbation of $\mathscr{H}'(\boldsymbol{k})$. This is the concept of k·p perturbation.

In the above $\boldsymbol{k}=0$ is taken to the unperturbed point assuming some avoided level crossing due to the high symmetry. Such avoided level crossing results in $\partial E/\partial k=0$ and the approximation is practically good around the point because $\mathscr{H}'(\boldsymbol{k})$ is small around it. Similar may occur in other points with high symmetry and k·p expansions around such points are also available. Also as we have seen so far, physical properties of semiconductors are determined with band structures around such symmetric points. We cannot perform, of course, the infinite summation hence cut the summation around the band n which is under consideration. The accuracy of the k·p perturbation usually determined by the number of bands taken into account.

(a) the case of non-degenerate $u_{i0}(\boldsymbol{r})$

$$u_{i\boldsymbol{k}}(\boldsymbol{r})=u_{i0}(\boldsymbol{r})+\sum_{j\neq i}\frac{\langle j|\mathscr{H}'|i\rangle}{E_i-E_j}u_{i0}(\boldsymbol{r}),\quad E_i(\boldsymbol{k})=E_i(0)+\langle i|\mathscr{H}'|i\rangle+\sum_{j\neq i}\frac{|\langle i|\mathscr{H}'|j\rangle|^2}{E_i-E_j}\tag{2.52}$$

are obtained as the first order perturbation. Here we have used abbreviation $|i\rangle$ for $|u_{i0}(\boldsymbol{r})\rangle$. From $\langle i|j\rangle=\delta_{ij}$ and $\langle i|\nabla|i\rangle=0$,

$$E_i(\boldsymbol{k})=E_i(0)+\frac{\hbar^2\boldsymbol{k}^2}{2m_0}-\frac{\hbar^4}{m_0^2}\sum_{j\neq i}\frac{\langle i|\boldsymbol{k}\cdot\boldsymbol{\nabla}|j\rangle\langle j|\boldsymbol{k}\cdot\boldsymbol{\nabla}|i\rangle}{E_i-E_j}.\tag{2.53}$$

(b) the case $u_{i0}(\boldsymbol{r})$ has degeneracy

When $u_{00}(\boldsymbol{r})$ has n-fold degeneracy, we take an orthogonal basis $\{u_{00}^j(\boldsymbol{r})\}\quad(j=1,\cdots,n)$ and write the functions in short form as $|0j\rangle$. Perturbed wavefunction is approximated with the linear combination $|u_{0\boldsymbol{k}}^i\rangle=\sum_{j=1}^{n}A_{ij}(\boldsymbol{k})|0j\rangle$. Substituting this into (2.50) gives $[\mathscr{H}_0+\mathscr{H}'-E_0(\boldsymbol{k})]u_{0\boldsymbol{k}}=0$. With taking inner product with $|0i\rangle$, equation

$$\sum_{j=1}^{n}A_{ij}(\boldsymbol{k})[\langle 0i|\mathscr{H}_0|0j\rangle+\langle 0i|\mathscr{H}_0'|0j\rangle-\langle 0i|E_0(\boldsymbol{k})|0j\rangle]$$
$$=\sum_{j=1}^{n}A_{ij}(\boldsymbol{k})[\langle 0i|\mathscr{H}'|0j\rangle+(E_0-E_0(\boldsymbol{k}))\delta_{ij}]=0\tag{2.54}$$

is obtained. The secular equation for this simultaneous equation to have non-trivial solution is

$$|\langle 0i|\mathscr{H}'|0j\rangle+(E_0-E_0(\boldsymbol{k}))\delta_{ij}|=0,\tag{2.55}$$

which gives the dispersion relation $E_0(\boldsymbol{k})$. From the solution $A_{ij}(\boldsymbol{k})$, we obtain approximate set of eigenfunctions corresponding to $\boldsymbol{k}$.

2.2.6 Spin-orbit interaction

For rigorous derivation of **spin-orbit interaction** we should go back to Dirac equation, for which we do not have enough time unfortunately. Here without any derivation, we adopt the Hamiltonian for spin-orbit interaction as

$$\mathscr{H}_{\mathrm{so}}=-\frac{\hbar}{4m_0^2c^2}\boldsymbol{\sigma}\cdot\boldsymbol{p}\times(\nabla V).\tag{2.56}$$

And just discuss the effect on the band structure. $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ is a vector which has Pauli matrices:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{2.57}$$

as its elements. (2.56) is added to (2.50) and with (1.5), we obtain

$$\left[\frac{p^2}{2m_0} + V + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar}{m_0}\boldsymbol{k}\cdot\boldsymbol{\pi} + \frac{\hbar}{4m_0^2c^2}\boldsymbol{p}\cdot\boldsymbol{\sigma}\times\nabla V\right]|n\boldsymbol{k}\rangle = E_n(\boldsymbol{k})|n\boldsymbol{k}\rangle,$$
$$\boldsymbol{\pi} \equiv \boldsymbol{p} + \frac{\hbar}{4mc^2}\boldsymbol{\sigma}\times\nabla V. \tag{2.58}$$

We expand the solution, again with the basis of the band bottom $|\nu 0\rangle$. This time we need to take care of spin freedom and write $|\nu,\sigma\rangle \equiv |\nu 0\rangle \otimes |\sigma\rangle$, and expand as $|n\boldsymbol{k}\rangle = \sum_{\nu',\sigma'} c_{n,\nu\sigma}|\nu',\sigma'\rangle$. With taking inner product with $\langle\nu,\sigma|$, we obtain the eigen equation as

$$\sum_{\nu',\sigma'}\left\{\left[E_{\nu'}(0) + \frac{\hbar^2k^2}{2m}\right]\delta_{\nu\nu'}\delta_{\sigma\sigma'} + \frac{\hbar}{m}\boldsymbol{k}\cdot\boldsymbol{P}^{\nu\nu'}_{\sigma\sigma'} + \Delta^{\nu\nu'}_{\sigma\sigma'}\right\}c_{n\nu'\sigma'} = E_n(\boldsymbol{k})c_{n\nu\sigma}. \tag{2.59}$$

where

$$\boldsymbol{P}^{\nu\nu'}_{\sigma\sigma'} \equiv \langle\nu\sigma|\boldsymbol{\pi}|\nu'\sigma'\rangle, \quad \Delta^{\nu\nu'}_{\sigma\sigma'} \equiv \frac{\hbar^2}{4m^2c^2}\langle\nu\sigma|[\boldsymbol{p}\cdot\sigma\times(\nabla V)]|\nu'\sigma'\rangle. \tag{2.60}$$

The dispersion relation again is obtained with solving the eigen value problem. In this stage, it is often a good approximation to drop the spin-orbit part. In such cases it can be written as $\boldsymbol{\pi} = \boldsymbol{p}$, $\boldsymbol{P}^{\nu\nu'}_{\sigma\sigma'} = \delta_{\sigma\sigma'}\boldsymbol{P}_{\nu\nu'}$.

2.2.7 Wavefunctions at Γ-point in fcc semiconductors

In empirical $k\cdot p$ method, without detailed knowledge of wavefunctions, the parameters required for the band calculation are extracted theoretically and the values are obtained from experiments. Many of such parameters are zero around highly symmetrical points making the calculation easier. Hence the knowledges of spatial symmetries in crystal and in atomic orbitals are important. Though the theory of space group gives systematic discussion to this problem, again due to the time limitation, we restrict ourselves to the discussion around Γ-point in fcc semiconductors.

Bravais lattice is fcc for group IV semiconductors with diamond structure and group III-V semiconductors with zinc blende structure. Here we name them “DZB" semiconductors. As is guessed from the structure in Fig.5.10(b), in chemical bond theory the crystal formation can be understood along covalent bonding between neighboring sp^3 hybrid orbitals. In the group III-V semiconductors, for each atom to form sp^3 hybrid, it needs to be ionized. Hence the crystals are also formed with the ionic bonding. The most effective atomic orbitals on the band structures in these semiconductors are s and p. In DZB structure, there are two atoms per a single lattice point in the simplest fcc structure (Fig.5.10). In Fig.5.9, substituting $2 \times 4 = 8$ into Z, we see that the energy gap opens around the degeneracy points in the distance around $\sqrt{3}$ from Γ point.

We consider a function $|S\rangle$, which has the lattice translational symmetry though also has the same angular symmetry as s orbital in the vicinities of nuclei. For that, we first take a linear combination of atomic orbitals (LCAO) of s orbital $|s\rangle$

$$|u_s\rangle = \sum_{i,\beta} a_{i\beta}|s_{i\beta}\rangle,$$

where i is the index of unit cells, β is the relative index of atoms in a unit cell (as is in the pseudo potential calculation). Though the above function satisfies the crystal translational symmetry, it is not a solution for the Schrödinger equation with lattice potential. Hence we assume that we can modify the form of $|s\rangle$ to make the linear combined function a

solution for the Schrödinger equation with keeping the rotational symmetry in $|s\rangle$ characteristic to the s-orbital. We write thus obtained LCAO wavefunction as $|S\rangle$, which must satisfy

$$\mathscr{H}_0|S\rangle = \left[-\frac{\hbar^2\boldsymbol{\nabla}^2}{2m_0} + V(\boldsymbol{r})\right]|S\rangle = E_c|S\rangle. \tag{2.61}$$

In the same way we define $|X\rangle$, $|Y\rangle$, $|Z\rangle$, which have angular symmetries of p_x, p_y, p_z respectively around nuclei, translational symmetry at the same time.

At Γ-point, the bottom of conduction band is mostly made from s orbitals while the top p orbitals. Hence, though the approximation is rough, we assume the above defined functions satisfy the unperturbed ($\boldsymbol{k} = 0$) equation

$$\mathscr{H}_0|\zeta\rangle = \left[-\frac{\hbar^2\boldsymbol{\nabla}^2}{2m_0} + V(\boldsymbol{r})\right]|\zeta\rangle = E_b|\zeta\rangle, \tag{2.62}$$

where $\zeta \in \{S, X, Y, Z\}$, E_b is E_c for $\zeta = S$ and E_v for others. It may be a problem whether such functions as $|S\rangle$, $|X\rangle$, $\cdots$ actually exist. The space group theory says we can adopt lattice periodic functions with the same angular symmetries as s or p_α orbitals around the point at which parabola with bottoms at $(\pm1, \pm1, \pm1)$ degenerate in the empty lattice approximation. [*1].

For the convenience to take into account the spin-orbit interaction, we transform basis from $|X\rangle$, $|Y\rangle$, $|Z\rangle$ to

$$|+\rangle \equiv (|X\rangle + i|Y\rangle)/\sqrt{2}, \quad |0\rangle \equiv |Z\rangle, \quad |-\rangle \equiv (|X\rangle - i|Y\rangle)/\sqrt{2},$$

which correspond to eigen functions of angular momentum $|p_{+1}\rangle$, $|p_0\rangle$, $|p_{-1}\rangle$ respectively. With the direct product of these four basis functions for the orbital part and two for the spin part ($\uparrow$, $\downarrow$), eight basis functions in total, roughest $k \cdot p$ perturbation calculation, in which the orbital degeneracy and the spin-orbit interaction are taken into account can be performed.

The perturbation Hamiltonian to $|n\boldsymbol{k}\rangle$ is taken as

$$\mathscr{H}' + \mathscr{H}_{\mathrm{SO}} = -i\frac{\hbar^2}{m_0}\boldsymbol{k}\cdot\boldsymbol{\nabla} - \frac{\hbar}{4m_0^2c^2}\boldsymbol{\sigma}\cdot(\boldsymbol{p}\times\nabla V), \tag{2.63}$$

in which we have dropped higher order terms from (2.58) and put $\boldsymbol{\pi} = \boldsymbol{p}$. The matrix elements between $|S\rangle$, $|X\rangle$, $\cdots$ are

$$P \equiv \frac{\hbar}{m_0}\langle S|p_x|X\rangle = \frac{\hbar}{m_0}\langle S|p_y|Y\rangle = \frac{\hbar}{m_0}\langle S|p_z|Z\rangle, \tag{2.64}$$

$$\Delta \equiv -\frac{3i\hbar}{4m_0^2c^2}\langle X|[\nabla V\times\boldsymbol{p}]_y|Z\rangle = (\text{cyclic replacement of } xyz) \tag{2.65}$$

and their conjugate elements. Others are zero due to the symmetries around nuclei.

Hence non-zero matrix elements of $\mathscr{H}'$ are

$$\langle S\alpha|\mathscr{H}'|0\alpha'\rangle = Pk_z\delta_{\alpha\alpha'}, \ \langle S\alpha|\mathscr{H}'|\pm\alpha'\rangle = \frac{\mp P}{\sqrt{2}}(k_x \pm ik_y)\delta_{\alpha\alpha'}, \tag{2.66}$$

and their conjugate elements $\mathscr{H}'_{ji} = (\mathscr{H}'_{ij})^*$, where α, α' are spin coordinates. As for $\mathscr{H}_{\mathrm{SO}}$,

$$\begin{gathered}\langle\pm\uparrow|\mathscr{H}_{\mathrm{SO}}|\pm\uparrow\rangle = -\langle\pm\downarrow|\mathscr{H}_{\mathrm{SO}}|\pm\downarrow\rangle = \pm\Delta/3,\\ \langle\pm\alpha|\mathscr{H}_{\mathrm{SO}}|0\alpha'\rangle = (1-\delta_{\alpha\alpha'})\sqrt{2}\Delta/3,\end{gathered} \tag{2.67}$$

and others are zero. From (2.62), unperturbed Hamiltonian $\mathscr{H}_0$ has

$$\langle S\alpha|\mathscr{H}_0|S\alpha'\rangle = \delta_{\alpha\alpha'}E_c, \quad \langle\{+,0,-\}\alpha|\mathscr{H}_0|\{+,0,-\}\alpha'\rangle = \delta_{\alpha\alpha'}E_v. \tag{2.68}$$

[*1] *e.g.* see Inui, Tanabe, Onodera, "Applied group theory" (Shokabo, 1976) Chapter 11 (in Japanese).

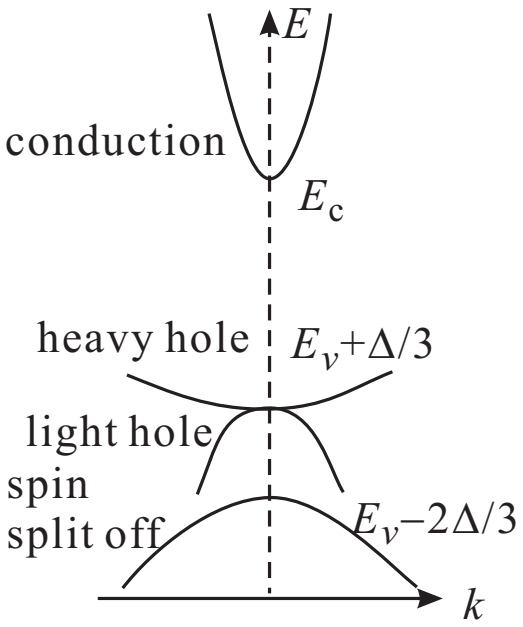


Fig. 2.12 Band structure for diamond and zinc blende semiconductors calculated from the lowest order k·p perturbation with adopting only S and P orbitals. Spin-orbit splitting exists though the heavy hole mass is the same as that of the vacuum electron, that is, the hole mass is negative in this calculation.

From the above we obtain the secular equation and thus the energy eigen values $E_n(\boldsymbol{k})$. $\mathscr{H}'$ is an 8×8 matrix in the present basis though if we fix the wavenumber vector to z direction, *i.e.*, $\boldsymbol{k} = (0, 0, k)$, it becomes

$$\begin{bmatrix} H_d & 0 \\ 0 & H_d \end{bmatrix},$$

thus is broken down to 4×4 matrices and

$$H_d = \begin{bmatrix} E_c & 0 & kP & 0 \\ 0 & E_v - \Delta/3 & \sqrt{2}\Delta/3 & 0 \\ kP^* & \sqrt{2}\Delta/3 & E_v & 0 \\ 0 & 0 & 0 & E_v + \Delta/3 \end{bmatrix}. \tag{2.69}$$

From this the secular equation to obtain the eigenvalue λ is obtained as

$$\lambda = E_v + \frac{\Delta}{3},$$
$$(\lambda - E_c)\left(\lambda - E_v + \frac{2\Delta}{3}\right)\left(\lambda - E_v - \frac{\Delta}{3}\right) - |P|^2k^2\left(\lambda - E_v + \frac{\Delta}{3}\right) = 0.$$

In the second equation we approximate that the term of $|P|^2k^2$ is small then obtain the energies for the conduction band $E_c(\boldsymbol{k})$, and the valence band $E_{vj}(\boldsymbol{k})$ as

$$E_c(\boldsymbol{k}) = E_c + \frac{\hbar^2k^2}{2m} + \frac{|P|^2k^2}{3}\left[\frac{2}{E_\mathrm{g}} + \frac{1}{E_\mathrm{g} + \Delta}\right], \tag{2.70}$$
$$E_{v1}(\boldsymbol{k}) = E_v + \frac{\Delta}{3} + \frac{\hbar^2k^2}{2m_0}, \tag{2.71}$$
$$E_{v2}(\boldsymbol{k}) = E_v + \frac{\Delta}{3} + \frac{\hbar^2k^2}{2m_0} - \frac{2|P|^2k^2}{3E_\mathrm{g}}, \tag{2.72}$$
$$E_{v3}(\boldsymbol{k}) = E_v - \frac{2\Delta}{3} + \frac{\hbar^2k^2}{2m_0} - \frac{|P|^2k^2}{3(E_\mathrm{g} + \Delta)}. \tag{2.73}$$

The band structure around $k = 0$ thus far obtained is displayed in Fig. 2.12. Small mass of the conduction band, two different masses at the top of the valence band, and lowered energy of spin split-off band due to the spin-orbit coupling in the valence band, which properties are well known from optical measurements, cyclotron measurements, etc, are reproduced qualitatively though in particular, the heavier valence band mass is that of the vacuum electron, that is, the

hole effective mass is predicted to be negative apparently different from the real band structure. This is, of course, due to the coarse approximation, which is to the first order perturbation based on the degenerate four bands. The accuracy is enhanced by enhancing the order of perturbation to second, and by taking the surrounding bands into account. At present front of calculation, due to algorithm developments, and enhancement in computational performance have made it possible to perform calculations including over 20 bands and results with high accuracy which can be even used at comparatively high k[4].

Another way to utilize the result of k·p "empirically" is, as is in the pseudo potential method, to represent the results of second order k·p perturbation with a small number of parameters (*e.g.* Luttinger parameters) and to determine them fitting to the experiments. In the case of valence band in diamond and zinc blende semiconductors, the energies can be expressed as

$$E_v(\boldsymbol{k}) = E_v + \frac{\Delta}{3} + Ak^2 \pm \sqrt{B^2k^4 + C^2(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2 + k_x^2)}, \tag{2.74}$$

$$E_{vsp}(\boldsymbol{k}) = E_v - \frac{2\Delta}{3} + Ak^2, \tag{2.75}$$

and A, B, C are obtained from, *e.g.* cyclotron resonance.

2.3 Band structure of graphene

One of the ways to form a two dimensional electron system is to utilize two-dimensional crystals (two-dimensional materials). Graphene is the representative two-dimensional material. Graphene provides a good expample for the application of tight-binding calculation and we would like to see how the things go in a practical (though simplest) example.

The crystal structure of single-layer graphene is show in Fig. 2.13(a), which is a simple honeycomb structure of carbon atoms. The diamond drawn in the figure is the unit cell and the primitive lattice vectors and the primitive reciprocal lattice vectors are written as

$$\boldsymbol{a}_1 = \begin{pmatrix} \sqrt{3}a/2 \\ a/2 \end{pmatrix}, \quad \boldsymbol{a}_2 = \begin{pmatrix} 0 \\ a \end{pmatrix}, \quad \boldsymbol{b}_1 = \begin{pmatrix} 4\pi/\sqrt{3}a \\ 0 \end{pmatrix}, \quad \boldsymbol{b}_2 = \begin{pmatrix} -2\pi/\sqrt{3}a \\ 2\pi/a \end{pmatrix}. \tag{2.76}$$

Henceforce we calculate the electronic states of graphene under simplest approximation. Because the approximation is rough, quantitative comparison with experiments is difficult. However, the results help understanding properties of graphene, *e.g.* the Dirac points appear at the Fermi level in pure graphene. Carbon belongs to group-IV and the outmost electrons exist in the orbitals $2s$, $2p_x$, $2p_y$, $2p_z$. It is easy to see that these orbitals form sp^2-hybrids and the electronic states separate to σ-electrons (sp^2) and π-electrons (p_z). σ-electrons form the honeycome through covalent bonding and the energy bands lie at low energy region. Then the electronic states placed around the Fermi level are π-electrons. Hence we consider Schrödinger equation on π-electrons on the honeycomb lattice.

We write the equation as

$$\psi = \mathscr{H}\psi, \tag{2.77}$$

and as Fig. 2.13(a), we separate the lattice sites to A-sites and B-sites on different sub-lattices. We consider a kind of tight-binding approximation between the two-sites. That is

$$\psi = \zeta_{\mathrm{A}}\psi_{\mathrm{A}} + \zeta_{\mathrm{B}}\psi_{\mathrm{B}}, \tag{2.78}$$

$$\psi_{\mathrm{A}} = \sum_{j\in A} \exp(i\boldsymbol{k}\boldsymbol{r}_j)\phi(\boldsymbol{r} - \boldsymbol{r}_j), \tag{2.79a}$$

$$\psi_{\mathrm{B}} = \sum_{j\in B} \exp(i\boldsymbol{k}\boldsymbol{r}_j)\phi(\boldsymbol{r} - \boldsymbol{r}_j), \tag{2.79b}$$

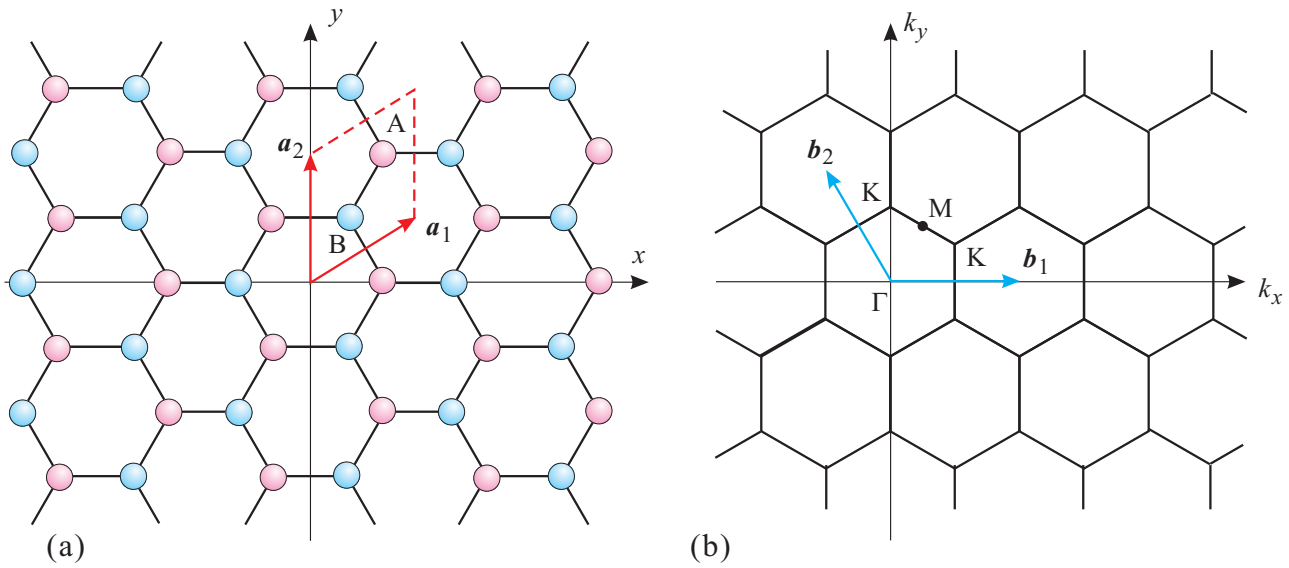


Fig. 2.13 (a) Two dimensional cryatal structure of graphene. Carbon atoms form a honeycomb lattice. It can be also viewed as an overlap of two face-centered square lattices placed at A and B positions. (b) Reciprocal lattice of (a). $\boldsymbol{b}_1, \boldsymbol{b}_2$ are the primitive reciprocal lattice vector corresponding to $\boldsymbol{a}_1, \boldsymbol{a}_2$. The centtral point of the first Brillouin zone is Γ-point and as other points with high symmetries, K-point and M-point are indicated in the figure.

where $\phi(\boldsymbol{r})$ is atomic wavefunction of π-electrons, $\boldsymbol{r}_j$ are the lattice points. Here we write the matrix elements of the Hamiltonian between the each sub-lattice wavefunctions as

$$H_{\mathrm{AA}} = \langle \psi_{\mathrm{A}} | \mathscr{H} | \psi_{\mathrm{A}} \rangle, \quad H_{\mathrm{BB}} = \langle \psi_{\mathrm{B}} | \mathscr{H} | \psi_{\mathrm{B}} \rangle, \quad H_{\mathrm{AB}} = H_{\mathrm{BA}}^* = \langle \psi_{\mathrm{A}} | \mathscr{H} | \psi_{\mathrm{B}} \rangle. \tag{2.80}$$

And the number of atoms in the system is $2N$, that is

$$\langle \psi_{\mathrm{A}} | \psi_{\mathrm{A}} \rangle = \langle \psi_{\mathrm{B}} | \psi_{\mathrm{B}} \rangle = N. \tag{2.81}$$

Let $\langle \psi_{\mathrm{A}} | \psi_{\mathrm{B}} \rangle$ be zero. We substitute (2.78) to (2.77). The condition of have non-trivial $(\zeta_{\mathrm{A}}, \zeta_{\mathrm{B}})$ givies the cecular equation

$$\begin{vmatrix} H_{\mathrm{AA}} - NE & H_{\mathrm{AB}} \\ H_{\mathrm{BA}} & H_{\mathrm{BB}} - NE \end{vmatrix} = 0. \tag{2.82}$$

Lastly

$$E = (2N)^{-1} \left(H_{\mathrm{AA}} + H_{\mathrm{BB}} \pm \sqrt{(H_{\mathrm{AA}} - H_{\mathrm{BB}})^2 + 4|H_{\mathrm{AB}}|^2} \right) \equiv h_{\mathrm{AA}} \pm |h_{\mathrm{AB}}|, \tag{2.83}$$

where we have used $H_{\mathrm{AA}} = H_{\mathrm{BB}}$, which comes from the symmetry, and we use lower cases for the quantities per atom with being devided by $(2N)^{-1}$.

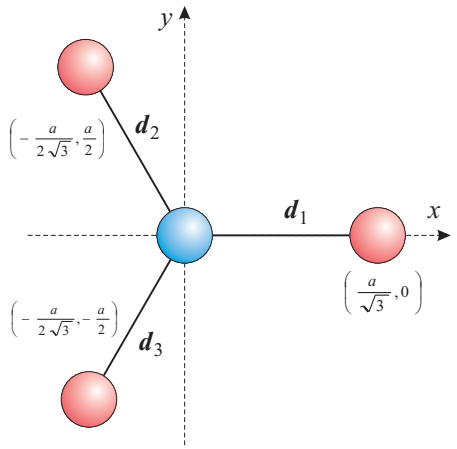


Fig. 2.14 Vectors indicating three directional couplings between nearest neighbor carbon atoms.

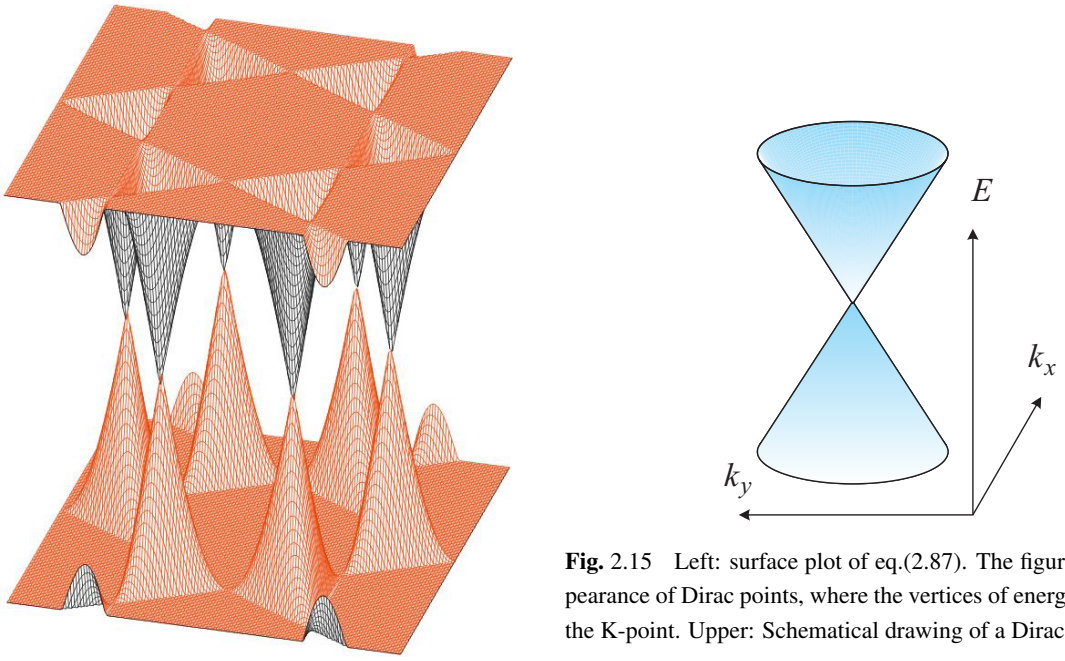


Fig. 2.15 Left: surface plot of eq.(2.87). The figure shows the appearance of Dirac points, where the vertices of energy corns crash at the K-point. Upper: Schematical drawing of a Dirac point.

$$H_{\mathrm{AB}} = \sum_{l\in \mathrm{A}, j\in \mathrm{B}} \exp\left[i\boldsymbol{k}(\boldsymbol{r}_j - \boldsymbol{r}_l)\right] \langle \phi(\boldsymbol{r}-\boldsymbol{r}_l)|\mathscr{H}|\phi(\boldsymbol{r}-\boldsymbol{r}_j)\rangle_{\boldsymbol{r}}. \tag{2.84}$$

We further approximate that the off-diagonal matrix elements of $\mathscr{H}$ just exist between the nearest neighbor sites. For the calculation we take the atom indicated as A in Fig. 2.13(a) as the center atom. The vectors from A to the nearest neighbor atoms 1, 2, 3 are $\boldsymbol{d}_i (i = 1, 2, 3)$ respectively. As is apparent from the figure,

$$\boldsymbol{k}\cdot\boldsymbol{d}_1 = \frac{k_x a}{\sqrt{3}}, \quad \boldsymbol{k}\cdot\boldsymbol{d}_2 = \left(-\frac{k_x}{2\sqrt{3}} + \frac{k_y}{2}\right) a, \quad \boldsymbol{k}\cdot\boldsymbol{d}_3 = \left(-\frac{k_x}{2\sqrt{3}} - \frac{k_y}{2}\right) a, \tag{2.85}$$

where $a = |\boldsymbol{a}_1| = |\boldsymbol{a}_2|$. The terms $\langle \phi(\boldsymbol{r}-\boldsymbol{r}_l)|\mathscr{H}|\phi(\boldsymbol{r}-\boldsymbol{r}_j)\rangle_{\boldsymbol{r}}$ should be equal due to the symmetry and we write it as ξ. Consequently the residual resonant integral from the crystal structure is the repetition of the above and

$$h_{\mathrm{AB}} = \left(\sum_{j=1}^{3} \exp(i\boldsymbol{k}\cdot\boldsymbol{d}_j)\right)\xi. \tag{2.86}$$

Substituting eqs.(2.85), (2.86) into eq.(2.83), we get the following expressio for the energy.

$$E = h_{\mathrm{AA}} \pm \xi\sqrt{1 + 4\cos\frac{\sqrt{3}k_x a}{2}\cos\frac{k_y a}{2} + 4\cos^2\frac{k_y a}{2}}. \tag{2.87}$$

The second term is the perturbation from the nearest neighbor resonant integral, which vanishes at K-point in the reciprocal space

$$(k_x, k_y) = \left(0, \pm\frac{4\pi}{3a}\right), \quad \left(\frac{2\pi}{\sqrt{3}a}, \pm\frac{2\pi}{3a}\right), \quad \left(-\frac{2\pi}{\sqrt{3}a}, \pm\frac{2\pi}{3a}\right). \tag{2.88}$$

We write $k_y = 4\pi/3a$ and around $k_x = 0$(one of the K-points), eq. (2.87) can be approximated as

$$E\left(k_x, \frac{4\pi}{3a}\right) \approx h_{\mathrm{AA}} + \frac{\sqrt{3}\xi a}{2}|k_x|. \tag{2.89}$$

Namely, at the K-point the upper band has a lower pointed shape. Because the same for the lower band and as a result, at the K-point, as shown in Fig. 2.15, the band structure called **Dirac point**, which has no energy gap, no effective mass, appears.

Equation (2.87) is for a very simplified model. Just like a cosine band appeared in the tight-binding model in one-dimension, the model itself does not have realistic meaning. However the model tells that the reason why we have the Dirac points at K-points is that the exsittence of three equivalent resonant integrals in eq. (2.86). The inference holds for the band calculation with any level precision since it is based on the symmetry. That meas the K-points in real graphene are really Dirac points.

Appendix 2A: Band structure calculation based on density function theory

There are many electrons in an actual substance, and the wave function that expresses that state has fermion symmetry called antisymmetry with respect to the particle exchange operation. This causes an electron correlation effect. In addition, a Coulomb repulsive force acts between the electrons. In the semi-empirical band calculation, the effect of these electron-electron interactions is taken into account when the parameters of the one-electron band picture are obtained from the fit to the experimental values, but in the so-called first principle (ab initio) calculation, direct treatment of the electron-electron interaction is required. Calculations that incorporate electron-electron interactions require enormous amounts of calculation for high accuracy, computer resource savings are hence required. The density functional theory (DFT) is very advantageous in that point, thus used in many ab-intio calculations. Nowadays, calculation packages sometimes give us answers even without knowledge of calculations inside. Here, hoewever, we will have a brief look at very basics of ab-initio band calculations[6].

It has become clear that the ab-intio calculations with various approximations reproduce the qualitative features of the semiconductor band structure, but on the other hand, even the band gap, which is the most basic quantity cannot be reproduce without taking into account the quantum correlation effect though this is not an easy task. Incorporating the correlation effect properly is not easy even with DFT, and various theoretical modifications are made that can be said to be ad hoc, which is far from the “first principle" in some cases. It is necessary to pay attention to what kind of approximation is used for the calculation and how accurate it is.

2A.1 Kohn-Sham equation

Hohenberg and Kohn showed that the energy of interacting electron gas with electron density distribution $\rho(\boldsymbol{r})$ in the external potential $v(\boldsymbol{r})$ can be expressed with a universal functional of density $F\{\rho(\boldsymbol{r})\}$ as

$$E\{\rho(\boldsymbol{r})\} = F\{\rho(\boldsymbol{r})\} + \int v(\boldsymbol{r})n(\boldsymbol{r})d\boldsymbol{r}, \tag{2A.1}$$

and also showed that $E\{\rho(\boldsymbol{r})\}$ takes minimum for the true electron density $\rho(\boldsymbol{r})$. The proof is for the case of ground state without degeneracy but the limitation was removed by Levy. Herer we skip the proof.

Let the Hamiltonian without $v(\boldsymbol{r})$ be

$$\mathscr{H}_{\rm f} = -\frac{\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2 + \sum_{i>j}^{N} V(\boldsymbol{r}_i - \boldsymbol{r}_j), \tag{2A.2}$$

and the density functional $F\{\rho\}$ is generally written as

$$F\{\rho\} = \langle\Psi_N|\mathscr{H}_{\rm f}|\Psi_N\rangle. \tag{2A.3}$$

Ψ_N is the wavefunction to give $\rho(\boldsymbol{r})$. To obtain the form of $F\{\rho\}$, we write it in the form

$$F\{\rho\} = T\{\rho\} + U\{\rho\} + E_{\rm xc}^{(0)}\{\rho\}. \tag{2A.4}$$

$T\{\rho\}$ is the kinetic energy, $U\{\rho\}$ is the mean field expression of the Coulomb interaction between the electrons as

$$U\{\rho\} = \frac{e^2}{8\pi\epsilon_0}\int \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}d\boldsymbol{r}d\boldsymbol{r}', \tag{2A.5}$$

and is called Hartree term. The residucal $E_{\rm ex}\{\rho\}^{(0)}$ is called **exchange-correlation energy**.

Even the kinetic energy $T\{\rho\}$ is difficult to be expressed explicitly with ρ and we make modification as follows. We consider an imaginary electron system without interaction in an effective potential $v_{\rm eff}(\boldsymbol{r})$.

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v_{\rm eff}(\boldsymbol{r})\right]\psi_i = \epsilon_i\psi_i(\boldsymbol{r}). \tag{2A.6}$$

We also assume the electron density of this system coinsides with that of the interacting electron system.

$$\rho(\boldsymbol{r}) = \sum_i |\psi_i(\boldsymbol{r})|^2. \tag{2A.7}$$

The kinetic energy of the imaginary system is

$$T_{\rm S}\{\rho\} = -\frac{\hbar^2}{2m}\sum_{i=1}^{N}\int \psi_i^*(\boldsymbol{r})\nabla^2\psi_i(\boldsymbol{r})d\boldsymbol{r}. \tag{2A.8}$$

We now renormalize $E_{\rm ex}$ to include the difference between $T_{\rm S}\{\rho\}$ and $T\{\rho\}$ as

$$E_{\rm ex}\{\rho\} = E_{\rm ex}^{(0)} + T\{\rho\} - T_{\rm S}\{\rho\}. \tag{2A.9}$$

Then the total energy can be written as

$$E\{\rho\} = T_{\rm S}\{\rho\} + U\{\rho\} + E_{\rm ex}\{\rho\} + \int v(\boldsymbol{r})n(\boldsymbol{r})d\boldsymbol{r}. \tag{2A.10}$$

We minimize $E\{\rho\}$ with the variation of $\psi_i(\boldsymbol{r})^*$. For that the normalization condition $\langle\varphi_i|\varphi_i\rangle = 1$ leads to the introduction of Lagrange multipliers $-\epsilon_i$. Then the minimum condition is written as

$$\frac{\delta E\{\rho\}}{\delta\psi_i^*} = \left[-\frac{\hbar^2}{2m}\nabla^2 + v(\boldsymbol{r}) + \frac{e^2}{4\pi\epsilon_0}\int\frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}d\boldsymbol{r}' + \mu_{\rm ex}(\boldsymbol{r}) - \epsilon_i\right]\psi_i(\boldsymbol{r}) = 0, \tag{2A.11}$$

where $\mu_{\rm ex}(\boldsymbol{r})$, defined as

$$\mu_{\rm ex}(\boldsymbol{r}) = \frac{\delta E_{\rm ex}\{\rho\}}{\delta\rho(\boldsymbol{r})} \tag{2A.12}$$

is the quantity called **exchange-correlation potential**. The condition can be written in an eingenvalue equation as

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v(\boldsymbol{r}) + \frac{e^2}{4\pi\epsilon_0}\int\frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}d\boldsymbol{r}' + \mu_{\rm ex}(\boldsymbol{r})\right]\psi_i(\boldsymbol{r}) = \epsilon_i\psi_i(\boldsymbol{r}). \tag{2A.13}$$

Eq. (2A.13) is called Kohn-Sham equation.

In summary, the many-body effects are put into $E_{\rm ex}\{\rho\}$ in the above. The next step is how to calculate this term. In most frequently used method is **local density approximation** (LDA),

$$E_{\rm ex}\{\rho\} = \int \epsilon_{\rm ex}\{\rho(\boldsymbol{r})\}\rho(\boldsymbol{r})d\boldsymbol{r}, \tag{2A.14}$$

then for $\epsilon_{\rm ex}\{\rho(\boldsymbol{r})\}$, the exchange-correlation energy for uniform electron gas with the density ρ. There are many other methods for the calculation.

References

[1] G. Dresselhaus, A. F. Kip, and C. Kittel, Phys. Rev. **98**, 368 (1955).

[2] M. L. Cohen and J. R. Chelikowsky, “Electronics Structure and Optical Properties of Semiconductors" (Springer, 1988).

[3] C. Hamaguchi, “Basic Semiconductor Physics" 3rd ed. (Springer, 2017).

[4] L. C. Lew Van Voon, M. Willatzen, “The k·p method" (Springer, 2009).

[5] P. Hohenberg and W. Kohn, Phys. Rev. B **136**, 864 (1964).

[6] D. Sholl and J. A. Steckei, “Density Functional Theory: A Practical Introduction" (Wiley, 2009).

2.3.1 Effective mass approximation

Let us consider the effect of spatially non-uniform perturbation. For that we add the perturbation potential $U(\boldsymbol{r})$ to the Schrödinger equation in crystals to obtain

$$\left[-\frac{\hbar^2\nabla^2}{2m}+V(\boldsymbol{r})+U(\boldsymbol{r})\right]\zeta(\boldsymbol{r})=[\hat{H}_0+U(\boldsymbol{r})]\zeta(\boldsymbol{r})=E\zeta(\boldsymbol{r}). \tag{2.90}$$

$\zeta(\boldsymbol{r})$ can be expanded with the Bloch functions $\psi_{n\boldsymbol{k}}$, which are the eigenstates of $\hat{H}_0$ as

$$\zeta(\boldsymbol{r})=\sum_{n,\boldsymbol{k}}f(n,\boldsymbol{k})\psi_{n\boldsymbol{k}}(\boldsymbol{r})=\sum_{n,\boldsymbol{k}}f(n,\boldsymbol{k})u_{n\boldsymbol{k}}(\boldsymbol{r})e^{i\boldsymbol{k}\cdot\boldsymbol{r}}. \tag{2.91}$$

With taking the inner product with $\psi_{n'\boldsymbol{k}'}$ after substitution of (2.91) to (2.90),

$$[E_0(n',\boldsymbol{k}')-E]f(n',\boldsymbol{k}')+\sum_{n,\boldsymbol{k}}\langle n',\boldsymbol{k}'|U|n,\boldsymbol{k}\rangle f(n,\boldsymbol{k})=0, \tag{2.92}$$

where $\psi_{n\boldsymbol{k}}$ is written as $|n,\boldsymbol{k}\rangle$. The second term, the transition mediated by U (we write it as $U_{n'\boldsymbol{k}',n\boldsymbol{k}}$), represents the scattering from $|n,\boldsymbol{k}\rangle$ to $|n',\boldsymbol{k}'\rangle$. U and $u^*_{n'\boldsymbol{k}'}u_{n\boldsymbol{k}}$ are Fourier transformed into

$$U(\boldsymbol{r})=\int d\boldsymbol{q}U_{\boldsymbol{q}}e^{-i\boldsymbol{q}\cdot\boldsymbol{r}},\quad u_{n'\boldsymbol{k}'}(\boldsymbol{r})u_{n\boldsymbol{k}}(\boldsymbol{r})=\sum_{\boldsymbol{G}}b_{n'\boldsymbol{k}'n\boldsymbol{k}}(\boldsymbol{G})e^{i\boldsymbol{G}\cdot\boldsymbol{r}}.$$

The transformation of $u^*_{n'\boldsymbol{k}'}u_{n\boldsymbol{k}}$ is a Fourier series on the reciprocal lattice because the term has the lattice periodicity. The coefficients $b_{n'\boldsymbol{k}'n\boldsymbol{k}}$ are written with the unit cell space Ω_0, the unit cell volume v_0 as

$$b_{n'\boldsymbol{k}'n\boldsymbol{k}}(\boldsymbol{G})=\int_{\Omega_0}\frac{d\boldsymbol{r}}{v_0}e^{-i\boldsymbol{G}\cdot\boldsymbol{r}}u^*_{n'\boldsymbol{k}'}(\boldsymbol{r})u_{n\boldsymbol{k}}(\boldsymbol{r}).$$
$$\therefore U_{n'\boldsymbol{k}',n\boldsymbol{k}}=\int d\boldsymbol{q}U_{\boldsymbol{q}}\sum_{\boldsymbol{G}}b_{n'\boldsymbol{k}'n\boldsymbol{k}}(\boldsymbol{G})\int d\boldsymbol{r}e^{i(\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}.$$

The last integral can be performed to be $(2\pi)^3\delta(\boldsymbol{k}-\boldsymbol{k}'+\boldsymbol{q}+\boldsymbol{G})$, and the integration over $\boldsymbol{q}$ gives

$$U_{n'\boldsymbol{k}',n\boldsymbol{k}}=(2\pi)^3\sum_{\boldsymbol{G}}U_{\boldsymbol{k}'-\boldsymbol{k}-\boldsymbol{G}}\,b_{n'\boldsymbol{k}'n\boldsymbol{k}}(\boldsymbol{G}). \tag{2.93}$$

$U(\boldsymbol{r})$ is assumed to have much slower spatial variation than the lattice potential. Then as $U_{\boldsymbol{q}}$, it is enough to restrict ourseleved to $|q|\ll\pi/a$, *i.e.* much smaller values than that of the Brillouin zone edge. The approximation corresponds to $\boldsymbol{k}'-\boldsymbol{k}\sim\boldsymbol{G}$. We further assume that U does not cause strong scattering that drives the state to the zone edge, then $\boldsymbol{G}$ takes only $\vec{0}$. Also $|U|$ is smaller than the band gap, then there is no interband scattering via U, *i.e.* there is no matrix element for $n\neq n'$. Then we can approximate

$$U_{n'\boldsymbol{k}',n\boldsymbol{k}}\approx U_{\boldsymbol{k}'-\boldsymbol{k}}\delta_{n'n}. \tag{2.94}$$

(2.92) is written as

$$[E_0(\boldsymbol{k}')-E]f(n,\boldsymbol{k}')+\sum_{\boldsymbol{k}}U_{\boldsymbol{k}'-\boldsymbol{k}}f(n,\boldsymbol{k})=0. \tag{2.95}$$

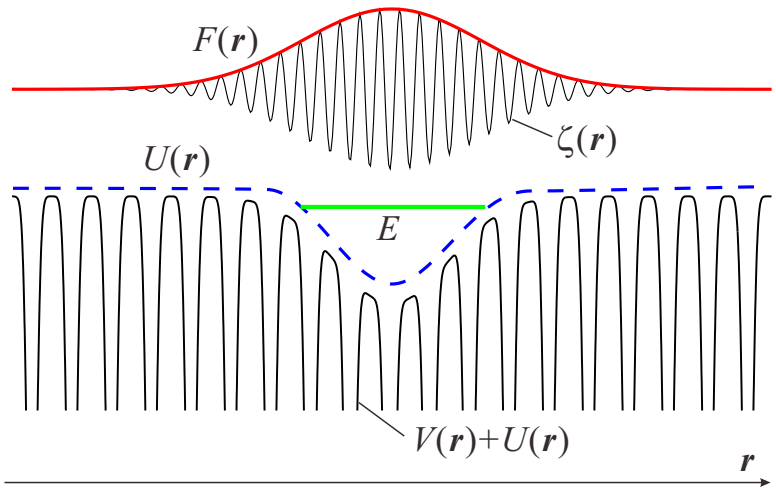


Fig. 2.16 A perturbation pontetial $U(\boldsymbol{r})$ is superposed on the crystal potential $V(\boldsymbol{r})$. The figure illustrates the potential and the wavefunction $\zeta(\boldsymbol{r})$, the envelope function $F(\boldsymbol{r})$ for the system with the crystal potential $V(\boldsymbol{r})$ the slowly varing perturbation potential $U(\boldsymbol{r})$.

Nest we consider the expansion in (2.91). In the present approximation, only the region $\boldsymbol{k} \sim 0$ is considered for $u_{n\boldsymbol{k}}$, and u is almost constant for $\boldsymbol{k}(\approx u_{n0})$. Then we take it out from the sum over $\boldsymbol{k}$.

$$\zeta_n(\boldsymbol{r}) = u_{n0} \sum_{\boldsymbol{k}} f(n,\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = u_{n0} F_n(\boldsymbol{r}), \tag{2.96}$$

where the index n is attached to ζ with ignoring the intermixing of the bands. Here, $F_n(\boldsymbol{r})$ defined as

Envelope function

$$F_n(\boldsymbol{r}) \equiv \sum_{\boldsymbol{k}} f(n,\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \tag{2.97}$$

is the inverse Fourier transformation of $f(n,\boldsymbol{k})$, and called **envelope function**. $F_n(\boldsymbol{r})$ should be a slowly varing function over the scale of lattice constant(Fig. 2.16).

For the unperturbed dispersion relation, we apply that of a particle with the effective mass, and for simplicity the effective mass m^* is assumed to be isotropic. Then $E_0(\boldsymbol{k}) = \hbar^2\boldsymbol{k}^2/2m^*$ is substituted to (2.95) to give

$$\frac{\hbar^2 k \boldsymbol{k}'^2}{2m^*} f(\boldsymbol{k}) + \sum_{\boldsymbol{k}} U_{\boldsymbol{k}'-\boldsymbol{k}} f(\boldsymbol{k}) = E f(\boldsymbol{k}'). \tag{2.98}$$

Here we omit writing n. The above can be inverse-Fourier transfomed with the care that the second term becomes convolution because it already has the summation over $\boldsymbol{k}$ to give

Effective mass equation

$$\left[\frac{\hbar^2\nabla^2}{2m^*} + U(\boldsymbol{r})\right] F(\boldsymbol{r}) = E F(\boldsymbol{r}). \tag{2.99}$$

The equation takes the form of the Schrödinger equation of the paticle with the mass m^* and the potential $U(\boldsymbol{r})$. That is, for the envelop function, the problem is now a particle with the effective mass in the perturbation potential. This way of handling the problems on the level of envelope function is called **effective mass approximation**, and eq.(2.99) is called effective mass equation. In this sense, the Bloch states can be viewed as plane waves in the effective mass approximation.

The viewpoint is very usuful in designing various quantum systems in solids with semiconductor technologies. We test the approximation for the shallow impurity states in the next chapter. We also use it in many places in this lecture. We should be careful, however, that the envelope function is not the wavefunction itself. The difference becomes clear, particularly when the perturbation potential has a sharp spatial variation.

Chaper 3 Carrier statistics and impurity doping

In this chapter we consider the energy distribution of **carriers** in semiconductors. We introduce the concept of carrier doping with very little amount of impurities, which brings drastic changes in the electric conduction.

3.1 Carrier statistics in intrinsic semiconductors

We call a pure semiconductor without any impurity as an **intrinsic semiconductor**. Of course this is just an idea, but *e.g.* non-doped Si for LSIs' can be considered as an intrinsic semiconductor. And under some conditions other semiconductors can also be treated as intrinsic semiconductors.

3.1.1 Density of states

We consider a simple lattice system which has a state per a unit cell with an edge length of a. We take the system size as $L = Na$ in one dimension. For an n-dimensional system, the volume $(2\pi/L)^n$ contains a single state in k-space(Fig. 3.1(a)). Given the kinetic energy as $E(k) = \hbar^2 k^2/2m$, the number of states per volume between E and $E + dE$ (Fig. 3.1(b)) devided by dE is

$$\mathscr{D}(E) = \frac{1}{L^d}\left(\frac{L}{2\pi}\right)^d \frac{dV_d(k)}{dE} = \frac{1}{(2\pi)^d}\frac{dV_d(k)}{dk}\frac{dk}{dE} = \frac{1}{(2\pi)^d}\frac{m_0}{\hbar^2}\frac{dV_d(k)}{kdk}, \tag{3.1}$$

where $V_d(k)$ is the volume of d-dimensional sphere with the radius of k. This $\mathscr{D}(E)$ is called **energy density of state**. Because $V_1 = 2k,\ \ V_2 = \pi k^2,\ \ V_3 = 4\pi k^3/3$ (Fig. 3.2),

$$\mathscr{D}^{(0)}_{d=1} = \frac{1}{\pi\hbar}\sqrt{\frac{2m_0}{E}}, \quad \mathscr{D}^{(0)}_{d=2} = \frac{m_0}{\pi\hbar^2}, \quad \mathscr{D}^{(0)}_{d=3} = \frac{\sqrt{2m_0^3}}{\pi^2\hbar^3}\sqrt{E}, \tag{3.2}$$

where the factor 2 comes from the freedom of spin.

In the case of electrons in crystals, the above expressions for density of states are applicable with replacing the mass with the effective mass where non-parabolicity of the band is ignorable, *e.g.*, around tops and bottoms of the bands. When we cannot apply the parabolic approximation, we need to go back to the definition of the density of states. For a three

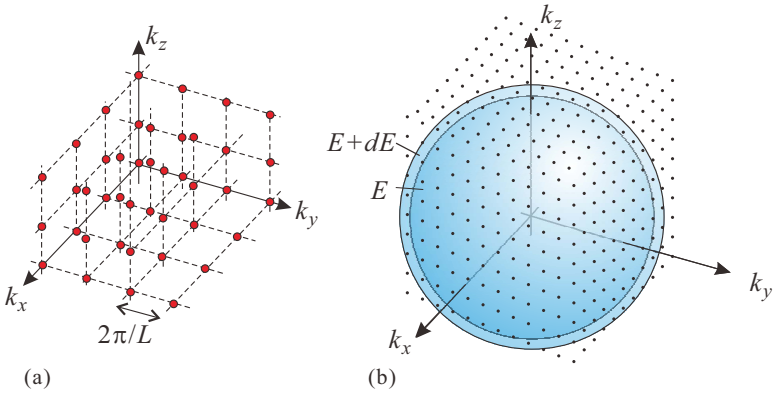


Fig. 3.1 (a) The red dots represent possible wavenumber in k-space in 3d empty lattice approximation for simple cubic. (b) Counts the number of dots in the spherical shell from E to $E + dE$.

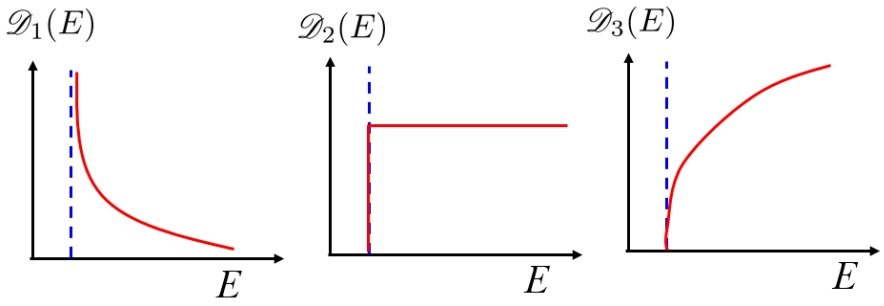


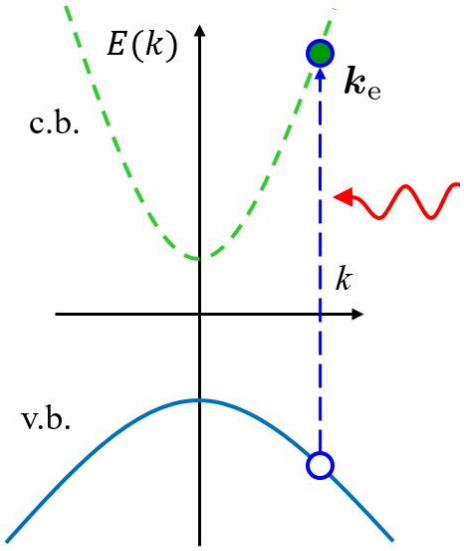
Fig. 3.2 Schematic diagrams of density of states for 1, 2, 3 dimentions in (3.2).

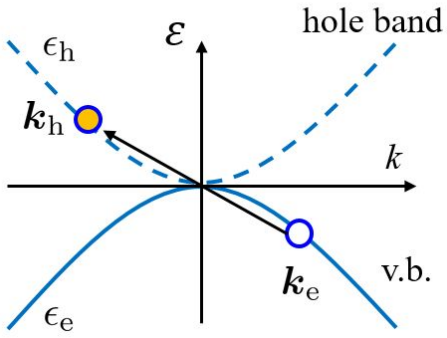
dimensional system it is given from

$$\mathscr{D}(E) = \int_{E(\boldsymbol{k})=E} \frac{dS_k}{(2\pi)^3} \frac{2}{\nabla_{\boldsymbol{k}} E(\boldsymbol{k})}. \tag{3.3}$$

The integral is over the equi-energy surface $E(\boldsymbol{k}) = E$ in k-space.

3.1.2 Concept of holes





The total current $\boldsymbol{J}_{\rm v.b.}$ carried by a full valence band is zero by canceling of counter-going electrons ($\boldsymbol{J}_{\rm v.b.} = \sum_{\rm v.b.}(-e)\boldsymbol{v}_{\boldsymbol{k}} = 0$). When a state with crystal momentum $\boldsymbol{k}$ is empty as seen in the left, the current is

$$\boldsymbol{J}_{\rm v.b.}(\boldsymbol{k}) = \sum_{\rm v.b.}(-e)\boldsymbol{v}_{\boldsymbol{k}'} - (-e)\boldsymbol{v}_{\boldsymbol{k}} = e\boldsymbol{v}_{\boldsymbol{k}}, \tag{3.4}$$

as if there is a particel with the charge $+e$ and the velocity $\boldsymbol{v}_{\boldsymbol{k}}$. Such a many-body state in valence band is called **hole**.

We write the wavenumber of hole as $\boldsymbol{k}_{\rm h}$, then it should be the variation of the total wavenumber (momentum) due to the creation of the hole.

$$\boldsymbol{k}_{\rm h} = \sum_{\rm v.b.} \boldsymbol{k}'_{\rm e} - \boldsymbol{k}_{\rm e} = -\boldsymbol{k}_{\rm e}. \tag{3.5}$$

When an electric field $\boldsymbol{E}$ is applied, the valence electrons are accelerated and move in the k-space. The “hole" follows the movement, that is, the equation of motion for holes is the same as that for electrons. However, we define a hole has the charge $+e$, then the acceleration by the electric field should be opposite to that for electrons and for consistency the sign of the effective mass should be opposite.

$$m^* \frac{d\boldsymbol{v}}{dt} = (-e)\boldsymbol{E} \;\rightarrow\; (-m^*)\frac{d\boldsymbol{v}}{dt} = e\boldsymbol{E}.$$

The kinetic energy decreases with the creation of a hole and if we take the origin of energy at the top of valence band, we get

$$\left(\frac{1}{m_h^*}\right)_{ij} = -\left(\frac{1}{m_e^*}\right)_{ij}, \quad E_{\rm h}(\boldsymbol{k}_{\rm h}) = E_{\rm h}(-\boldsymbol{k}_{\rm e}) = -E_e(\boldsymbol{k}_{\rm e}). \tag{3.6}$$

Form (3.6), m_h^* is positive around the valence band top and the dispersion is obtained with 180° rotation of the electron dispersion. The density of states $\mathscr{D}_h(E)$ is the same as $\mathscr{D}_e(E)$. The above definitions make it possible to treat the holes as positive charge carriers. The hole band drawn in the lower left is used to be consistent with the electron dispersion and the hole picture. We need to be careful that the frequently-used whilte hole picture (actually in the left) which is really an electron dispersion and the “white hole" is placed at $\boldsymbol{k}_{\rm e}$ which is $\boldsymbol{k}_{\rm h}$.

3.1.3 Carrier distribution in thermal equilibrium

Let us see how electrons and holes distribute in energy space at a finite temperature obeying the Fermi distribution function. For a while we treat general properties, which hold also for doped semiconductors. The effect of doping can be included in the position of the Fermi level $E_{\rm F}$. The numbers of electrons and holes which exist in $E \sim E + dE$ are

$$g_e(E)dE = \mathscr{D}_e(E)f(E)dE, \tag{3.7a}$$
$$g_h(E)dE = \mathscr{D}_h(E)[1 - f(E)]dE \equiv \mathscr{D}_h(E)f_h(E)dE. \tag{3.7b}$$

Here we introduced the hole distribution function as (Fig. 3.3(c)),

$$f_h(E) = 1 - f(E) = \frac{1}{1 + \exp(E_{\rm F} - E)/k_{\rm B}T)}. \tag{3.8}$$

For the density of states, we use those of particles with the effective masses. From (3.2),

$$\mathscr{D}_e(E) = \frac{\sqrt{2m_e^{*3}}}{\pi^2\hbar^3}\sqrt{E - E_c} \quad \text{(conduction band)}, \tag{3.9a}$$
$$\mathscr{D}_h(E) = \frac{\sqrt{2m_h^{*3}}}{\pi^2\hbar^3}\sqrt{E_v - E} \quad \text{(valence band)}. \tag{3.9b}$$

Here E_c, E_v are the bottom of conduction band and the top of valence band respectively as in Fig. 3.3(a).

Hence the distributions of electrons and holes at a finite temperature should be as in Fig. 3.3(b), giving the electron concentration in the conduction band n, the hole concentration p in the valence band as

$$n = \int_{E_c}^{\infty} g_e(E)dE = \frac{\sqrt{2m_e^{*3}}}{\pi^2\hbar^3}\int_{E_c}^{\infty}\frac{\sqrt{E - E_c}dE}{1 + \exp(E - E_{\rm F})/k_{\rm B}T}, \tag{3.10a}$$
$$p = \int_{-\infty}^{E_v} g_h(E)dE = \frac{\sqrt{2m_h^{*3}}}{\pi^2\hbar^3}\int_{-\infty}^{E_v}\frac{\sqrt{E_v - E}dE}{1 + \exp(E_{\rm F} - E)/k_{\rm B}T}. \tag{3.10b}$$

In the case of $f_{\rm F}(E) \ll 1 (E \geq E_c)$, $f_h(E) \ll 1 (E \leq E_v)$, the distribution can be approximated by Maxwellian as

$$f_{\rm F}(E) \sim \exp(E_{\rm F} - E)/k_{\rm B}T, \quad f_h(E) \sim \exp(E - E_{\rm F})/k_{\rm B}T. \tag{3.11}$$

We apply the identity

$$\int_0^{\infty}\sqrt{x}e^{-x}dx = \frac{\sqrt{\pi}}{2}$$

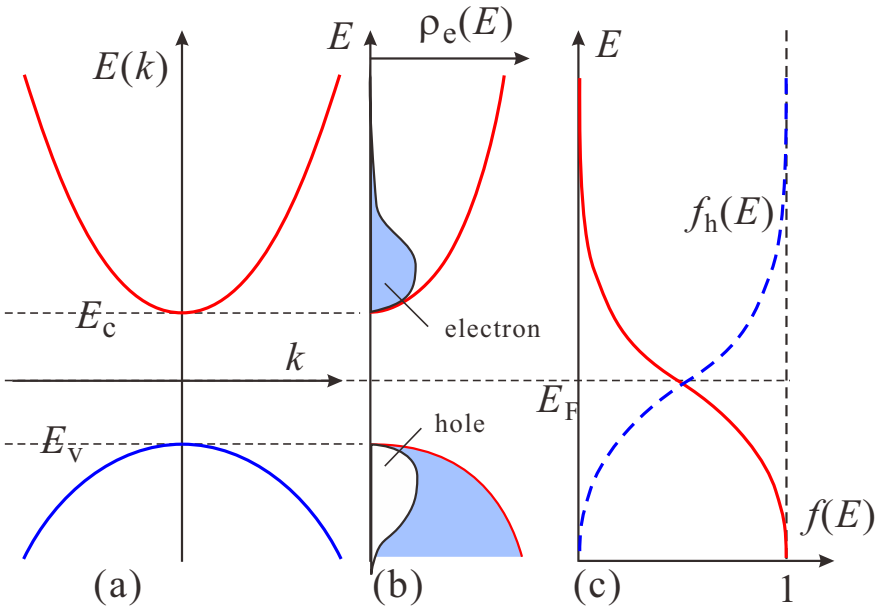


Fig. 3.3 (a) Schematic diagram of energy bands. (b) Density of states and the distributions of electrons $n(E)$(blue-gray), $p(E)$(white). (c) Electron distribution function $f(E)$(red solid line), and hole distribution function $f_h(E)$(blue broken line).

with $x = (E - E_{\rm F})/k_{\rm B}T$ to obtain

$$n = 2\left(\frac{m_e^* k_{\rm B}T}{2\pi\hbar}\right)^{3/2} \exp\left(\frac{E_{\rm F} - E_c}{k_{\rm B}T}\right) \equiv N_c \exp\left(\frac{E_{\rm F} - E_c}{k_{\rm B}T}\right), \tag{3.12a}$$

$$p = 2\left(\frac{m_h^* k_{\rm B}T}{2\pi\hbar}\right)^{3/2} \exp\left(\frac{E_v - E_{\rm F}}{k_{\rm B}T}\right) \equiv N_v \exp\left(\frac{E_v - E_{\rm F}}{k_{\rm B}T}\right). \tag{3.12b}$$

N_c and N_v are the coefficients which give n, p respectively in the situation the energy states are concentrated at E_c and E_v. They are called **effective density of states**. From (3.10a) and (3.10b) we obtain

Law of mass action

$$np = N_c N_v \exp\left(\frac{E_v - E_c}{k_{\rm B}T}\right) = N_c N_v \exp\left(-\frac{E_g}{k_{\rm B}T}\right) = n_{\rm i}^2 \tag{3.13}$$

Here the width of forbidden band $E_g \equiv E_c - E_v$ is called **energy gap**, and $n_{\rm i}$ は真性半導体の場合のキャリア濃度である．Equation (3.13) does not depend on the position of $E_{\rm F}$, which varies, *e.g.* with doping. In other words, the product np in the thermal equilibrium is determined only by the temperature and the species of semiconductor.

In intrinsic semicondutors, there is no space charge and the charge neutral condition leads to $n = p$ hence is written as $n_{\rm i}$ in the above law of mass action. The relation $n = p$ in intrinsic semiconductors leads to

$$E_{\rm F} = \frac{E_c + E_v}{2} + \frac{k_{\rm B}T}{2}\ln\frac{N_v}{N_c} = \frac{E_c + E_v}{2} + \frac{3k_{\rm B}T}{4}\ln\frac{m_h}{m_e}, \tag{3.14}$$

which gives the position of $E_{\rm F}$. At low temperatures the second term gets small and $E_{\rm F}$ comes close to the middle of the band gap.

3.2 Impurity doping

In semiconductors, very small amount of impurities give drastic change in the material properties. Such addition of impurities is called **doping** [*1].

3.2.1 Donors and acceptors

As a typical example, the case of Si is shown schematically in Fig. 3.4. In Si pure crystal, as in (a), a Si atom has four nearest neighbor atoms, which have 4 covalent electrons. As a result each atom has eight electrons in the outmost shell, which fill up $3s$ and $3p$ orbits forming the closed shell geometry. When the center atom is replaced with an Sb (group-V) atom, there is an excess electron for the closed shell structure as in (b). On the other hand, the positive charge in the nucleous excesses the negative one of surrounding electrons by $+e$, which forms a Coulomb potential around the Sb atom. The excess electron is excited to the conduction band or loosely trapped in the bound state in the Coulomb potential. An impurity that emits electrons to the conduction band or the shallow levels is called **donor**.

When the center atom is replaced with a B atom (group-III), which is just the opposite of Sb, there are not enough electrons to form a closed shell structure. Therefore, holes are created in the valence band to supplement the electrons, but as a result, the electron charge becomes extra around the B atom, and a Coulomb potential of only $-e$ is generated. Impurities that emit holes into the valence band and shallow levels in this way are called **acceptors**(acceptor).

[*1] In so called studies in strongly correlated systems, which began with the studies of high-$T_{\rm c}$ superconductors, addtion of atoms wtih concentrations even far above 1% is called “doping" as far as the crystal structure is unchanged. Such regions are called “alloying" in the semiconductor fields. Furthermore, enhancement of carrier concentrations with application of strong electric field is sometimes called “electric field doping", and the addition of impurities is called “chemical doping." Here, however, I follow tranditional epxression in the field of semiconductors.

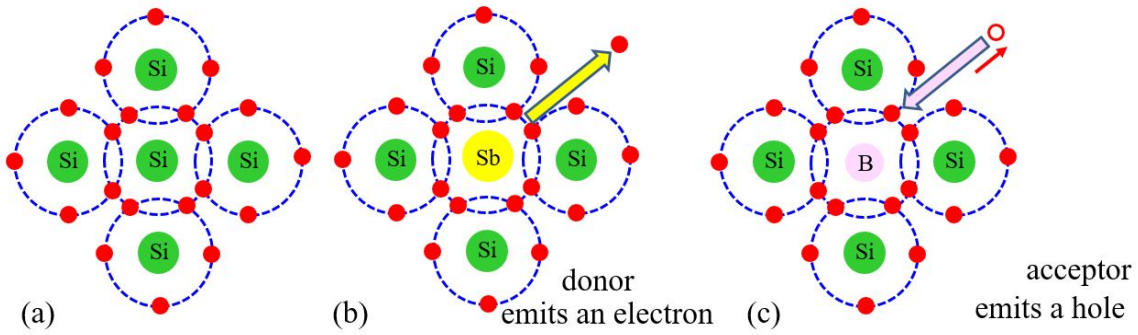


Fig. 3.4 (a) Illustration of electronic structure in the outmost shell in a Si atom in a Si crystal. (b) In the case of replacement with an Sb atom. There is an excess electron for Sb to form the closed shell structure. The excess electron should go out from the covalent system. (c) In the case of replacement with a B atom, which opens a “hole" in the valence band.

The situation is a little more complicated in the case of compound semiconductors than in Group IV elemental semiconductors. For example, when a group III-V semiconductor is doped with a group IV element, replacing the group III site becomes a donor, and replacing the group V site becomes an acceptor. Elements whose donor / acceptor changes depending on the doping method are called amphoteric.

3.2.2 Effective mass approximation for shallow hydrogen-like levels

Regarding the Coulomb potential formed by donors and acceptors, semiconductors generally have a relatively large permittivity due to the polarization of valence electrons, and this impurity attraction potential is considerably weaker than in vacuum. Therefore, the binding energy of the impurity bound state is smaller than that of the hydrogen atom, and in many cases, it spreads over several unit cells, and the effective mass approximation can be applied.

For the case of isotropic effective mass, taking the origin at the impurity potision we write the effective mass equation from $U(r) = -e^2/4\pi\epsilon_0\epsilon r$ as

$$\left[-\frac{\hbar^2\nabla^2}{2m^*} - \frac{e^2}{4\pi\epsilon_0\epsilon r}\right] F(\boldsymbol{r}) = EF(\boldsymbol{r}), \tag{3.15}$$

which as the same form as that of hydrogen atom other than the effective mass m^* and the relative permittivity ϵ. Hence we can readily apply the results for hydrogen atom. Here we write the **effective Rydberg constant** and the **effective Bohr radius** as

$$Ry^* = \frac{e^2 m^*}{2(4\pi\epsilon\epsilon_0)^2\hbar^2} = \frac{m^*}{m}\frac{1}{\epsilon^2}Ry, \quad a_{\mathrm{B}}^* = \frac{4\pi\epsilon\epsilon_0\hbar^2}{m^* e^2} = \frac{m}{m^*}\epsilon a_{\mathrm{B}} \tag{3.16}$$

respectively. The eigenenergy is then represented as

$$E_n = E_c - \frac{Ry^*}{n^2} \quad (n = 1, 2, \cdots) \tag{3.17}$$

and the wavefunction corresponding to $1s$ state is

$$\psi_{1s}(\boldsymbol{r}) = \sqrt{\frac{1}{\pi a_{\mathrm{B}}^{*3}}} \exp\left(-\frac{\boldsymbol{r}}{a_{\mathrm{B}}^*}\right). \tag{3.18}$$

An expample of semiconductor with such an isotropic effective mass is GaAs. At the conduction band minimum placed at Γ-point, $\epsilon \approx 11.5$, $m^* \approx 0.067m$. $a_{\mathrm{B}}^* = 172a_{\mathrm{B}} = 91$ Åis sufficiently longer than the lattice constant 5.65 Åand guarantees the legitimacy of the effective mass approximation. From $Ry^* = 5.07 \times 10^{-4} Ry = 5.57 \times 10^3$ m^{-1} the binding energy of $1s$ state is as small as 6.9 meV.

Semiconductor	Calculated binding energy (meV)	Experimental binding energy (meV)
GaAs	5.72	Si_{Ga}(5.84); Ge_{Ga}(5.88) S_{As}(5.87);
InP	7.14	7.14
InSb	0.6	Te_{Sb} (0.6)
CdTe	11.6	In_{Cd} (14); Al_{Cd} (14)
ZnSe	25.7	Al_{Zn} (26.3); Ga_{Zn} (27.9) F_{Se} (29.3); Cl_{Se} (26.9)

Tab. 3.1 Effective mass approximation for hydrogen-like impurities and the measured value of binding energies.

Tab. 3.1 shows the comparison of the value given by (3.17) and experimentally measured values for isotropic effective mass condition. The agreement is satisfactory.

Then what if the effective mass is anisotropic and there are six conduction band valleys, as in Si? We consider the effective mass approximation for the valley along (0,0,1). The equation for the spheroidal surface is

$$E_1(\boldsymbol{k}) = \frac{\hbar^2}{2}\left[\frac{k_x^2+k_y^2}{m_t} + \frac{(k_z-k_0)^2}{m_l}\right]. \tag{3.19}$$

Then the effective mass equation is

$$\left[-\frac{\hbar^2}{2m_t}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right) - \frac{\hbar^2}{2m_l}\frac{\partial^2}{\partial z^2} - \frac{e^2}{4\pi\epsilon_0\epsilon r}\right] F(\boldsymbol{r}) = EF(\boldsymbol{r}). \tag{3.20}$$

The eigenfunction can be approximated by the variational method assuming an anisotropic exponential function. As a trial function we take a and b as the paramters and write down as

$$F_{1s}(\boldsymbol{r}) = \sqrt{\frac{1}{\pi a^2 b}}\exp\left(-\sqrt{\frac{x^2+y^2}{a^2}+\frac{z^2}{b^2}}\right). \tag{3.21}$$

The stationary condition gives numerical solutions as $a = 2.5$ nm, $b = 1.42$ nm, $E = 29$ meV. However the experiments give 33 meV for Li (the shallowest case), 45 meV for P manifesting that the approximation is not appropriate. Further discussion will be given in Appendix 3B.

3.3 Carrier statistics in doped semiconductors

Let us consider the case we dope donors uniformly with the density $N_{\rm D}$. At absolute zero all the electrons emitted from the donors are bound to the donors. [*2] At finite temperatures some of them are excited to the conduction band and can carry electric charges. We call them “carriers" or “electrons". Let n be the density of such electrons and $n_{\rm D}$ be the density of electrons bounded at the donors. From the charge neutrality condition we get $n + n_{\rm D} = N_{\rm D}$.

Now we estimate Helmholtz free energy $F = U - TS$ by considering the number of cases W for assigning $n_{\rm D}$ electrons to $N_{\rm D}$ states. From $S = k_{\rm B}\ln W$,

$$F = E_{\rm D}n_{\rm D} - k_{\rm B}T\ln\left[2^{n_{\rm D}}\frac{N_{\rm D}!}{n_{\rm D}!(N_{\rm D}-n_{\rm D})!}\right].$$

$E_{\rm D}$ is the position of the bound state measured from the bottom of the conduction band and $2^{n_{\rm D}}$ is due to the spin degeneracy. We assume that the Coulomb repulsion prevents double occupation of a localized state with two electrons.

[*2] In so called degenerate semiconductors the following discussion does not hold.

According to Starling approximation $\ln N! \sim N \ln N - N$, the chemical potential (Fermi energy) is given as

$$\mu = E_{\mathrm{F}} = \frac{\partial F}{\partial n_{\mathrm{D}}} = E_{\mathrm{D}} - k_{\mathrm{B}} T \ln \left[\frac{2(N_{\mathrm{D}} - n_{\mathrm{D}})}{n_{\mathrm{D}}} \right]. \tag{3.22}$$

And from this

$$n_{\mathrm{D}} = N_{\mathrm{D}} \left[1 + \frac{1}{2} \exp \left(\frac{E_{\mathrm{D}} - E_{\mathrm{F}}}{k_{\mathrm{B}} T} \right) \right]^{-1} \tag{3.23}$$

is obtained. The factor 1/2 on the exponential function is due to the spin degeneracy.

Similarly, for uniform doping of acceptors with density N_{A}, the density of electrons bounded to the acceptors n_{A} is

$$n_{\mathrm{A}} = N_{\mathrm{A}} \left[1 + 2 \exp \left(\frac{E_{\mathrm{A}} - E_{\mathrm{F}}}{k_{\mathrm{B}} T} \right) \right]^{-1}. \tag{3.24}$$

Here we have a factor 2 instead of 1/2 but the density of holes bounded to the acceptors is $p_{\mathrm{A}} = N_{\mathrm{A}} - n_{\mathrm{A}}$ and symmetrical with n_{D} having a factor 1/2.

From (3.22), if we dope only “shallow" donors, for which the effective mass approximation holds, E_{F} comes to E_{D} at $T \to 0$. E_{D} should be much smaller than E_{g}. Accordingly from (3.23), the electron concentration n becomes much higher than that of the intrinsic semiconductor at finite temperatures. This type of semiconductors are called **n-type**. Similarly doping of acceptors enhances the hole concentration p. We call them **p-type**.

When donors and acceptors co-exist, the semiconductor becomes n-type for $N_{\mathrm{D}} \gg N_{\mathrm{A}}$ and p-type for $N_{\mathrm{D}} \ll N_{\mathrm{A}}$. In the former, some of the electrons emitted from donors are captured to acceptors and almost all the acceptors are ionized. In the latter, the other way around. In both cases we say such semiconductors are **compensated**.

Remember the semiconductor equation (3.13) then the product np does not depend on the doping. If one of n , p increases with doping, then the other decreases. In the case of n-type semiconductor under $N_D \gg N_A$, n is much higher than p by many orders, and we call the electrons **majority carriers** and the holes **minority carriers**. The other way around in the case of p-type semiconductors.

Even in the presence of donors and acceptors eq.(3.12) hold and simultaneous satisfaction of them gives n, p and E_{F}. To obtain E_{F} with knowledge of n, p approximate expressions

$$E_{\mathrm{F}} \approx E_C + k_{\mathrm{B}} T \left[\ln \left(\frac{n}{N_C} \right) + 2^{-3/2} \left(\frac{n}{N_C} \right) \right], \tag{3.25a}$$

$$E_{\mathrm{F}} \approx E_V - k_{\mathrm{B}} T \left[\ln \left(\frac{p}{N_V} \right) + 2^{-3/2} \left(\frac{p}{N_V} \right) \right] \tag{3.25b}$$

are convenient. In the region where (3.23), (3.24) hold the last term can be omitted.

In an n-type semiconductor with compensation, p,n_{A} can be ignored and the electrically neutral condition is

$$n + N_A = N_{\mathrm{D}} - n_{\mathrm{D}}. \tag{3.26}$$

Substitution of eq.(3.23) gives

$$\frac{n + N_{\mathrm{A}}}{N_{\mathrm{D}} - N_{\mathrm{A}} - n} = \frac{1}{2} \exp \left(\frac{E_{\mathrm{D}} - E_{\mathrm{F}}}{k_{\mathrm{F}} T} \right). \tag{3.27}$$

Equation (2.22) holds for the case of doped semiconductors with shifts of E_{F}, multiplication of each side of the equation results in

$$\frac{n(n + N_{\mathrm{A}})}{N_{\mathrm{D}} - N_{\mathrm{A}} - n} = \frac{1}{2} N_c \exp \left(-\frac{\Delta E_{\mathrm{D}}}{k_{\mathrm{B}} T} \right), \quad \Delta E_{\mathrm{D}} \equiv E_c - E_{\mathrm{D}}. \tag{3.28}$$

The temperature dependence of carrier concentration n described by eq.(3.28) has the following four characteristic regions:

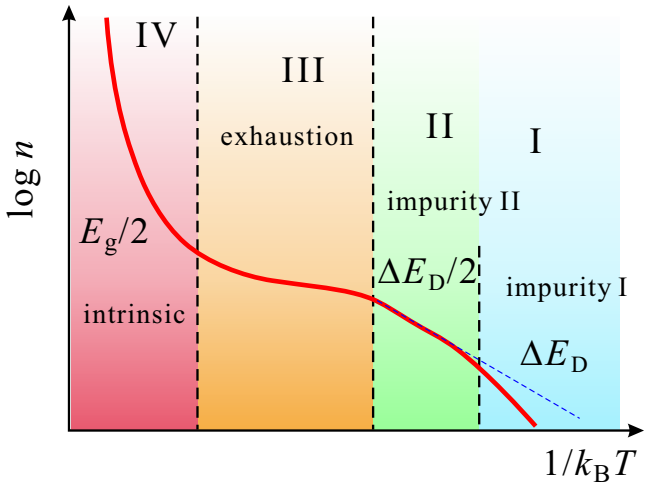


Fig. 3.5 Characteristic four temperature regions of an n-type semiconductor with compensation. Schematical temperature dependence of carrier concentration n is plotted versus $1/T$ in semi-log scale.

I. Impurity (Freeze-out) region I: At very low temperatures and the case of $n \ll N_A \ll N_D$,

$$n \approx \frac{N_D N_c}{2N_A} \exp\left(-\frac{\Delta E_D}{k_B T}\right), \tag{3.29}$$

where n decreases with lowering the temperature in an Arrhenius type with an activation energy of ΔE_D.

II. Impurity (Freeze-out) region II: In middle temperature range, in the case of $N_A \ll n \ll N_D$,

$$n \approx \left(\frac{N_c N_D}{2}\right)^{1/2} \exp\left(-\frac{\Delta E_D}{2k_B T}\right), \tag{3.30}$$

where the temperature dependence shows again an Arrhenius type but with a different activation energy, which is a half of that in the impurity region I.

III. Exhaustion (Saturation) region: Temperature is higher than ΔE_D ($k_B T > \Delta E_D$). The exponential function in eq.(3.28) is now almost a constant (~ 1) and

$$n \approx N_D - N_A. \tag{3.31}$$

Electrons once captured in donors are “exhaustively" excited to the conduction band and work as carriers.

IV. Intrinsic region: At higher temperatures where direct thermal excitation for the valence band to conduction band cannot be ignored in comparison with N_D, the temperature dependence of the carrier concentration asymptotically approaches to that in an intrinsic semiconductor described as eq.(3.10b), (3.14).

We show the behavior in Fig. 3.5 schematically. For semiconductor devices, the exhaustion region III is mostly used.

Appendix 2B: Wannier functions and the effective mass approximation

There is a way to derive the effective mass approximation by using expansion with the **Wannier function**. Thouhg it is essentially the same as in Sec.2.3.1, Wannier functions have several convenient points and we may use them afterwards. In that case, I will introduce it again, but let’s take a quick look at what it is as an appendix.

2B.1 Wannier function

The Wannier function is defined as Fourier transform of Bloch function as follows.

$$w_n(\boldsymbol{r}-\boldsymbol{R}_j)=\frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}}\exp(-i\boldsymbol{k}\cdot\boldsymbol{R}_j)\psi_{n\boldsymbol{k}}(\boldsymbol{r}). \tag{2B.1}$$

The Bloch function is usually given in the coordinate representation but here the spatial coordinate is a parameter and it is now taken as a function of wavenumber $\boldsymbol{k}$. The summation over $\boldsymbol{k}$ is inside the Brillouin zone. The Bloch function is a product of a lattice periodic function and a plane wave and sperad over the space. On the other hand, the Wannier function has tendency to localized to the lattice point $\boldsymbol{R}_j$. This is understood by making the lattice-periodic function in the Bloch function a constant, which makes the Wanner function completely localized on $\boldsymbol{R}_j$. Equation (2B.1) can be seen as the expantion of the Wannier function with the Bloch function. Conversely the Bloch function can be expanded by the Wannier function as

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r})=\frac{1}{\sqrt{N}}\sum_{\boldsymbol{k}}\exp(i\boldsymbol{k}\cdot\boldsymbol{R}_j)w_n(\boldsymbol{r}-\boldsymbol{R}_j). \tag{2B.2}$$

An advantage in the Wannier function is the orthgonality, that is

$$\langle w^*_{n'}(\boldsymbol{r}-\boldsymbol{R}_{j'})|w_n(\boldsymbol{r}-\boldsymbol{R}_j)\rangle=\delta_{jj'}\delta_{nn'}. \tag{2B.3}$$

We skip the proof but straightforwardly performed with using the summatoins on the lattice and the reciprocal lattice. The Wannier function is normalized if it is defined as (2B.1) and also forms a complete set.

2B.2 Derivation of effective mass approximation

The problem is the addition of perturbation potential $U(\boldsymbol{r})$ to the crystal Hamiltonian $\mathscr{H}_0$, that is

$$[\mathscr{H}_0+U(\boldsymbol{r})]\phi(\boldsymbol{r})=E\phi(\boldsymbol{r}). \tag{2B.4}$$

The derivation goes almost parallely as that with the Bloch funtion. First we expand the wavefunction with the Wannier functions as

$$\phi(\boldsymbol{r})=\sum_{n,j'}F_n(\boldsymbol{R}_{j'})w_n(\boldsymbol{r}-\boldsymbol{R}_{j'}). \tag{2B.5}$$

We assume that $\mathscr{H}_1$ does not have the elements for interband transition (the amplitude is too small) and we drop n henceforth. Tne Wannier function $w(\boldsymbol{r}-\boldsymbol{R}_j)$ is simply written as $|j\rangle$. Equation (2B.5) is substituted into eq. (2B.4) and with takeing the inner product with $\langle j|$, the orthogonality (2B.3) leads to

$$\sum_{j'}\langle j|\mathscr{H}_0|j'\rangle F(\boldsymbol{R}_{j'})+\sum_{j'}\langle j|U(\boldsymbol{r})|j'\rangle F(\boldsymbol{R}_{j'})=EF(\boldsymbol{R}_j). \tag{2B.6}$$

As seen above $|j\rangle$ is localized to $\boldsymbol{R}_j$ and because $U(\boldsymbol{r})$ is slowly varying function in the scale of lattice constant, we can approximate as

$$\sum_{j'}\langle j|U(\boldsymbol{r})|j'\rangle\approx\sum_{j'}U(\boldsymbol{R}_{j'})\langle j|j'\rangle=U(\boldsymbol{R}_j). \tag{2B.7}$$

The term of crystal Hamiltonian can be written with shifting the spatial origin as

$$\langle j|\mathscr{H}_0|j'\rangle=\langle w(\boldsymbol{r})|\mathscr{H}_0|w(\boldsymbol{r}-(-\boldsymbol{R}_{j'}+\boldsymbol{R}_j))\rangle\equiv h_0(\boldsymbol{R}_j-\boldsymbol{R}_{j'}). \tag{2B.8}$$

The Bloch funcition $\psi_{\boldsymbol{k}}(\boldsymbol{r})$ is the eigenstate of $\mathscr{H}_0$ and the application of the effective mass approximation to the eigenenergy gives

$$\langle\psi_{\boldsymbol{k}}(\boldsymbol{r})|\mathscr{H}_0|\psi_{\boldsymbol{k}}(\boldsymbol{r})\rangle=E_0(\boldsymbol{k})=\frac{\hbar^2\boldsymbol{k}^2}{2m^*}. \tag{2B.9}$$

$\psi_{\boldsymbol{k}}(\boldsymbol{r})$ can be expanded as (2B.2) and leads to

$$E_0(\boldsymbol{k}) = \frac{1}{N}\sum_{j,j'} \exp[-i\boldsymbol{k}\cdot(\boldsymbol{R}_j - \boldsymbol{R}_{j'})]\langle j|\mathscr{H}_0|j'\rangle = \frac{1}{N}\sum_{j,j'} \exp[-i\boldsymbol{k}\cdot(\boldsymbol{R}_j - \boldsymbol{R}_{j'})]h_0(\boldsymbol{R}_j - \boldsymbol{R}_{j'})$$
$$= \sum_j \exp(-i\boldsymbol{k}\cdot\boldsymbol{R}_j)h_0(\boldsymbol{R}_j). \tag{2B.10}$$

The inverse transformation gives

$$h_0(\boldsymbol{R}_j) = \frac{1}{N}\sum_{\boldsymbol{k}} E_0(\boldsymbol{k})\exp(i\boldsymbol{k}\cdot\boldsymbol{R}_j). \tag{2B.11}$$

Though $F(\boldsymbol{R}_j)$ is just defined on the lattice point as in (2B.5), since the spatial variation in $U(\boldsymbol{r})$ is slow, the differences between the values of F for the neighboring lattice point are small and smooth interpolation is possible. The operator of spatial shift by $\boldsymbol{a}$ is $\exp(-\boldsymbol{a}\cdot\nabla)$[*3], then we can write

$$F(\boldsymbol{r} - \boldsymbol{R}_j) = \exp(-\boldsymbol{R}_j\cdot\nabla)F(\boldsymbol{r})$$

to obtain

$$\sum_{j'} h_0(\boldsymbol{R}_{j'})F(\boldsymbol{r} - \boldsymbol{R}_{j'}) = \sum_{j'} h_0(\boldsymbol{R}_{j'})\exp(-\boldsymbol{R}_{j'}\cdot\nabla)F(\boldsymbol{r}). \tag{2B.12}$$

On the other hand for (2B.10),

$$E_0(\boldsymbol{k})F(\boldsymbol{r}) = \sum_{j'} h_0(\boldsymbol{R}_{j'})\exp(-i\boldsymbol{k}\cdot\boldsymbol{R}_{j'})F(\boldsymbol{r}). \tag{2B.13}$$

We formally inverse Fourier transform (2B.12) and (2B.13). Then these equations have the common right hand side if we make replacement of $\boldsymbol{k} \to -\nabla$. Therefore we can write

$$\sum_{j'} h_0(\boldsymbol{R}_{j'})F(\boldsymbol{r} - \boldsymbol{R}_{j'}) = E_0(-i\nabla)F(\boldsymbol{r}). \tag{2B.14}$$

All the above results are restored into (2B.6) and we replace $\boldsymbol{R}_j$ with a continuous variable $\boldsymbol{r}$. Then we obtain

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2 + U(\boldsymbol{r})\right]F(\boldsymbol{r}) = EF(\boldsymbol{r}). \tag{2B.15}$$

Here I have introduce the proof in the textbook [1]. It has, though, a small jump in the logic from (2B.13)→(2B.14). Also the description of $\boldsymbol{R}_j \to \boldsymbol{r}$ is rather vague. The derivation in [2] is more strict but it needs the width of paper. For the transformation $\boldsymbol{R}_j \to \boldsymbol{r}$, clearly written as “to be strict, this should be done in variations."

Appendix 3A: Methods for impurity doping

Various methods have been developed for impurity doping. A part of it is introduced in the following. Impurities to be doped are called dopants, and base crystal is called host.

3A.1 Mixing of impurities to the raw material

We have introduced the semiconductor crystal growth method, but especially in the method of growing bulk crystals from a material melt, if impurities are mixed in the raw material in advance, doping may be performed with relatively good uniformity. In many cases, for example, even in the Czochralski method, a concentration gradient is generated in the crystal growth direction by segregation. For this reason, various growth measures are taken, such as adding a dopant

[*3] This can be confirmed, for example by the Taylor expansion.

to the melt crucible in order to obtain a uniform doping concentration. In addition, in the case of amphoteric impurities, there is a possibility that some impurities will be compensated depending on the growth conditions.

In epitaxial thin film growth, by controlling the dopant and irradiating it on the growth surface, it is possible to start a stepwise distribution and create various concentration distributions with high non-equilibrium while maintaining crystallinity. This bf modulation doping method plays a major role in occupying an extremely important position in the semiconductor industry for semiconductor thin films.

3A.2 Thermal diffusion method

A method in which the host is kept at a high temperature in a state where the dopant is present in a high concentration on the surface of the host, and is mixed inside by heat diffusion. Methods for increasing the concentration of the surface include contacting with the vapor of the dopant and pre-depositing a thin film of the dopant on the host surface. In the figure below, the dopant and host wafer are simultaneously enclosed in a quartz tube and heated together so that the vapor of the dopant flows to the surface of the high-temperature wafer. In some cases, the whole is put into the furnace without creating a flow. In addition, if some of the constituent elements of the host have a high vapor pressure, it is necessary to suppress the separation from the surface by mixing the vapor of this element.

In the thermal diffusion method, the concentration is high near the surface and low as it goes inside. It is usually used to form a device near the surface. It is used for integrated circuit formation because the doping region can be patterned by masking the wafer when the vapor is applied.

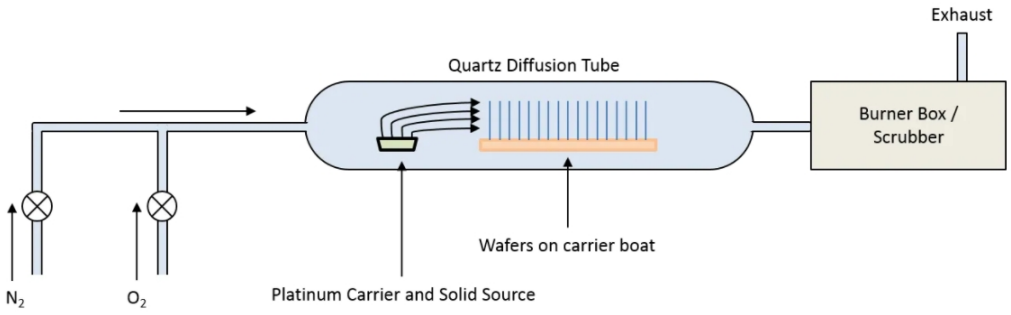


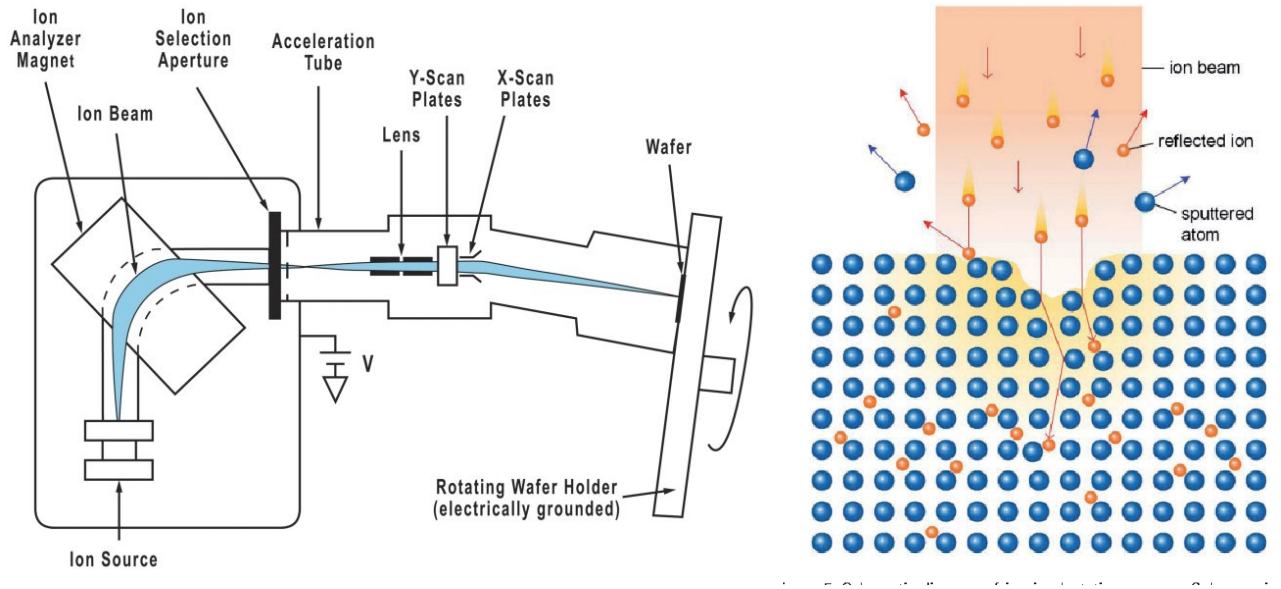
Fig. 3A.1 Schematic diagram of thermal diffution doping.

3A.3 Ion implantation

The ion implantation method is used not only for doping but also for cutting and oxidation of the inner layer. As in Fig. 3A.2(a), ions entering from the source are bent by a magnetic field, passed through a diaphragm for mass spectrometry, then narrowed down by a lens, irradiated on a wafer, and scanned by an XY voltage.

As in the imaginary figure in (b), Since the ions that reach the surface have high kinetic energy, they invade the crystal in a non-equilibrium manner and stop at a depth corresponding to the average kinetic energy. Since the crystallinity of the passing region decreases due to the collision of ions and the dopant is not always in the stable position, it is often annealed after implantation. The distribution of dopants is generally represented by the Gaussian distribution after annealing.

In addition to doping, it was once actively used as one of the Silicon on Insulator (SOI) techniques that form an oxide film inside by implanting oxygen ions and annealing as described above. At present, the seemingly primitive method of bonding after surface oxidation is mainly used.



(a) (b)

Fig. 3A.2 (a) Illustration of ion implantation doping. Ions coming out of the source are subjected to mass spectrometry using a magnetic field to sort them, and then the focused beam is scanned onto the wafer. (b) Imaginary illustration of the host surface during the ion implantation.

3A.4 Irradiation with neutrons

Currently, it is rarely used, but there is an interesting doping method that uses neutrons. For that the following nuclear reaction is used.

$$[n + {}^{30}\mathrm{Si}] \rightarrow {}^{31}\mathrm{Si} \rightarrow [{}^{31}\mathrm{P} + \beta]$$

With reactor neutrons, extremely uniform doping can be performed without compromising crystallinity. However, due to problems such as throughput, this method remains at the research level.

Appendix 3B: Shallow donors in Si

For the improvement of the effective mass approximation for the shallow donor levels in Si, we consider the effect of multiple (6) valleys. Since the impurity potential also has a significant magnitude and steepness near the center, it is possible that it has matrix elements between the eigenfunctions attached to the degenerate valleys. We consider the donor wavefunction $\chi(\boldsymbol{r}) = F(\boldsymbol{r})\psi(\boldsymbol{r})$, where $F(\boldsymbol{r})$ is the envelope function and $\psi(\boldsymbol{r})$, for each valley and obtain the donor function as a linear combination of them as

$$\phi^{(i)}(\boldsymbol{r}) = \sum_{j=1}^{6} \alpha_j^{(i)} \chi_j, \tag{3B.1}$$

where j is the index of valley and i is the index of symmetry reflecting that of surrounding atoms. Here we assume there is no mixing between the eigenstates with different quantum number in the trap potential.

While the isotropic potential approximation does not hold in the vicinity of impurities, the spatial symmetry of the **crystal field** created by the atoms around the impurity ions must be taken into consideration when taking a linear bond. It is convenient to use point group theory for the discussion, and I will use it here as well. Regarding the point group, if I have time, I would like to supplement the minimum knowledge in the appendix of lecture notes etc.[**?**, 3]. In the

case of Si, the nearest neighbor atoms are at the apexes of a regular tetrahedron containing impurity ions. The point group corresponding to the symmetry is expressed as the symbol $T_{\rm d}$. The elements of point group have correspondences to the representations, which express symmetry operations. The independent elements have one-to-one correspondence to the reduced representations. The reduced representations in $T_{\rm d}$ group are A_1, E and T_1, which have single, double, triple degeneracy respectively and there are six elements. The index i in eq.(3B.1) correspons to these six elements. The coefficients for these elements are as in the following table.

		j						
	normalization const.	1	2	3	4	5	6	expression
$\alpha_j^{(1)}$	$1/\sqrt{6}$	1	1	1	1	1	1	A_1
$\alpha_j^{(2)}$	1/2	1	1	-1	-1	0	0	E
$\alpha_j^{(3)}$	1/2	1	1	0	0	-1	-1	E
$\alpha_j^{(4)}$	$1/\sqrt{2}$	1	-1	0	0	0	0	T_1
$\alpha_j^{(5)}$	$1/\sqrt{2}$	0	0	1	-1	0	0	T_1
$\alpha_j^{(6)}$	$1/\sqrt{2}$	0	0	0	0	1	-1	T_1

Tab. 3.2 Linear combination coefficients for the donor states in Si.

Then we index the donor states with the quantum number of valley wavefunction chi_j and the above reduced representation. Here the qunantum number (qn) of χ_j is determined by main qn n, directional qn l, magnetic qn m but with anisotropy. Hence the indices are like $1s(A_1)$, $1s(E)$ etc.

From Tab. 3.2, we see that in the elements other than A_1, the wavefunctions are superposed in the inverse phase and the amplitude at the origin is small. In A_1, all the wavefunctions are superposed in phase and the amplitude at the origin is large. Hence the state of $1s(A_1)$ largely deviates from the effective mass approximation and that causes large decrease in the eigenenergy. The larger the positive charge in the nuclear the larger the binding energy. This eneryg splitting is called **valley-orbit splitting**.

Effective mass theory	Li	P	As	Sb	Bi
32	32.5	45	53.7	43	70.6

Tab. 3.3 Donor binding energy in Si (meV)

As in Tab. 3.3, such tendency really appears in $1s(A_1)$. Pantelides and Sah gave theoretical calculation of the valley-orbit splitting, which reproduces the experiments well[4].

References

[1] C. Hamaguchi, “Basic Semiconductor Physics" 3rd ed. (Springer, 2017).

[2] P. Yu and M. Cardona, “Fundamentals of Semiconductors", (4th ed. Springer, 2010).

[3] M. Tinkham, “Group Theory and Quantum Mechanics" (Dover, 2003).

[4] S. Pantelides and C. T. Sah, Phys. Rev. B 10, 621-637 (1974); *ibid.* 638-658 (1974).

3.3.1 Degenerate semiconductors

So far we treat the impurity states in semiconductors as isolated. In such cases, as shown in Fig.3.2, charge carriers disappear at low temperatures, the conductivity is lost, and the system is insulating. Now we consider doping to higher impurity densities, where the average distance between the dopants is similar to or less than the spatial size of the wavefunctions. Then the overlapping of wavefunctions enables tunneling between the impurity sites. Such tunnleings may form a kind of conducting network in the crystal and with further increasing the impurity concentration finally spreads the network over the whole crystal, which now has a finite conductance at the lowest temperature, thus is a metal.

This problem – **metal-insulator transitions**, MIT – has been one of the most important problems in condensed matter physics, and huge amount of efforts have been devoted for years. The field of MIT extends over various phenomena in condensed matter physics, far beyond the doped semiconductors. We have not reached the final answer through great amount of knowledges have been accumulated. There are so many textbooks, very few of which are listed in references ([1]∼[5]).

In the above we have defined the MIT as the spatial size of the wavefunctions at the Fermi level. The phenomenon is observed in the energy space as follows. With overlapping of neighboring wavefunctions, as we have seen in the tight binding model (regular, coherent case), the energy levels broadened and a band is formed, which we call an **impurity band**. Even under the formation of impurity band, in which the density of states is continuous, the wavefunction at the Fermi level is not necessarily spread over the entire crystal. It was first pointed out by Anderson that the electrons in a potential with a certain degree of disorder are spatially localized. This is called **Anderson localization**. Hence, some lower part of the impurity band is usually localized and the boundary is called a mobility edge.

It is well known that as shown in Fig. 3.6, in many matrix crystals and species of dopants, an empirical relation,

$$n_c^{1/3} a_{\mathrm{B}}^{*} = 0.26 \tag{3.32}$$

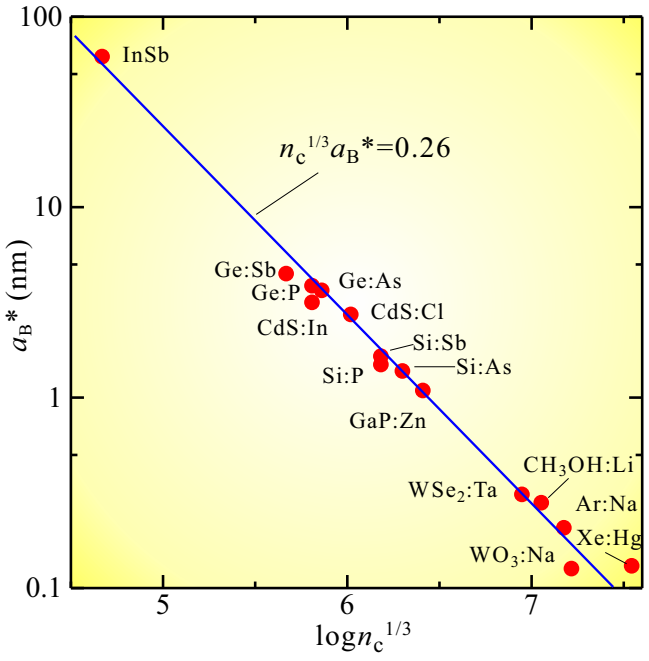


Fig. 3.6 Experimental values for the critical concentrations of the MIT and the effective Bohr radiuses for various matrix semiconductors and dopants (the element symbols put after the colons). The data are plotted in log scale. The unit for n_c is cm^{-3}. The line indicates the empirical relation in eq.(3.32). The data are taken from P. Edwards and M. Sienko, Phys. Rev. B **17**, 2575 (1978).

holds between the critical impurity concentration n_c for the MIT and the effective Bohr radius $a_{\rm B}^*$. This criterion is natural from the view of impurity band formation and there are many trials to derive it from more rigorous theoretical background.

The largest difficulty in solving this problem lies in the treatment of disorder, which makes it impossible to utilize the coherence of the scattering from the crystal lattice. In the band theory, the coherence brings about great simplicity. In the case of MIT in disordered systems, one should directly treat the disorder itself.

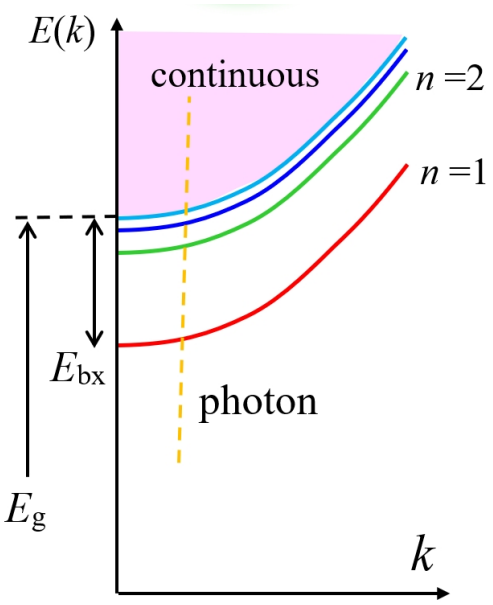
Though the final answer has not been found, or it is not even known whether there is such one or not, many new physical idea have been developed, which have greatly expanded our knowledges on random systems. The concepts have been applied various fields such as organic semiconductors. We do not go into this problem further due to the space-time limitations.

The MIT in ordered systems is also an important and difficult problem, particularly in so called strongly correlated systems. Here I just list a review paper[6], which was published a quarter century ago. I will not go into this problem in this lecture.

There are a number of devices which utilize such degenerate semiconductors. A representative is the Esaki diode (tunnel diode), which first attracted attensions as a device for fast switching and actually was used in counter circuits in the experiments of high energy physics. Recently, the Esaki diode is also used in the interfacial connection of multiple junction solar cells. In many of ordinary solar cells, highly-doped degenerate semiconductors are used as the upper layers of the junctions. The high level doping is also used in the IMPATT diodes for high-frequency use, and p-i-n type photo-diode, etc.

3.3.2 Excitons

Here we introduce the concept of exciton, which is a bit tail subject as "carrier statistics" but we can view it as an application of the effective mass approximation. Exciton has long been the central theme of optical properties[7], but even more extensive research is still underway, such as the BEC of exciton polaritons. In solids, the bound states formed by Coulomb force between quasiparticles of positive and negative charges are called **exciton**.



When the quasiparticles are spread over several lattice constants, the exciton is called "Wanner type." When the charge polarization occurs within a molecule or over very few lattice points, it is called "Frenkel type." The latter often found in organic semiconductors, in which the molecules at the lattice points are comparatively well separated. Here we concentrate ourselves on the Wannier type.

Let us consider an exciton state with an electron excited to the conduction band, and a hole excited to the valence band. These spread over several lattice points or more, and effective mass approximation can be applied. Based on the free state of both electrons and holes, even if they create a bound state, the degree of freedom of the movement of the center of mass remains, and the "wave number" and kinetic energy due to this remain. This wavenumber should be derived from the overall wavenumber conservation since the concept of holes is also introduced by considering the conservation of the total wavenumber. The "mass" of the exciton also would be introduced simply by taking the sum of effective masses of electrons and holes, as $m_{\rm e} + m_{\rm h}$. I have used the expression "would" because the Coulomb force, which is most natural candidate

for the force in the equation of motion for electrons and holes, works in the opposite directions, the accelerations for two kinds of particles with different effectve masses is a bit complicated for the treatment. Anyway we assume that the effective mass approximation holds. Then the creatio energy of an exciton from the state without the electron and the hole, hence including the electron-hole pair creation energy, is written as

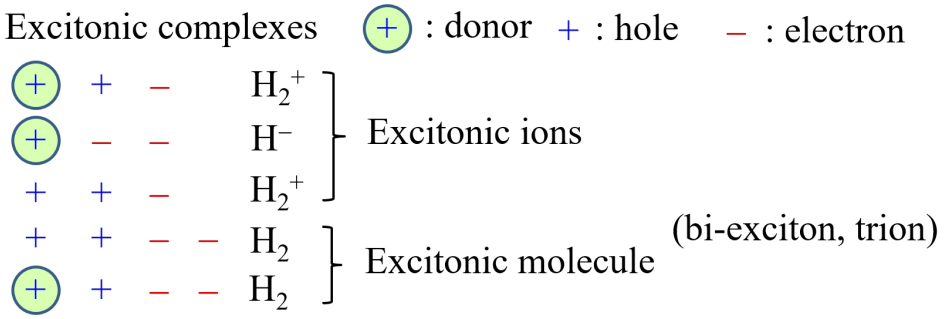
$$E_{\rm ex}^{(n)}(k) = E_{\rm g} + \frac{\hbar^2 k^2}{2(m_{\rm e}+m_{\rm h})} - \frac{e^4 m_{\rm r}^*}{2\hbar^2}\frac{1}{n^2}, \quad n = 1, 2, \cdots, \tag{3.33}$$

where the third term is the binding energy of the electron and the hole. We assume the system is isotropic and the exciton is hydrogen-like. $m_{\rm r}$ is the reduced mass defined as

$$\frac{1}{m_{\rm r}^*} = \frac{1}{m_{\rm e}^*} + \frac{1}{m_{\rm h}^*}. \tag{3.34}$$

The second term in (3.33) is the kinetic energy of the parallel motion. The dispersion described in eq.(3.33) is illustrated in the figure in the previous page. The existence of such bound states can be confirmed by checking, *e.g.* the optical absorption spectra, which we will see in the next chapter.

However, for example in optical absorption experiments, there often appear many absorption peak spectra which cannot be interpreted simply with eq.(3.33). The candidates for the interpretation of those observations are, the excitons trapped by impurity potentials (bound excitons), the exciton molecules made of more than two excitons, or such complicated excited states.



In the above illustration, the cases for the number of charged excitation including zero or single donor is from three to four (corresponding to hydrogen molecule or its charged state) are listed. Such excited states are called **exciton comlexes**.

In this chapter, we introduced electrons excited in the conduction band, holes excited in the valence band, and excitons, which are bound states of these excitations.

All of them are many-body states of electrons, but they can be treated as if "particles" are freely moving in the space of a crystal, which is different from the vacuum. Such free particle-like pictures, in which many-body effects are renormalized are called **quasi-particle**.

Chapter 4 Optical response of bulk semiconductors

Many of the substances called “semiconductors” have bandgaps around the energy region of electromagnetic wave called “light”, and have characteristic optical responses. The optical response is one of the most important subject as well as the carrier transport. In the optical devices such as detectors or emitters, semiconductors are mostly used as active materials. In this chapter, as the first look, we see basic optical properties of semiconductor bulk materials.

4.1 Optical response of two-level systems

In order to consider the optical response of semiconductor bulk, we should investigate the relationship between light and the transition between the extended electronic states of the valence band and conduction band that we have seen so far. But here, we begin with the optical response of a much simpler “two-level system.” The reason why we devote our pages to such basic matters here is that we want to confirm the zero-point oscillation of the electromagnetic field and the state of photons in particular. The following two sections are for the lecture to be just self-contained. For more complete description, see the textbooks listed in [8]. If the reader already has such knowledge, the skip to Sec.4.1.3 is recommended. In addition, if he/she is already used to the two-level systems, a further skip to Sec.4.2 is also OK.

4.1.1 Quantizationo of electromagnetic field

We have a very short look at the quantization of electromagnetic field to consider the states of photons[8]. As the basics we start with the one-dimensional harmonic oscillator, which subject appears in the beginning part of elementary quantum mechanics. The problem is discribed as the Schrödinger equation;

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}+\frac{m\omega_{\rm h}^2x^2}{2}\right]\phi=E\phi. \tag{4.1}$$

The second term in the parenthis in the left hand side represents the potential characteristic for the harmonic oscillator. We define a dimensionless variable q with

$$x=\sqrt{\frac{\hbar}{\omega_{\rm h}m}}q, \tag{4.2}$$

and rewrite (4.1) as

$$\frac{\hbar\omega_{\rm h}}{2}\left(-\frac{d^2}{dq^2}+q^2\right)\phi=E\phi. \tag{4.3}$$

We introduce down and up operators

$$a=\frac{1}{\sqrt{2}}\left(\frac{d}{dq}+q\right),\quad a^\dagger=\frac{1}{\sqrt{2}}\left(-\frac{d}{dq}+q\right),\quad [a,a^\dagger]=1,\ (\text{others})=0, \tag{4.4}$$

where the commutation relation is readily derived from the definition. Then eq.(4.3) is furthre rewritten as

$$\hbar\omega_{\rm h}\left(a^\dagger a+\frac{1}{2}\right)\phi(\equiv\hat{H}\phi)=E\phi. \tag{4.5}$$

If we define the **number operator** as

$$\hat{n}\equiv a^\dagger a. \tag{4.6}$$

Because $\hat{n}$ and $\hat{H}$ commute with each other ($[\hat{n},\hat{H}]=[\hat{H},\hat{n}]=0$), they have common eigenfunctions. Here we assume $|w\rangle$ is a eigen function common for $\hat{H}$ and $\hat{n}$ with eigen[values ϵ, γ respectively. From the commutation relation eq.(4.4), we see

$$\hat{H}(a^\dagger|w\rangle)=(\gamma+\hbar\omega_{\rm h})(a^\dagger|w\rangle),\quad \hat{H}(a|w\rangle)=(\gamma-\hbar\omega_{\rm h})(a|w\rangle), \tag{4.7}$$

that is, $a^\dagger|w\rangle$, $a|w\rangle$ are also such simultaneous eigenfunctions with the energy eigenvalues of up and down shifts by $\hbar\omega_{\rm h}$ respectively. Let $|0\rangle$ be the simultaneous eigenstate with the lowest energy eigenstate ϵ_0. Since there is no eigenstate with energy eigenvalue of $\epsilon_0-\hbar\omega_{\rm h}$, the above equation leads to $a|0\rangle=0$. Furthermore, $\epsilon_0=\hbar\omega_{\rm h}/2$ is concluded.

On the other hand, the eigenstates with higher energy eigenvalues than ϵ_0 can be obtained by sequential application of $a^\dagger$ to $|0\rangle$. The eigenvalues are

$$E_n=\hbar\omega_{\rm h}\left(n+\frac{1}{2}\right)\quad (n=0,1,2,\cdots). \tag{4.8}$$

And the commutation relation tells that the operatore $a^n(a^\dagger)^n$ works as multiplication of $n!$. Then the normalized eigenfunction for the eigenvalue E_n can be obtained from the normalized $|0\rangle$ is written as

$$|n\rangle=\frac{(a^\dagger)^n}{\sqrt{n!}}|0\rangle. \tag{4.9}$$

Also from $a|0\rangle=0$, a solution of $|0\rangle=\varphi_0(q)$ is straightforwardly obtained as

$$\frac{d\varphi_0}{dq}+q^2\varphi_0=0\quad \therefore\varphi_0=\frac{1}{\pi^{1/4}}\exp\left(-\frac{q^2}{2}\right). \tag{4.10}$$

Based on the above knowledge, we go to the electromagnetic field. Our starting point here is the fact that the electromagnetic field is mathematically equivalent to a set of harmonic oscillators [*1]. We take Coulomb gage ($\mathrm{div}\boldsymbol{A}=\vec{0}$), and expand the vector potential $\boldsymbol{A}$ with the plane waves as follows.

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{r},t)&=\sum_{\boldsymbol{k},\lambda}(\boldsymbol{A}_{\boldsymbol{k}\lambda}e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega_{\boldsymbol{k}\lambda}t)}+\boldsymbol{A}^*_{\boldsymbol{k}\lambda}e^{-i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega_{\boldsymbol{k}\lambda}t)}),\\ &(\omega_{\boldsymbol{k}}=c|\boldsymbol{k}|,\quad \boldsymbol{A}^*_{\boldsymbol{k}\lambda}=\boldsymbol{A}_{-\boldsymbol{k}\lambda}).\end{aligned} \tag{4.11}$$

Here λ represents the freedom of polarization. From the selection of Coulomb gauge, the electromagnetic wave should be transverse and λ represents two-dimensional freedom. From the Maxwell equaiton $\boldsymbol{E}=\partial\boldsymbol{A}/\partial t$, $\boldsymbol{B}=\mathrm{rot}\boldsymbol{A}$, the energy of electromagnetic field $\mathcal{E}$ in volume V is written as

$$\mathcal{E}=\int_V[\epsilon_0\boldsymbol{E}^2(\boldsymbol{r},t)+\mu_0^{-1}\boldsymbol{B}^2(\boldsymbol{r},t)]\frac{d^3\boldsymbol{r}}{2}=2\epsilon_0V\sum_{\boldsymbol{k},\lambda}\omega_{\boldsymbol{k}\lambda}(\boldsymbol{A}_{\boldsymbol{k}\lambda}\cdot\boldsymbol{A}^*_{\boldsymbol{k}\lambda}), \tag{4.12}$$

because the terms with $\exp(\pm2i\boldsymbol{k}\cdot\boldsymbol{r})$ vanish with spatial integration.

Then we introduce variables(vectors) $\boldsymbol{Q}_{\boldsymbol{k}\lambda}$, $\boldsymbol{P}_{\boldsymbol{k}\lambda}$ as

$$\boldsymbol{Q}_{\boldsymbol{k}\lambda}=\sqrt{\epsilon_0V}(\boldsymbol{A}_{\boldsymbol{k}\lambda}e^{-i\omega_{\boldsymbol{k}\lambda}t}+A^*_{\boldsymbol{k}\lambda}e^{i\omega_{\boldsymbol{k}\lambda}t}),\quad \boldsymbol{P}_{\boldsymbol{k}\lambda}=d\boldsymbol{Q}_{\boldsymbol{k}\lambda}/dt=i\omega_{\boldsymbol{k}\lambda}\sqrt{\epsilon_0V}(-\boldsymbol{A}_{\boldsymbol{k}\lambda}e^{-i\omega_{\boldsymbol{k}\lambda}t}+A^*_{\boldsymbol{k}\lambda}e^{i\omega_{\boldsymbol{k}\lambda}t}). \tag{4.13}$$

$\mathcal{E}$ is expressed as

$$\mathcal{E}=\frac{1}{2}\sum_{\boldsymbol{k}\lambda}(\boldsymbol{P}^2_{\boldsymbol{k}\lambda}+\omega^2_{\boldsymbol{k}\lambda}\boldsymbol{Q}^2_{\boldsymbol{k}\lambda}), \tag{4.14}$$

which tells the electromagnetic field is described as a set of harmonic oscillators in $\boldsymbol{Q}$ space. Then the field can be quantized with changing $\boldsymbol{P}$, $\boldsymbol{Q}$ to operators and require the following commutation relations.

$$[\hat{Q}_{\boldsymbol{k}'\lambda'},\hat{P}_{\boldsymbol{k}\lambda}]=i\hbar\delta_{\boldsymbol{k}\boldsymbol{k}'}\delta_{\delta\delta'},\quad (\text{others})=0. \tag{4.15}$$

[*1] According to the literature[10], this is called "Jeans theorem." Actually, in ref.[9], that "theorem" is proven. The discussion then leads to the Rayleight-Jeans law. However, there is more famous "Jeans theorem", which is on the distribution of particles with gravitational interactions[9, 10]

The Hamiltonian is in the same form with (4.3).

$$\hat{H} = \frac{1}{2}\sum_{\boldsymbol{k}\lambda}(\hat{P}^2_{\boldsymbol{k}\lambda} + \omega^2_{\boldsymbol{k}}\hat{Q}^2_{\boldsymbol{k}\lambda}). \tag{4.16}$$

creation/annihilation operators, which corresponds to up/down operators, are

$$a^\dagger_{\boldsymbol{k}\lambda} = \frac{1}{\sqrt{2\hbar\omega_{\boldsymbol{k}}}}(\omega_{\boldsymbol{k}}\hat{Q}_{\boldsymbol{k}\lambda} - i\hat{P}_{\boldsymbol{k}\lambda}), \quad a_{\boldsymbol{k}\lambda} = \frac{1}{\sqrt{2\hbar\omega_{\boldsymbol{k}}}}(\omega_{\boldsymbol{k}}\hat{Q}_{\boldsymbol{k}\lambda} + i\hat{P}_{\boldsymbol{k}\lambda}). \tag{4.17}$$

From (4.15), the commutation relations

$$[a_{\boldsymbol{k}\lambda}, a^\dagger_{\boldsymbol{k}'\lambda'}] = \delta_{\boldsymbol{k}\boldsymbol{k}'}\delta_{\lambda\lambda'}, \quad (\text{others}) = 0 \tag{4.18}$$

are derived. Finally (4.16) can be quantized in the Hamiltonian form as

$$\hat{H} = \sum_{\boldsymbol{k}\lambda}\hbar\omega_{\boldsymbol{k}}\left(a^\dagger_{\boldsymbol{k}\lambda}a_{\boldsymbol{k}\lambda} + \frac{1}{2}\right). \tag{4.19}$$

The vector potential, for example, can also be written in the form of operator as

$$\hat{\boldsymbol{A}}(\boldsymbol{r}, t) = \sum_{\boldsymbol{k}\lambda}\sqrt{\frac{\hbar}{2\epsilon_0\omega_{\boldsymbol{k}}V}}\boldsymbol{e}_{\boldsymbol{k}\lambda}\left[a_{\boldsymbol{k}\lambda}e^{i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega_{\boldsymbol{k}}t)} + a^\dagger_{\boldsymbol{k}\lambda}e^{-i(\boldsymbol{k}\cdot\boldsymbol{r}-\omega_{\boldsymbol{k}}t)}\right]. \tag{4.20}$$

4.1.2 States of photons

Corresponding to eq.(4.6), the operator

$$\hat{n}_{\boldsymbol{k}\lambda} \equiv a^\dagger_{\boldsymbol{k}\lambda}a_{\boldsymbol{k}\lambda} \tag{4.21}$$

represents the revel of excitation in the mode $(\boldsymbol{k}, \lambda)$ from the ground state $|0\rangle_{\boldsymbol{k}\lambda}$. Single step of the excitation corresponds to the energy of $\hbar\omega_{\boldsymbol{k}}$, which is also the energy of single photon in the Einstein relation of photon quantum. Hence $\hat{n}_{\boldsymbol{k}\lambda}$ in (4.21) can be interpreted as **photon number operator**in mode $(\boldsymbol{k}, \lambda)$. As in (4.9), there are eigenfunctions of (4.19) in which the number of photons in mode $(\boldsymbol{k}, \lambda)$ is $n_{\boldsymbol{k}\lambda}$. We use the expression that the symbol $\{\cdots_\alpha\}$ represents the set of elements with α as the index. Then the state is represented as $|\{n_{\boldsymbol{k}\lambda}\}\rangle$.

From (4.9), we call the state described in the form

Number state

$$|\{n_{\boldsymbol{k}\lambda}\}\rangle = \left[\prod_{\boldsymbol{k}\lambda}\frac{(a^\dagger_{\boldsymbol{k}\lambda})^{n_{\boldsymbol{k}\lambda}}}{\sqrt{n_{\boldsymbol{k}\lambda}!}}\right]|0\rangle \tag{4.22}$$

as **number state** .

The expectation value of the energy of the number state is

$$\langle\{n_{\boldsymbol{k}\lambda}\}|\hat{H}|\{n_{\boldsymbol{k}\lambda}\}\rangle = \sum_{\boldsymbol{k}\lambda}\hbar\omega_{\boldsymbol{k}}\left(n_{\boldsymbol{k}\lambda} + \frac{1}{2}\right). \tag{4.23}$$

A state with multiple mode is a superpositon of eigenstates with different eigenvalues and is not an eigenstate of the total Hamiltonian. On the other hand, the expectation value of the electric field is from (4.20),

$$\langle\{n_{\boldsymbol{k}\lambda}\}|\hat{\boldsymbol{E}}|\{n_{\boldsymbol{k}\lambda}\}\rangle = -\langle\{n_{\boldsymbol{k}\lambda}\}|(\partial\hat{\boldsymbol{A}}/\partial t)|\{n_{\boldsymbol{k}\lambda}\}\rangle = 0. \tag{4.24}$$

That is the expectation value of the electric field is zero. This does not mean the time-average makes it zero. Even for the measurements in very short time, the average over many measurements is zero. Just the same for the magnetic field. On the other hand the quantum fluctuation in the electric field is

$$\langle\{n_{\boldsymbol{k}\lambda}\}|\hat{\boldsymbol{E}}^2|\{n_{\boldsymbol{k}\lambda}\}\rangle = \sum_{\boldsymbol{k}\lambda}\frac{\hbar\omega_{\boldsymbol{k}}}{\epsilon_0 V}\left(n_{\boldsymbol{k}\lambda} + \frac{1}{2}\right) = \frac{1}{\epsilon_0 V}\langle\{n_{\boldsymbol{k}\lambda}\}|H|\{n_{\boldsymbol{k}\lambda}\}\rangle, \tag{4.25}$$

which is non-zero. Furthermore, even for the photon number zero state, each mode has the fluctuation of $\hbar\omega_{\boldsymbol{k}}/(2\epsilon_0 V)$, which is called **zero-point motion**of electromagnetic field. The zero-point fluctuation corresponds to 1/2 in the energy expression of (4.23). This is very important property for the spontaneous emission of photon. The reason of using space for free electromagnetic field is to describe this clearly.

The properties of number state described above indicate that it is difficult to coherently superimpose the oscillating electromagnetic field of multiple photons to obtain the oscillating electromagnetic field as in the classical picture in the energy eigenstate where the number of photons is fixed. On the other hand, by superimposing several states, it is possible to create a state with a finite expected value of the electromagnetic field. For example, the number states for a single mode (hence for a while we omit writing the mode index as $|n\rangle$) can be summed up with Gaussian weight to get

Coherent state

$$|\alpha\rangle = \exp\left(-\frac{|\alpha|^2}{2}\right)\exp(\alpha a^\dagger)|0\rangle = \exp\left(-\frac{|\alpha|^2}{2}\right)\sum_{n=0}^{\infty}\frac{\alpha^n}{\sqrt{n!}}|n\rangle, \tag{4.26}$$

where α is a complex parameter. The state expressed as (4.26) is called **coherent state**. When the annihilation operator is applied, from $a|n\rangle = \sqrt{n}|n-1\rangle$,

$$a|\alpha\rangle = \alpha|\alpha\rangle, \tag{4.27}$$

that is the coherent state is the eigenstate of the annihilation operator with the eigenvalue of α. This means that the coherent state is a superposition of an infinite number of states, and even if quantum mechanical "measurement" is performed on the single photon in it, the whole state remains unchanged. If we measure the photon number in this state, the probability of detecting n-photons is

$$P(n) = \langle n|\alpha\rangle = \frac{e^{-|\alpha|^2}|\alpha|^{2n}}{n!}, \tag{4.28}$$

which is a Poissonian distribution. We write the complex parameter α in the amplitude and the phase as $\alpha = |\alpha|e^{i\phi}$. Then the expectation values of the electric field and the magnetic field are

$$\langle\alpha|\hat{\boldsymbol{E}}(\boldsymbol{r},t)|\alpha\rangle = -\sqrt{\frac{2\hbar\omega_{\boldsymbol{k}}}{\epsilon_0 V}}|\alpha|\boldsymbol{e}_{\boldsymbol{k}\lambda}\sin(\boldsymbol{k}\cdot\boldsymbol{r}-\omega_{\boldsymbol{k}}t+\phi), \tag{4.29a}$$

$$\langle\alpha|\hat{\boldsymbol{B}}(\boldsymbol{r},t)|\alpha\rangle = -\sqrt{\frac{2\hbar}{\epsilon_0\omega_{\boldsymbol{k}}V}}|\alpha|\boldsymbol{k}\times\boldsymbol{e}_{\boldsymbol{k}\lambda}\sin(\boldsymbol{k}\cdot\boldsymbol{r}-\omega_{\boldsymbol{k}}t+\phi). \tag{4.29b}$$

This means classical electromagnetic wave is reproduced in the coherent state.

4.1.3 Basic optical processes in two-level systems

The **two-level system** composed of two qunatum states is also called qubit in the field of quantum information and is the most basic quantum system. As in Fig. 4.1, we consider a two-level electronics system of $(|a\rangle, |b\rangle)$ with the energy eigenvalues (E_a, E_b). We take these

$$\mathscr{H}_0|a\rangle = E_a|a\rangle, \quad \mathscr{H}_0|b\rangle = E_b|b\rangle \tag{4.30}$$

as the basis and the general state can be written as

$$\psi(t) = c_a(t)e^{-E_a t/\hbar}|a\rangle + c_b(t)e^{-E_b t/\hbar}|b\rangle. \tag{4.31}$$

In Fig. 4.1, three basic optical processes in the two-level system are illustrated. (a) is the **optical absorption**, in which the electron absorbs the photon energy and makes transition of $|a\rangle \to |b\rangle$. (b) is the **spontaneous emission** of photon associated with the transition $|b\rangle \to |a\rangle$ of the electron initially excited to $|b\rangle$. (c) is the **slimulated emission**, in which the first photon comes to the excited state $|b\rangle$ to sitimulate the emission of the second photon coherent to the first one.

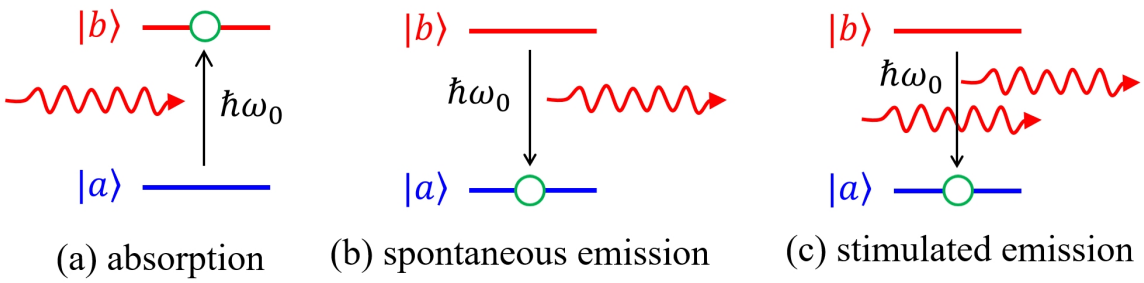


Fig. 4.1 Three basic optical processes in two-level systems (a) optical absorption, (b) spontaneous emission of a photon, (c) stimulated emission of a photon

We write the Hamiltonian of the system with electromagnetic field in non-relativistic approximation as

$$\mathscr{H}_{\rm op} = \frac{(\boldsymbol{p} + e\boldsymbol{A})^2}{2m} + V(\boldsymbol{r}), \tag{4.32}$$

where $\boldsymbol{A}$ is the vector potential and we treat it as perturbation. We then drop the term of $\boldsymbol{A}^2$ and the perturbation Hamiltonian $\mathscr{H}'$ can be defined as

$$\mathscr{H}_{\rm op} \approx \mathscr{H}_0 + \frac{e}{m}\boldsymbol{A}\cdot\boldsymbol{p} \equiv \mathscr{H}_0 + \mathscr{H}'. \tag{4.33}$$

For simplicity, we assume $\mathscr{H}'$ does not have the diagnal terms.

$$\langle a|\mathscr{H}'|a\rangle = \langle b|\mathscr{H}'|b\rangle = 0. \tag{4.34}$$

We consider the case that a plane electromagnetic wave is applied to the two-level system, which wave is described in Coulomb gauge ($\mathrm{div}\boldsymbol{A} = 0$) as

$$\boldsymbol{A} = A_0\boldsymbol{e}_{\rm p}\cos(\boldsymbol{k}_{\rm p}\cdot\boldsymbol{r} - \omega t). \tag{4.35}$$

As we saw in the previous section, this means a coherent state comes to the two-level system. The perturbation Hamiltonian is

$$\mathscr{H}' = \frac{eA_0}{m}\boldsymbol{e}_{\rm p}\cdot\hat{\boldsymbol{p}}\cos(\boldsymbol{k}_{\rm p}\cdot\boldsymbol{r} - \omega t). \tag{4.36}$$

This approximation is called **dipole approximation** from the following reason. The matrix element of $\mathscr{H}'$ for $|a\rangle \to |b\rangle$ is with writing $A_0\cos(\boldsymbol{k}_{\rm p}\cdot\boldsymbol{r} - \omega t)$ as A,

$$\frac{eA}{m}\boldsymbol{e}_{\rm p}\cdot\langle b|\hat{\boldsymbol{p}}|a\rangle = \frac{eA}{m}\langle b|\boldsymbol{e}_{\rm p}\cdot\frac{m}{i\hbar}[\hat{\boldsymbol{r}}, \mathscr{H}_0]|a\rangle = \frac{iA}{\hbar}(E_b - E_a)\boldsymbol{e}_{\rm p}\cdot\langle b|(-e)\hat{\boldsymbol{r}}|a\rangle. \tag{4.37}$$

The last term is the transition element of the electric dipole momen operator and the transition by the Hamiltonian (4.36) is called dipole transition.

Substituting (4.30) to the Schrödinger equation $i\hbar\partial\psi/\partial t = (\mathscr{H}_0 + \mathscr{H}')\psi$, we obtain

$$i\hbar\left[\frac{dc_a}{dt}|a\rangle e^{-iE_at/\hbar} + \frac{dc_b}{dt}|b\rangle e^{-iE_bt/\hbar}\right] = c_a\mathscr{H}'|a\rangle e^{-iE_at/\hbar} + c_b\mathscr{H}'|b\rangle e^{-iE_bt/\hbar}. \tag{4.38}$$

Taking inner products with $\langle a|$ and $\langle b|$ leads to the following simultaneous differential equations for (c_a, c_b).

$$\begin{cases} \dfrac{dc_a}{dt} = -\dfrac{i}{\hbar}c_b\langle a|\mathscr{H}'|b\rangle e^{-i\omega_0 t}, \\ \dfrac{dc_b}{dt} = -\dfrac{i}{\hbar}c_a\langle b|\mathscr{H}'|a\rangle e^{i\omega_0 t}, \end{cases} \qquad \omega_0 \equiv \frac{E_b - E_a}{\hbar}. \tag{4.39}$$

4.1.4 Optical absorption, emission

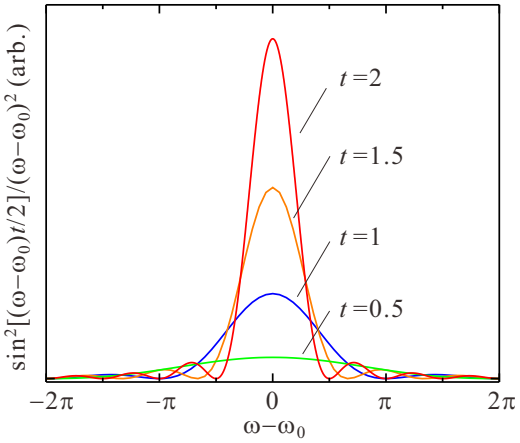
Frist we consider the optical absorption in (a). We take $c_a(0) = 1$, $c_b(0) = 0$ as the initial condition and $c_a^{(1)}(t) = 1$ as the starting point of the sequential substitution method to get the approximate solution.

$$c_a^{(1)}(t) = 1,$$

$$c_b^{(1)}(t) = -\frac{i}{\hbar}\int_0^t \langle b|\mathscr{H}'|a\rangle(t')e^{i\omega_0 t'}dt', \tag{4.40a}$$

$$c_a^{(2)}(t) = 1 - \frac{1}{\hbar^2}\int_0^t dt'\langle a|\mathscr{H}'|b\rangle(t')e^{-i\omega_0 t'}\left[\int_0^{t'} dt''\langle b|\mathscr{H}'|a\rangle(t'')e^{i\omega_0 t''}\right]. \tag{4.40b}$$

The expression $\langle b|\mathscr{H}'|a\rangle(t)$ is to clarify that $\langle b|\mathscr{H}'|a\rangle$ is a function of t.



To consider $\langle b|\mathscr{H}'|a\rangle(t)$, we drop the term of photon wavenumber $\boldsymbol{k}_\mathrm{p}$ from (4.36) because of the following reason. We consider the process in which the electron wavenumber varies as $0 \to k_\mathrm{e}$ to enhance the kinetic energy by $\hbar^2 k_\mathrm{e}^2/2m_0$. If we consider a photon with the energy that matches to this energy enhancement, the photon wavenumber k_p is obtained from $k_\mathrm{e} = \sqrt{2k_\mathrm{ec}k_\mathrm{p}}$. Here $k_\mathrm{ec} = m_0 c/\hbar$ is the product of electron rest mass m_0 and the speed of light c, corresponding to a very large momentum. Then $k_\mathrm{e} \gg k_\mathrm{p}$ and we can ignore the contribution of photon momentum.

Now we write

$$V_{ba} \equiv \langle b|\frac{eA_0}{m}\boldsymbol{e}_\mathrm{p}\cdot\hat{\boldsymbol{p}}|a\rangle \tag{4.41}$$

and from (4.40a),

$$c_b(t) \simeq -\frac{i}{\hbar}V_{ba}\int_0^t dt'\cos\omega t' e^{i\omega_0 t'} = -\frac{V_{ba}}{2\hbar}\left[\frac{e^{i(\omega_0+\omega)t}-1}{\omega_0+\omega} + \frac{e^{i(\omega_0-\omega)t}-1}{\omega_0-\omega}\right]$$
$$\simeq -i\frac{V_{ba}}{\hbar}\frac{\sin[(\omega_0-\omega)t/2]}{\omega_0-\omega}e^{i(\omega_0-\omega)t/2}. \tag{4.42}$$

Then if we apply the oscillating electromagnetic field from $t = 0$, the probability amplitude of $|b\rangle$ at time t is

$$P_b(t) = |c_b(t)|^2 \simeq \frac{|V_{ba}|^2}{\hbar^2}\frac{t}{2}\frac{\sin^2[(\omega_0-\omega)t/2]}{(\omega_0-\omega)^2(t/2)} \tag{4.43}$$

As is well known the last factor goes to a delta function in the limit of $t \to \infty$ ($\lim_{t\to\infty}\sin^2[(\omega_0-\omega)t/2]/(\omega_0-\omega)^2(t/2) = \pi\delta(\omega-\omega_0)$). Actually in the plot of $\sin^2[(\omega_0-\omega)t/2]/(\omega_0-\omega)^2$, the peak at $\omega-\omega_0 = 0$ grows high and sharp with the increase of t. Hence the factor represents the energy conservation.

When we take the initial condition as $c_b(0) = 1$, $c_a(0) = 0$, that is, the excited state $|b\rangle$ with photon field of ω, the transition $|b\rangle \to |a\rangle$, which is the reversed process of the optical absorption, occurs with the emission of a photon of ω_0. This emitted photon is coherent to the existing coherent photon state[11]. This is the stimulated emission in (c).

Then how we treat spontaneous emission in (b) without photon field outside? As we saw in Sec. 4.1.1, even in the vacuum with zero number of photons, the zero-point quantum fluctuation of electromagnetic field exists for each mode. In the spontaneous emission process, the photon emission is "stimulated" by these zero-point fluctuations. Most of the light emission from semiconductor devices (other than lasers) is by this spontaneous emission. In that sense, we are looking at zero-point quantum fluctuation when we are facing electric displays.

4.1.5 Rabi oscillation

In the perturbation term $V_{ba}\cos(\omega t)$, the cosine part can be expressed as $(e^{i\omega t}+e^{-i\omega t})/2$, which means the decomposition into two terms rotation on the complex plane with the angular frequency of $\pm\omega$. Among them, the component important for the transition is with $\omega\sim\omega_0$ and the term of rotation with $-\omega$ has the frequency very far from ω_0 and can be ignored as

$$\langle a|\mathscr{H}'|b\rangle = \frac{V_{ab}}{2}e^{i\omega t}. \tag{4.44}$$

This kind of approximation is called **rotation wave approximation**, in which the simultaneous differential equations are

$$\begin{cases} \dfrac{dc_a}{dt} = -\dfrac{i}{2\hbar}c_b V_{ab}e^{-i(\omega_0-\omega)t}, \\ \dfrac{dc_b}{dt} = -\dfrac{i}{2\hbar}c_a V_{ba}e^{i(\omega_0-\omega)t}. \end{cases} \tag{4.45}$$

This can be written into a differential equation for a single variable as

$$\frac{d^2c_b}{dt^2} + i(\omega-\omega_0)\frac{dc_a}{dt} + \frac{|V_{ab}|^2}{(2\hbar)^2} = 0. \tag{4.46}$$

The solution is obtained straightforwardly as

$$c_b(t) = c_+e^{i\lambda_+ t} + c_-e^{\lambda_- t}, \quad \lambda_\pm \equiv \frac{1}{2}(\delta \pm \sqrt{\delta^2+|V_{ab}|^2/\hbar^2}), \quad \delta\equiv\omega_0-\omega. \tag{4.47}$$

Under the initial condition $|c_a(0)|=1$, $c_b(0)=0$,

$$\begin{cases} c_b(t) = \dfrac{i|V_{ab}|}{\omega_{\mathrm{R}}\hbar}e^{i\delta t/2}\sin(\omega_{\mathrm{R}}t/2), & \text{(4.48a)} \\ c_a(t) = e^{i\delta t/2}\left[\cos\left(\dfrac{\omega_{\mathrm{R}}t}{2}\right) - i\dfrac{\delta}{\omega_{\mathrm{R}}}\sin\left(\dfrac{\omega_{\mathrm{R}}t}{2}\right)\right] & \text{(4.48b)} \end{cases}$$

are obtained where

$$\omega_{\mathrm{R}} \equiv \sqrt{\delta^2+|V_{ab}|^2/\hbar^2} \tag{4.49}$$

is called **Rabi frequency**.

The oscillation between the two-levels caused by the electromagnetic wave (photon) with energy close to the energy difference between the two levels is called **Rabi oscillation**. When the photon energy is tuned to the energy difference ($\delta=0$), the Rabi frequency is proportional to the magnitude of the electromagnetic irradiation. The Rabi oscillation is widely used in various resonance phenomena utilized to get information inside materials, or quantum information processing, etc.

4.1.6 Oscillator strength, selection rule

For the one-dimensional harmonic oscillator, which we consider in the beginning of this chapter, from (4.2), (4.4), we can write

$$\hat{x} = \sqrt{\frac{\hbar}{2m\omega_{\mathrm{h}}}}(a+a^\dagger), \tag{4.50}$$

which leads to the dipole transition elements of (4.37), corresponding to $|0\rangle\to|1\rangle$ is

$$\langle 1|(-e)\hat{x}|0\rangle = -e\sqrt{\frac{\hbar}{2m\omega_{\mathrm{h}}}} \equiv \mu_{10}. \tag{4.51}$$

The probability of the dipole transition $|a\rangle \to |b\rangle$ is indicated by the transition dipole moment;

$$\boldsymbol{\mu}_{ba} \equiv \langle b|(-e)\hat{\boldsymbol{r}}|a\rangle. \tag{4.52}$$

Then for the "unit" of the strength, we take the transition dipole moment μ_{10} for the one-dimensional harmonic oscillator with the characteristic frequency $\omega_{\rm h} = (E_b - E_a)/\hbar$. For the probability we need to take the square of the absolute value, we define **oscillator strength** as

$$f_{ba} = \frac{|\mu_{ba}|^2}{|\mu_{10}|^2} = \frac{2m\omega_{ba}}{e^2\hbar}|\mu_{ba}|^2. \tag{4.53}$$

The character f is commonly used and we also call it as "f-value."

When there are multiple possible final states $|b\rangle$, we use b as the index of all such states. Then the f-values sutisfy the following **sum rule**.

$$\sum_b f_{ba} = 1. \tag{4.54}$$

For the system with N-electrons, the right hand side is N.

When the system has multiple directional oscillators with random directions, the effective transition dipole mement $\langle\mu_{\rm eff}\rangle$ is given by taking the avarage as $\mu_{ba}/3$. Then the oscillator strength is expressed as

$$f'_{ba} = \frac{2m\omega_{ba}}{3e^2\hbar}|\mu_{ba}|^2. \tag{4.55}$$

When the system has spatial inversion symmetry, the eigenstates of the Hamiltonian should have the parity for the spatial inversion operation. That is, for an eigenstate $\phi_n(\boldsymbol{r})$, the following should hold.

$$\phi_n(-\boldsymbol{r}) = \pm\phi_n(\boldsymbol{r}). \tag{4.56}$$

$+$, $-$ correspond to even and odd parity respectively. In the expression (4.52), $\boldsymbol{r}$ has the odd parity then if $|a\rangle$, $|b\rangle$ have the same parity, the integration gives zero for μ_{ba} and the dipole transition is forbidden. As above, the rule that dominates the possibility of a transition along with symmetry, quantum number etc. is called **selection rule**.

4.2 Interband transition and optical response

So far we have seen very basic knowledges on the optical response of two-level systems. We now expand the concepts and the discussions to the electronic states in solids, in which both ground states and excited states are extended over the crystals.

4.2.1 Absorption of light with interband transition

Materials absorb electromagnetic wave in various ways. Free carrier absorption, impurity absorption, absorption by lattice vibration, etc. though the main absorption used in the optical devices is the absorption due to the interband transition of electrons. Thus in this sub-section, we will see the very basics of the interband transition absorption.

For simplicity, we write a plane electromagnetic wave with a linear polarization propagating along z axis with vector potential $\boldsymbol{A}$ as

$$\boldsymbol{A} = A_0\boldsymbol{e}\exp[i(\boldsymbol{k}_p\cdot\boldsymbol{r} - \omega t)]. \tag{4.57}$$

The wavenumber $\boldsymbol{k}_p$ is $(0,0,k_p)$, $\boldsymbol{e}$ is the polarizaiton vector and we put $\boldsymbol{e}_x = (1,0,0)$. The electric field $\boldsymbol{E} = -\partial\boldsymbol{A}/\partial t$, the magnetic field $\boldsymbol{H} = \mu^{-1}{\rm rot}\boldsymbol{A}$ (μ is the permeability of the medium), then the energy flow density (Poynting vector) is

$$\boldsymbol{I} = \langle\boldsymbol{E}\times\boldsymbol{H}\rangle = \frac{\epsilon_0 c\bar{n}\omega^2 A_0^2}{2}\boldsymbol{e}_z. \tag{4.58}$$

$\bar{n}$ is the refractive index (light speed in the medium is $c' = 1/\sqrt{\epsilon_1\epsilon_0\mu_1\mu_0}$ (ϵ_1, μ_1 are the ratio of dielectric constant and that of magnetic permeability to those of vacuum) $\bar{n} = c/c' = \sqrt{\epsilon_1\mu_1}$), $\boldsymbol{e}_z = (0,0,1)$.

The absorption of light causes the exponential damping of the intensity $|\boldsymbol{I}|$ as $I(z) = I_0\exp(-\alpha z)$. The damping constant α is the **absorption coefficient**. From this definition $\alpha = -dI/Idz = -dI/Ic'dt$. Thus if we assign the averaged number of photons absorbed in the unit time and the unit volume as W, then the decreasing rate of I is written as $\hbar\omega W$ giving

$$\alpha = \frac{\hbar\omega W}{I} = \frac{2\hbar\omega W}{\epsilon_0 c\bar{n}\omega^2 A_0^2}. \tag{4.59}$$

Among the various absorption mechanisms, that caused by a valence electron absorbing a photon and being excited to the conduction band, is called **fundamental absorption**. The fundamental absorption begins just above the band gap. The absorption just at the band gap is called "band edge absorption".

We write the Hamiltonian of the system as $\mathscr{H} = (\boldsymbol{p}+e\boldsymbol{A})^2/2m_0 + V(\boldsymbol{r})$ and treat $\boldsymbol{A}$ as a perturbation. With ignoring $\boldsymbol{A}^2$, $\mathscr{H} = \mathscr{H}_0 + (e/m_0)\boldsymbol{A}\cdot\boldsymbol{p}$. Bloch functions in conduction band and valence band are written as $|c\boldsymbol{k}\rangle = u_{c\boldsymbol{k}}e^{i\boldsymbol{kr}}$, $|v\boldsymbol{k}\rangle = u_{v\boldsymbol{k}}e^{i\boldsymbol{kr}}$ respectively and the perturbation term causes the transition from the valence band to the conduction band with the probability $W_{\rm vc}$ per unit volume in the Fermi's golden rule approximation as

$$W_{\rm vc} = \frac{2\pi e}{\hbar m_0}|\langle c\boldsymbol{k}|\boldsymbol{A}\cdot\boldsymbol{p}|v\boldsymbol{k}'\rangle|^2\delta(E_c(\boldsymbol{k}) - E_v(\boldsymbol{k}') - \hbar\omega) = \frac{\pi e^2}{2\hbar m_0^2}A_0^2|M|^2\delta(E_c(\boldsymbol{k}) - E_v(\boldsymbol{k}') - \hbar\omega), \tag{4.60}$$

$$\begin{aligned} M = \int_V \frac{d^3r}{V}e^{i(\boldsymbol{k}_p+\boldsymbol{k}'-\boldsymbol{k})\cdot\boldsymbol{r}}u^*_{c\boldsymbol{k}}(\boldsymbol{r})\boldsymbol{e}\cdot(\boldsymbol{p}+\hbar\boldsymbol{k}')u_{v\boldsymbol{k}'}(\boldsymbol{r}) &= \frac{\sum_l e^{i(\boldsymbol{k}_p+\boldsymbol{k}'-\boldsymbol{k})\cdot\boldsymbol{R}_l}}{V}\int_\Omega d^3r u^*_{c\boldsymbol{k}}(\boldsymbol{r})\boldsymbol{e}\cdot(\boldsymbol{p}+\hbar\boldsymbol{k}')u_{v\boldsymbol{k}'}(\boldsymbol{r}) \\ &= \frac{N}{V}\delta_{\boldsymbol{k}_p+\boldsymbol{k}'-\boldsymbol{k},\boldsymbol{K}}\int_\Omega d^3r u^*_{c\boldsymbol{k}}(\boldsymbol{r})\boldsymbol{e}\cdot(\boldsymbol{p}+\hbar\boldsymbol{k}')u_{v\boldsymbol{k}'}(\boldsymbol{r}). \end{aligned} \tag{4.61}$$

Here, l is the label of lattice points, V, Ω are the volumes of the system and the unit cell respectively. $\boldsymbol{K}$ is a reciprocal lattice vector, $\boldsymbol{k}_p$ a photon wavenumber, N the total number of the lattice points, $N\Omega = V$.

In eq.(4.61) we implicitly consider a direct excitation of an electron by the electromagnetic field of a photon. Such a transition is called a **direct transition**. The necessary condition for the momentum conservation in a fundamental absorption is $\boldsymbol{k}_p + \boldsymbol{k}' - \boldsymbol{k} = \boldsymbol{K}$, though practically from common values of band gaps, effective masses and lattice constants, it turns to be $\boldsymbol{K} = \boldsymbol{0}$. Also within the dipole transition approximation, $\boldsymbol{k}_p$ can be ignored and we can put

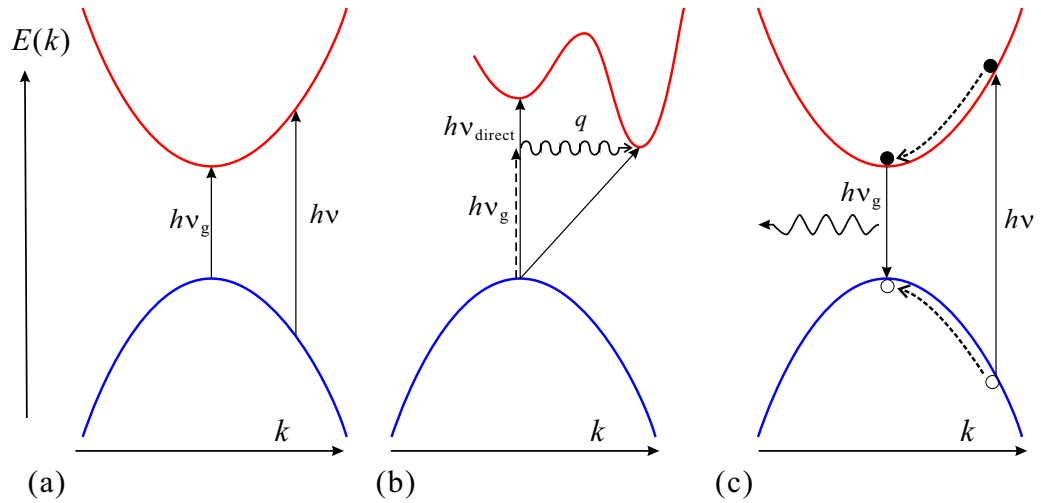


Fig. 4.2 Illustrations of optical response due to the interband transition of an electron. (a) Optical absorption with direct interband transition. (b) Optical absorption with indirect interband transition. (c) Photoluminescence, in which optically excited electron-hole pair recombine for the light emission.

$\boldsymbol{k} = \boldsymbol{k}'$. $u_{c\boldsymbol{k}}(\boldsymbol{r})$, $u_{v\boldsymbol{k}}(\boldsymbol{r})$ belong to different eigenvalues hence the term of $\hbar\boldsymbol{k}'$ in (4.61) vanishes giving

$$M = \int_\Omega \frac{d^3r}{\Omega} u^*_{c\boldsymbol{k}}(\boldsymbol{r}) \boldsymbol{e} \cdot \boldsymbol{p} u_{v\boldsymbol{k}}(\boldsymbol{r}). \tag{4.62}$$

From (4.59) and (4.62), we assume $\boldsymbol{k}$-dependence of M is weak and obtain the expression for the absorption coefficient for direct transition as

$$\alpha_{\rm da} = \frac{\pi e^2}{\bar{n}\epsilon_0 \omega c m_0^2} |M|^2 \sum_{\boldsymbol{k}} \delta(E_c(\boldsymbol{k}) - E_v(\boldsymbol{k}) - \hbar\omega). \tag{4.63}$$

The summation part on $\boldsymbol{k}$ is called **joint density of states**. Let us write it as $J_{cv}(\hbar\omega)$ and $E_c(\boldsymbol{k}) - E_v(\boldsymbol{k})$ as $E_{cv}(\boldsymbol{k})$, and turn the summation on $\boldsymbol{k}$ in an integral form to get

$$J_{cv}(\hbar\omega) = \sum_{\boldsymbol{k}} \delta(E_{cv}(\boldsymbol{k}) - \hbar\omega) = 2 \int \frac{d^3k}{(2\pi)^3} \delta(E_{cv}(\boldsymbol{k}) - \hbar\omega). \tag{4.64}$$

We transform the integral in $\boldsymbol{k}$-space into that on the infinitesimal area dS on an equi-energy surface and on the energy E_{cv}. By writing the k-component perpendicular to the equi-energy surface as $k_\perp$, the integration can be transformed into

$$\begin{aligned} d^3k = dS dk_\perp = dS \frac{dk_\perp}{dE_{cv}} dE_{cv} = dS |\boldsymbol{\nabla}_{\boldsymbol{k}} E_{cv}|^{-1} dE_{cv}, \\ \therefore J_{cv}(\hbar\omega) = \frac{2}{(2\pi)^3} \int \frac{dS}{|\boldsymbol{\nabla}_{\boldsymbol{k}} E_{cv}(\boldsymbol{k})|_{E_{cv}=\hbar\omega}}. \end{aligned} \tag{4.65}$$

From the above we see that we have absorption anomalies around the points where the integrand of (4.65) vanishes. Let us consider the case of a direct gap semiconductor as illustrated in Fig. **??**(a), and assume $E_{cv} = E_{\rm g}$, $\boldsymbol{\nabla}_{\boldsymbol{k}} E_{cv} = \boldsymbol{0}$ at $\boldsymbol{k} = \boldsymbol{k}_0$. In the expansion of E_{cv} around $\boldsymbol{k}_0$, the first order term is zero and taking the second order term we get

$$E_{cv}(\boldsymbol{k}) = E_{\rm g} + \sum_i \frac{\hbar^2}{2\xi_i} (k_i - k_{i0})^2. \tag{4.66}$$

For simplicity let $\xi_i > 0 (i = 1, 2, 3)$. With variable translation $(\hbar/(2\xi_i)^{1/2})(k_i - k_{i0}) = s_i$,

$$E_{cv} = E_{\rm g} + \sum_i s_i^2 \equiv E_{\rm g} + s^2, \;\; d^3k = \frac{\sqrt{8\xi_1\xi_2\xi_3}}{\hbar^3} ds_1 ds_2 ds_3,$$

We also consider the integration in $\boldsymbol{s}$-space with that on equi-energy surfaces and on the energy. Because $|\boldsymbol{\nabla}_{\boldsymbol{s}} E_{cv}| = 2s$,

$$J_{cv} = \frac{2}{(2\pi)^3} \frac{\sqrt{8\xi_1\xi_2\xi_3}}{\hbar^3} \int \frac{dS}{2s} = \frac{1}{2\pi^2} \frac{\sqrt{8\xi_1\xi_2\xi_3}}{\hbar^3} \sqrt{\hbar\omega - E_{\rm g}} = \frac{\sqrt{2}}{\pi^2} \frac{m_r^{3/2}}{\hbar^3} \sqrt{\hbar\omega - E_{\rm g}}. \tag{4.67}$$

The calculation in the last line is for a direct gap semiconductor as illustrated in Fig. **??**(a), based on the assumption of isotropic effective mass at the band edge, *i.e.*, $\forall i \;\; \xi_i = m_r$. From $m_r^{-1} = m_e^{*-1} + m_h^{*-1}$ this is the reduced mass for an electron-hole pair. After all, (4.67) is the density of states in a three dimensional $\boldsymbol{k}$ space, that is just a re-calculation of the density of states in a three dimensional system (2.14). In this case, from the expression of the absorption coefficient in a direct transition (4.63),

$$\alpha(\hbar\omega) = \frac{e^2 (2m_r)^{3/2} |M|^2}{2\pi\epsilon_0 m_0^2 \bar{n}\omega c\hbar^3} \sqrt{\hbar\omega - E_{\rm g}}, \tag{4.68}$$

is obtained. The factor in the left hand side other than the joint density of states

$$f_{vc} = \frac{2|M|^2}{m_0 \hbar\omega}, \tag{4.69}$$

in which $|M|^2/\omega$ is representing the strength of the transition. The dimensionless quantity f_{vc} is called **oscillator strength**.

Appendix 4A: Rate of stimulated emission, spontaneous emission

Here we consider many identical two-level systems with states $(|a\rangle, |b\rangle)$. They are placed in the electromagnetic field with the energy density spectrum $U(\omega)$, where ω is the angular frequency. There is no direction interaction between the two-level systems while they are in equilibrium with the electromagnetic field, which is in thermal equilibrium, *i.e.* has the energy distribution of Planck law of radiation, and the momentum distributes isotopically. The rate of optical absorption (frequency per unit time) for $\omega \sim \omega_0$ is obtained from (4.37) as

$$|\langle b|\mathscr{H}'|a\rangle| = |E_0 \boldsymbol{e}_{\rm p} \cdot \langle b|(-e)\hat{\boldsymbol{r}}|a\rangle| = |E_0 \boldsymbol{e}_{\rm p} \cdot \boldsymbol{\mu}_{ba}|, \tag{4A.1}$$

where $E_0 = \omega A_0$ is the amplitude of the oscillation in electric field. From(4.43), the absorption probability is proportional to the square of the above, hence to $E_0^2 \propto U$. In the form of equation the absorption rate W_{ba} is written as

$$W_{ba} = B_{ba} U(\omega), \tag{4A.2}$$

with B_{ba} a coefficiant. We write the emission rate as the sum of the rate for spontaneous emission, which is independent of U and the rate for the stimulated emission, which is proportional to U.

$$W_{ab} = A + B_{ab} U((\omega). \tag{4A.3}$$

As seen in Sec.4.1.4, the optical absorption and the stimulated emission are in the relation of reversed process,

$$B_{ba} = B_{ab} \equiv B. \tag{4A.4}$$

We write $E_b - E_a = \hbar\omega$, and let the numbers of the two-level systems in the states a, b as N_a, N_b respectively. Then

$$N_b = N_a \exp\left(-\frac{\hbar\omega}{k_{\rm B}T}\right). \tag{4A.5}$$

Because the system is in equilibrium, the event frequencies of emissions and absorption should be the same, *i.e.*

$$BUN_a = (A + BU)N_b. \tag{4A.6}$$

These leads to the following expression for $U(\omega)$.

$$U(\omega) = \frac{A}{B} \frac{1}{\exp(\hbar\omega/k_{\rm B}T) - 1}. \tag{4A.7}$$

We request this to be equivalent to the Planck law of radiation

$$U(\omega) = \frac{\hbar\omega^3}{\pi^2 c^3} \frac{1}{\exp(\hbar\omega/k_{\rm B}T) - 1}, \tag{4A.8}$$

and obtain

$$\frac{A}{B} = \frac{\hbar\omega^3}{\pi^2 c^3}. \tag{4A.9}$$

These coefficient A, B are called **Einstein A coefficient, B coefficient** respectively.

In the discussion of transition probability (4.43), we have considered photons with single energy $\hbar\omega$. Now we consider a finite width $\delta\omega$ of the energy distribution around ω_0 with the photon density (4A.8). We write the electric field amplitude for ω_0 as E_0, then the energy density is $\epsilon_0 E_0^2/2$ [*2] .

$$\epsilon_0 \frac{E_0^2}{2} = \int_{\omega_0 - \delta\omega/2}^{\omega_0 + \delta\omega/2} U(\omega) d\omega. \tag{4A.10}$$

[*2] From (**??**), the energy of oscillating electromagnetic field is $\langle(\epsilon_0 E^2 + B^2/\mu_0)/2\rangle = \epsilon_0 \langle(E^2)\rangle$, and then the time average gives $\epsilon_0 E_0^2/2$.

And taking the directional average, we obtain

$$\left\langle |\boldsymbol{\mu}_{ab} \cdot \boldsymbol{e}_{\rm p}|^2 \right\rangle = \left\langle \mu_{12}^2 \cos^2\theta \right\rangle = \frac{\mu_{12}^2}{3}. \tag{4A.11}$$

The the transition probability (4.43) can be approximated as

$$|c_b(t)|^2 \simeq \frac{|\boldsymbol{\mu}_{ab}|^2}{3\hbar^2}\frac{1}{\epsilon_0}\int_{\omega_0-\delta\omega/2}^{\omega_0+\delta\omega/2} U(\omega)\frac{\sin^2[(\omega-\omega_0)t/2]}{(\omega-\omega_0)^2}d\omega \approx \frac{\pi\mu_{ab}^2}{3\epsilon_0\hbar^2}U(\omega_0)t. \tag{4A.12}$$

We replace the integral over the period $\delta\omega$ with the infinite integration. And we applied the identity $\lim_{\lambda\to\infty}\sin^2\lambda x/\lambda x^2 = \pi\delta(x)$. The transition probability is obtained as $|c_b(t)|^2/t$. Then the discussion leads to the expression of B coefficient as

$$B = \frac{\pi\mu_{ab}^2}{3\epsilon_0\hbar^2} = \frac{\pi e^2}{6\epsilon_0 m\hbar\omega_0}f_{ba}. \tag{4A.13}$$

If we use the frequency spectrum $\rho(\nu)$ ($2\pi\nu = \omega$) instead of the angular frequency spectrum $U(\omega)$, the expression needs correction of 2π, of course.

References

[1] N. F. Mott, “Metal-Insulator Transisions” (CRC Press, 1990); 和訳 「金属と非金属の物理」小野嘉之，大槻東巳 (丸善, 1996).

[2] 小野嘉之 「金属絶縁体転移」(朝倉書店，2002).

[3] 大槻東巳 「不規則電子系の金属-絶縁体転移」(現代物理最前線 (2) 共立出版，2000).

[4] 米沢富美子 「金属-非金属転移の物理」(朝倉書店，2012).

[5] D. Stauffer and A. Aharony, “Introduction to Percolation Theory” (2nd ed., Taylor & Francis, 2018); 和訳 「パーコレーションの基本原理」小田垣 孝 (吉岡書店, 2001).

[6] M. Imada, A. Fujimori, and Y. Tokura, Rev. Mod. Phys. **70**, 1039 (1998).

[7] D. C. Reynolds, “Excitons: Their Properties and Uses” (Academic Press, 1981).

[8] 本格的に学ぶには例えば，R. Loudon, “The Quanutm Theory of Light” (3rd ed., Oxford, 2000); P. Meystre and M. Sargent III, “Elements of Quantum Optics” (Springer, 1990); 松岡正浩 「量子光学」 (裳華房, 2000) など.

[9] J. H. Jeans, Phil. Mag. **10**, 91 (1905).

[10] 太田浩一 「マクスウェル理論の基礎」(東京大学出版会，2002).

[11] 霜田光一「レーザー物理入門」

4.2.2 Luminescence with interband transition

There are numerous types of semiconductor light emission. A typical example is light emission due to the recombination(pair annihilation) of electron-hole pairs. Minority carriers excited by various methods, including the above light absorption, emit their energy as photons by **radiative recombination** with the majority carriers. When the electron-hole pair does not emit a photon and the energy is dissipated to other freedoms, the process is called **non-radiative recombination**. Such emission of photons by radiative recombination is called **luminescence**. Amon them the ones with comparatively short lifetime are called **fluorescence** while those with very long lifetime is called **phosphorescence**. Luminescence is also classified with the origin of the electron-hole pair creation. The photon-absorption originated emission is called **photoluminescence**, the electrically stimulated emission (electric field activation of recombination center, injection of minority carriers, etc.) is called **electroluminescence**. By some reason minority carriers are trapped in impurities and some heat pulses cause release of them and lead to luminescence, which phenomenon is called **thermoluminescence**[*1].

As we saw in the previous section, there are two kinds of photon-emission, **stimulated emission** and **spontaneous emission**. In the former the emission probability is proportional to the photon density in the surrounding space while in the latter the probability is independent of that. If we include the zero-point fluctuation into the photon density, there is no difference in these two. In practice, however because the former is significant under limited conditions, causing peculiar phenomena like laser light emission etc., we usually discuss these separately. As this indicates, the density of photons is a very important factor in the treatment of light emission.

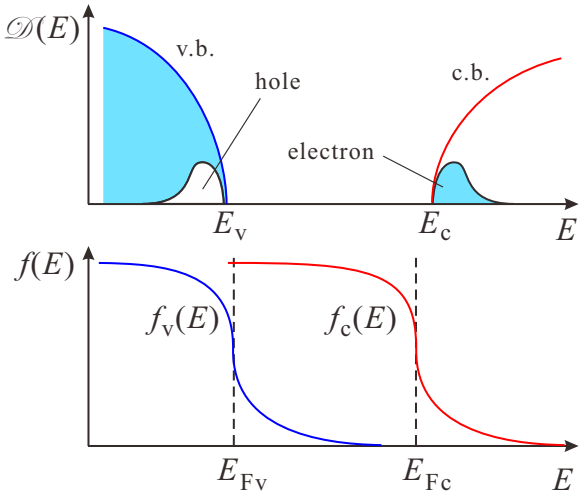


Fig. 4.3 Illustration of the concept of pseudo-Fermi levels.

The Planck's law of radiation gives the density of photons with energy E in a material with reflactive index $\bar{n}$ (we assume a real number ignoring the absorption) as

$$P(E) = \frac{8\pi\bar{n}^3E^3}{h^3c^3}\frac{1}{\exp(E/k_{\mathrm{B}}T)-1}. \tag{4.70}$$

When minority carriers are generated by photoexcitation under light irradiation, the carrier distribution deviates from the thermal equilibrium, which is described by single chemical potential and temperature. Even in such a case, if the system is steady in balance, we consider the energy distribution function f_c of electrons in the conduction band, that of electrons in the valence band f_v. Because in most cases, relaxation of distribution by carrier-to-carrier interaction and relaxation by intra-band carrier-lattice interaction are much faster processes than the inter-band carrier recombination, we adopt the approximation described in the following. In a semiconductor under steady light irradiation, the electrons in the conduction band and those in the valence band are in quasi-thermal equilibrium state described by the Fermi distribution functions with the same temperature but with the

[*1] There are many other excitation factors around us, such as electron beams, sound, friction, and chemical reactions, etc.

different chemical potentials called **quasi-Fermi levels**. The difference is caused by the exciation by the light and the slow inter-band transition. Then we write

$$f_c(E) = \left[\exp\left(\frac{E - E_{\mathrm{F}c}}{k_{\mathrm{B}}T}\right) + 1\right]^{-1}, \quad f_v(E) = \left[\exp\left(\frac{E - E_{\mathrm{F}v}}{k_{\mathrm{B}}T}\right) + 1\right]^{-1}. \tag{4.71}$$

Let us consider the process of the photon absorption (energy $\hbar\omega$) and the excitation of an electron from the valence band (energy E_1) to the conduction band (energy E_2). The frequency of such transition is written as

$$R(1 \to 2) = B_{12} f_v (1 - f_c) P(\hbar\omega), \tag{4.72}$$

where B_{12} is the transition probability of $1 \to 2$. Conversely, the frequency of spontaneous emission with the electron relaxation from E_2 to E_1 is

$$R(sp, 2 \to 1) = A_{21} f_c(E_2)(1 - f_v(E_1)), \tag{4.73}$$

independently of the photon density. The frequency of the stimulated emission is proportional to the photon density as

$$R(st, 2 \to 1) = B_{21} f_c(E_2)(1 - f_v(E_1)) P(\hbar\omega). \tag{4.74}$$

They should fulfill the balance equation

$$R(1 \to 2) = R(sp, 2 \to 1) + R(st, 2 \to 1). \tag{4.75}$$

Substituting equations (4.70)−(4.74) to the above and the comparison of LHS and RHS gives the following **Einstein relations**.

$$\begin{cases} A_{21} = \dfrac{8\pi\bar{n}^3 E_{21}^3}{h^3 c^3} B_{21}, & \text{(4.76a)} \\ B_{12} = B_{21}. & \text{(4.76b)} \end{cases}$$

These are identical with eq.(4A.9). Equation (4A.9) is for the angular frequency spectrum and there is the difference in the conversion factor $\hbar$.

4.3 Phenomenological treatment of electromagnetic field in materials

In the above we have considered the optical response caused by the photon absorption by interband transition of electrons based on the knowledge of two-level systems. This is very important of course, but there are many other optical processes in real crystals. It is also important to look at the optical phenomena from macroscopic perspectives. For example, the reflactive index can be viewed as a parameter that modifies the speed of light. We have a brief look at such a classical macroscopic approachs.

Let us begin with the Maxwell equations:

$$\mathrm{div}\boldsymbol{D} = \rho, \quad \mathrm{div}\boldsymbol{B} = 0, \tag{4.77a}$$

$$\mathrm{rot}\boldsymbol{E} = \frac{\partial \boldsymbol{B}}{\partial t}, \quad \mathrm{rot}\boldsymbol{H} = \boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{4.77b}$$

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}, \quad \boldsymbol{B} = \mu_0 \boldsymbol{H} + \boldsymbol{M}. \tag{4.77c}$$

Here we assume a non-magnetic insulating material and drop the magnetization $\boldsymbol{M} = \vec{0}$, and the current $|\boldsymbol{j}| \ll |\partial \boldsymbol{D}/\partial t|$[*2]. These simplifications leads to the following wave equation.

$$\Delta \boldsymbol{E} - \epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = \mu_0 \frac{\partial^2 \boldsymbol{P}}{\partial t^2}, \tag{4.78}$$

[*2] When these are finite, various interesting phenomena are expected even in this macroscopic level. On the microscopic level, we can find numerous subjects. These are called **magneto-optical effects**. They are the targets of researches as well as the sources of many usuful experimental techniques. [2, 3] are recommended for advanced study.

which is the same as that for the vacuum when $\boldsymbol{P} = \vec{0}$. This means that the polarization $\boldsymbol{P}$ represents the effect of dielectric material in this macroscopic model. In the linear response approximation, $\boldsymbol{P}$ is written with the electric susceptibility tensor $\boldsymbol{\chi}$ as

$$\boldsymbol{P} = \epsilon_0 \boldsymbol{\chi} \boldsymbol{E}. \tag{4.79}$$

Equation (4.77c) leads to $\boldsymbol{D} = \epsilon_0(1 + \boldsymbol{\chi}_\mathrm{r})\boldsymbol{E}$ and **relative dielectric functionor** relative permittivity is defined as follows[*3].

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{\epsilon}_\mathrm{r} \boldsymbol{E}, \quad \boldsymbol{\epsilon}_\mathrm{r} = 1 + \boldsymbol{\chi}. \tag{4.80}$$

Below for simplicity, we consider isotropic materials and the tensor $\boldsymbol{\epsilon}_\mathrm{r}$ can be treated as a scalar ϵ_r. From eq.(4.79) and (4.80), eq.(4.78) becomes

$$\Delta \boldsymbol{E} - \epsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = \epsilon_0 \mu_0 (\epsilon_\mathrm{r} - 1) \frac{\partial^2 \boldsymbol{E}}{\partial t^2}.$$

Then we obtain

$$\Delta \boldsymbol{E} - \frac{\epsilon_\mathrm{r}}{c^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0. \tag{4.81}$$

In the above simplest approximation, the effect of polarization in the material can be taken into account with changing the light speed c with $c' = c/\sqrt{\epsilon_\mathrm{r}}$. Hence the dispersion relation in the vacuum $\omega = ck$ is modified as

$$c^2 \boldsymbol{k}^2 = \omega^2 \epsilon_\mathrm{r}(\omega, \boldsymbol{k}). \tag{4.82}$$

Here ϵ_r depends on ω_r, $\boldsymbol{k}$, reflecting the properties of materials.

As above, the association of the polarization with the electromagnetic wave inside materials can be taken into account phenomenologically by considering the relative dielectric function $\epsilon_\mathrm{r}(\omega, \boldsymbol{k})$ or the refractive index $\tilde{n} = \sqrt{\epsilon_\mathrm{r}}$. Above that, the absorption we saw in Sec.4.2.1 can be phenomenologically taken into account with adding the imaginary part to the refractive index.Then the **complex refractive index** is defined as

$$\tilde{n}(\omega, \boldsymbol{k}) = n(\omega, \boldsymbol{k}) + i\kappa(\omega, \boldsymbol{k}). \tag{4.83}$$

Then from eq.(4.59), or from the definition $I(z) = I_0 \exp(-\alpha z)$, the absorption coefficient α is expressed as

$$\alpha = \frac{2\omega}{c} \kappa(\omega, \boldsymbol{k}). \tag{4.84}$$

Let us go into a bit "model" of materials. In the Lorentz model, the electromagnetic field in the materials is a set of harmonic oscillators. In the model the mass, the charge, and the spring constant is common as (m, e, ξ) and the electromagnetic wave interacts with the oscillators through the Coulomb interaction with the charges. The frequency of the electromagnetic wave is ω and the wavelength is much longer than the distance between the oscillators and the electromagnetic wave can be approximated by uniform time-dependent electric field, which is written as $eE_0 e^{-i\omega t}$. The equation of motion for each oscillator is written as

$$m\frac{d^2x}{dt^2} + \Gamma m \frac{dx}{dt} + \xi x = eE_0 \exp(-i\omega t), \tag{4.85}$$

where Γm is the coefficient representing the energy dissipation (friction in a classical model).

The eigenfrequency of each oscillator is $\omega_\mathrm{h} = \sqrt{\xi/m}$. In order to find the long-term stable solution of (4.85), we substitute $x(t) = x_\mathrm{p} \exp(-i\omega t)$. Then

$$x_\mathrm{p}(\omega) = \frac{eE_0}{m} \frac{1}{\omega_\mathrm{h}^2 - \omega^2 - i\omega\Gamma} \tag{4.86}$$

[*3] Various expressions are use for the dielectric funtion. Here we put the expression "relative" to clarify the unit is taken as the vacuum dielectric constant ϵ_0. The units in electromagnetism often cause confusions. Textbooks [5, 6, 7] are recommended for those who are intrested in the problem.

is obtained. Let N be the spatial density of the oscillators and we get

$$P = N(ex_{\rm p}(\omega)) = \frac{Ne^2}{m}\frac{1}{\omega_{\rm h}^2 - \omega^2 - i\omega\Gamma}E_0. \tag{4.87}$$

The coefficient of E_0 in r.h.s. corresponds to χ in (4.79). Then the definition in (4.80) leads to the relative dielectric function

$$\epsilon_{\rm r}(\omega) = 1 + \frac{Ne^2}{\epsilon_0 m}\frac{1}{\omega_{\rm h}^2 - \omega^2 - i\omega\Gamma}. \tag{4.88}$$

In the above we consider the case of single mode oscillator. If the oscillator has multiple mode and we write f_j as the portion of the mode indicated by index j, (4.88) は

$$\epsilon_{\rm r}(\omega) = 1 + \frac{Ne^2}{\epsilon_0 m}\sum_j \frac{f_j}{\omega_{\rm h}^2 - \omega^2 - i\omega\Gamma_j}. \tag{4.89}$$

This f_j is the oscillator strength we've already seen, but with this treatment we understand the wording of "oscillator strength."

4.4 Optical response of excitons

The excitons introduced at the end of the last chapter have discrete energy levels below the band gap. In many cases they appear as prominent peak structures in the absorption/emission spectrum. In bulk semiconductors, they appear mostly at low temperatures but the situation changes in quantum structures discussed later in this lecture. We do not have time to go into but the Frenkel-type excitons are now the main origin of the electroluminescence in organic semiconductors. Let us begin with the excitons in bulk semiconductors.

4.4.1 Absorption/emission by excitons

As we saw in Sec.3.3.2, the kinetic freedoms in excitions can be specified by the electon-hole relative spatial coordinate $\boldsymbol{r}$ and coordinate of the parallel motion $\boldsymbol{R}$. Then the wavefunction can be written in the effectve mass approximation as

$$\Phi_{n\boldsymbol{K}}(\boldsymbol{r}, \boldsymbol{R}) = \frac{1}{\sqrt{V}}\exp(i\boldsymbol{K}\cdot\boldsymbol{R})\phi_n(\boldsymbol{r}). \tag{4.90}$$

The Fourier transform of the above is

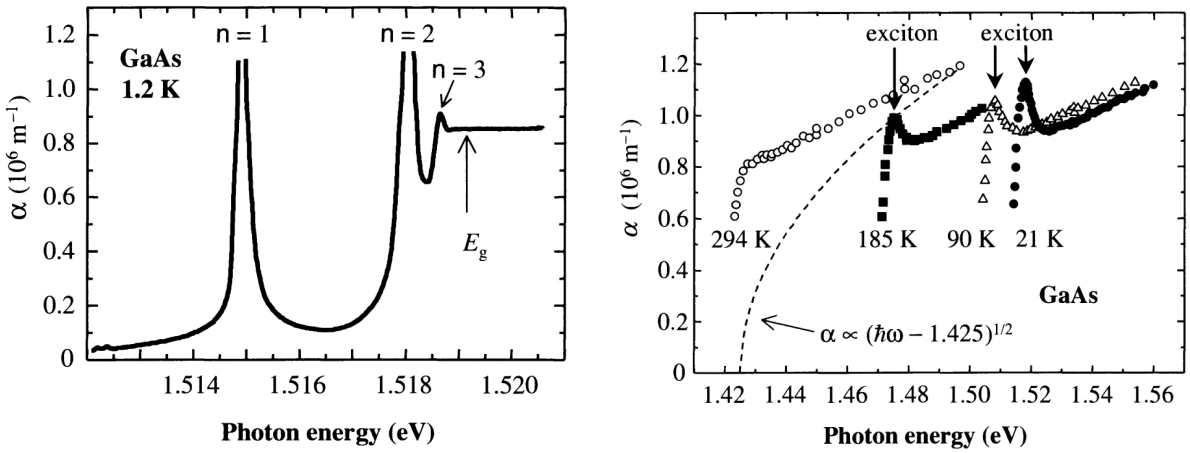
$$\begin{aligned} F_{n\boldsymbol{K}}(\boldsymbol{k}_{\rm e}, \boldsymbol{k}_{\rm h}) &= \frac{1}{V}\int d^3\boldsymbol{r}_{\rm e} d^3\boldsymbol{r}_{\rm h} e^{-i\boldsymbol{k}_{\rm e}\cdot\boldsymbol{r}_{\rm e}} e^{-i\boldsymbol{k}_{\rm h}\cdot\boldsymbol{r}_{\rm h}}\Phi_{n\boldsymbol{K}}(\boldsymbol{r}, \boldsymbol{R}) \\ &= \frac{1}{\sqrt{V}}\int d^3\boldsymbol{r} d^3\boldsymbol{R} e^{-i\boldsymbol{R}\cdot(\boldsymbol{k}_{\rm e}+\boldsymbol{k}_{\rm h}-\boldsymbol{K})}\phi_n(\boldsymbol{r})e^{-i\boldsymbol{k}^*\cdot\boldsymbol{r}} \\ &= \frac{1}{\sqrt{V}}\int d^3\boldsymbol{r} e^{-i\boldsymbol{k}^*\cdot\boldsymbol{r}}\phi_n(\boldsymbol{r})\delta_{\boldsymbol{K},\boldsymbol{k}_{\rm e}+\boldsymbol{k}_{\rm h}}, \quad \boldsymbol{k}^* \equiv \frac{m_{\rm h}\boldsymbol{k}_{\rm e} - m_{\rm e}\boldsymbol{k}_{\rm h}}{m_{\rm e}+m_{\rm h}}. \end{aligned} \tag{4.91}$$

The total wavenumber of the excition $\boldsymbol{K}$ is thus turned out to be

$$\boldsymbol{K} = \boldsymbol{k}_{\rm e} + \boldsymbol{k}_{\rm h}. \tag{4.92}$$

For the treatment of optical absorption, we take the initial state before the electron-hole excitation as the ground state $\Phi_0 = \phi_{{\rm c}\boldsymbol{k}_{\rm e}}\phi_{{\rm v}\boldsymbol{k}_{\rm e}}$ and calculate the transition probability $w_{\rm if}$ to the state represented as eq.(4.90) with taking $\boldsymbol{k}_{\rm p} = \vec{0}$, $\boldsymbol{k}_{\rm e} = -\boldsymbol{k}_{\rm h}$ and along with the line shown for the case of two-level systems.

$$\begin{aligned} w_{\rm if} &= \frac{2\pi}{\hbar}\frac{e^2}{m^2}|A_0|^2\frac{1}{V}\sum_\lambda |\langle\Phi_{\lambda\boldsymbol{K}}|\exp(i\boldsymbol{k}_{\rm p}\cdot\boldsymbol{r})\boldsymbol{e}\cdot\boldsymbol{p}|\Phi_0\rangle|^2\delta(E_{\rm g}+E_\lambda-\hbar\omega) \\ &= \frac{2\pi}{\hbar}\frac{e^2}{m^2}|A_0|^2\frac{1}{V}\sum_{\boldsymbol{k}_{\rm e}\lambda} |F_{\lambda\boldsymbol{K}}(\boldsymbol{k}_{\rm e}, -\boldsymbol{k}_{\rm e})\langle\phi_{{\rm c}\boldsymbol{k}_{\rm e}}|\boldsymbol{e}\cdot\boldsymbol{p}|\phi_{{\rm v}\boldsymbol{k}_{\rm e}}\rangle|^2\delta(E_{\rm g}+E_\lambda-\hbar\omega). \end{aligned} \tag{4.93}$$



(a) (b)

Fig. 4.4 (a) Optical absorption peaks by excitons at lower side in the energy than the fundamental absorption edge in GaAs[9]. (b) Absorption anomaly by excitons around the fundamental edge in GaAs[10].

From $\boldsymbol{k}_{\rm e} = -\boldsymbol{k}_{\rm h}$,

$$F_{n\boldsymbol{K}}(\boldsymbol{k}_{\rm e}, -\boldsymbol{k}_{\rm h}) = \frac{1}{V}\int d^3\boldsymbol{r}_{\rm e} d^3\boldsymbol{r}_{\rm h} \exp[-i\boldsymbol{k}_{\rm e}\cdot(\boldsymbol{r}_{\rm e}-\boldsymbol{r}_{\rm h})]\Phi_{\lambda\boldsymbol{K}}(\boldsymbol{r}_{\rm e},\boldsymbol{r}_{\rm h}). \tag{4.94}$$

In (4.93), the summation over $\boldsymbol{k}_{\rm e}$ results in $\boldsymbol{r}_{\rm e} = \boldsymbol{r}_{\rm h}$. $F_{n\boldsymbol{K}}$ takes large values only in a narrow region of $\boldsymbol{k}_{\rm e}$ around $\boldsymbol{k}_{\rm e} \approx \vec{0}$. In that region, $\langle \phi_{{\rm c}\boldsymbol{k}_{\rm e}} | \boldsymbol{e}\cdot\boldsymbol{p} | \phi_{{\rm v}\boldsymbol{k}_{\rm e}} \rangle$ is almost constant and is M in (4.62). We then obtain

$$w_{\rm if} = \frac{2\pi}{\hbar}\frac{e^2}{m^2}|A_0|^2 \sum_\lambda |M|^2 |\phi_\lambda(0)|^2 \delta(E_{\rm g} + E_\lambda - \hbar\omega). \tag{4.95}$$

Again for simplicity we consider an isotropic system. Because $\phi_\lambda(0)$ is not zero only for s-state,

$$|\phi_n(0)|^2 = \frac{1}{\pi a_{\rm ex}^3 n^3}, \quad E_n = -\frac{E_{\rm ex}}{n^2}. \tag{4.96}$$

The imaginary part of the relative dielectric function $\epsilon_{\rm r2}(\omega) = 2in(\omega)\kappa(\omega)$ is

$$\epsilon_{\rm r2}(\omega) = \frac{\pi e^2}{\epsilon_0 m^2 \omega^2}|M|^2 \frac{1}{\pi a_{\rm ex}^3}\sum_n \frac{1}{n^3}\delta\left(E_{\rm g} - \frac{E_{\rm ex}}{n^2} - \hbar\omega\right). \tag{4.97}$$

In the above the spin degree of freedom 2 is not considered and the result should be multiplied by two.

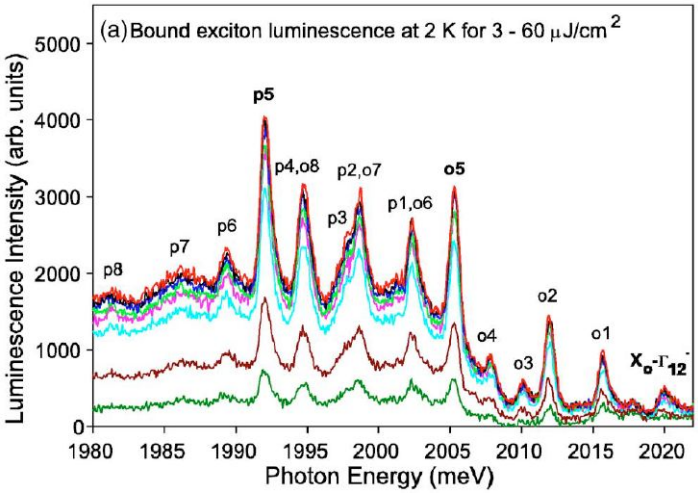


Fig. 4.5 Emission spectra of bound excitons in Cu_2O[11].

We do not go into calculation details (see *e.g.* [8]), but the twice of (4.97) agrees with (4.68) at the boundary $\hbar\omega = E_{\rm g}$ between discrete states and continuum. Hence we can confirm how good is the approximation by the comparison of the spectra in experiments.

In eq.(4.97), the part other than δ-functions is common and we write it as the constant C.

$$\epsilon_{\rm r2} = C\delta\left(E_{\rm g} - \frac{E_{\rm ex}}{n^2} - \hbar\omega\right). \tag{4.98}$$

Mathematical identity

$$\lim_{\Gamma\to+0}\frac{1}{x_0 - x - i\Gamma} = \mathcal{P}\frac{1}{x_0 - x} + i\pi\delta(x_0 - x) \tag{4.99}$$

tells

$$\epsilon_{\mathrm{r}2} = \mathrm{Im}\left\{ \frac{C/\pi}{E_{\mathrm{g}} - \dfrac{E_{\mathrm{ex}}}{n^2} - (\hbar\omega + i\delta)} \right\}. \tag{4.100}$$

Here we write $\Gamma \to +0$ as δ. And the Kramers-Kronig relation (4B.2) leads to

$$\epsilon_{\mathrm{r}} = \frac{C/\pi}{E_{\mathrm{g}} - \dfrac{E_{\mathrm{ex}}}{n^2} - (\hbar\omega + i\Gamma)}, \tag{4.101}$$

with which we can try fitting the data in, *e.g.* Fig. 4.4(b).

The emission is the reversal process of the absorption and just as the absorption, discrete emission peaks appear at lower energies than the funcamental emission edge. Figure 4.5 shows an example of photoluminescence spectra of Cu_2O.

4.4.2 Exciton-polariton

Well known as "polaritons" are the quasiparticle created by the combination of optical phonons and photons. Here we consider, however the combination of photons and excitons. The concept of **exciton-polariton** is illustrated in Fig. 4.6. As mentioned in the previous section, an absorption and an emission of photon with an exciton are reversal process to each other. In an exciton-polariton these processes form a continuous chain. The cycle period of the processes is as short as a few fs and both the exciton and the photon keep their quantum coherence and the resultant quasiparticle propagates inside the crystal as a coherent state.

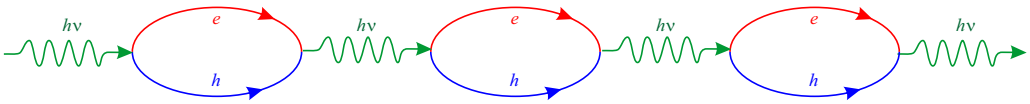


Fig. 4.6 Illustration of the concept of exciton-polariton. A photon creates an exciton and the recombination of the electron-hole pair recreates a photon. These processes occur in series.

We consider the ground state of $n = 1$ in eq.(4.101), define ω_0 as $E_{\mathrm{g}} - E_{\mathrm{ex}} \equiv \hbar\omega_0$, and the contribution to the dielectrin function other than the excitons as ϵ_{s}. Then with $\gamma = \Gamma/\hbar$, the relative dielectric function is written as

$$\epsilon_{\mathrm{r}}(\omega) = \epsilon_{\mathrm{s}}\left(1 + \frac{\Delta_{\mathrm{ex}}}{\omega_0 - \omega - i\gamma}\right). \tag{4.102}$$

For the transverse wave with $\boldsymbol{k} \cdot \boldsymbol{E} = 0$, the angular frequency $\omega_{\mathrm{t}} = \omega_0$, the polariton equation (4.82) holds. On the other hand, for the longitudinal wave $\epsilon_{\mathrm{r}}(\omega) = 0$, the angular frequency ω_{l} is given as

$$\omega_{\mathrm{l}} = \omega_0 + \Delta_{\mathrm{ex}} = \omega_{\mathrm{t}} + \Delta_{\mathrm{ex}}. \tag{4.103}$$

Δ_{ex} is called longitudinal-transverse splitting.

Now we consider the wavenumber $k = k_1 + ik_2$, then from (4.82), (4.102) we get

$$\begin{cases} \dfrac{\omega^2\epsilon_{\mathrm{s}}}{c^2}\left(1 + \dfrac{\Delta_{\mathrm{ex}}}{\omega_0 - \omega}\right) = k_1^2 - k_2^2,, & \text{(4.104a)} \\ \pi\delta(\omega - \omega_0)\dfrac{\omega_0^2\epsilon_{\mathrm{s}}}{c^2} = 2k_1k_2. & \text{(4.104b)} \end{cases}$$

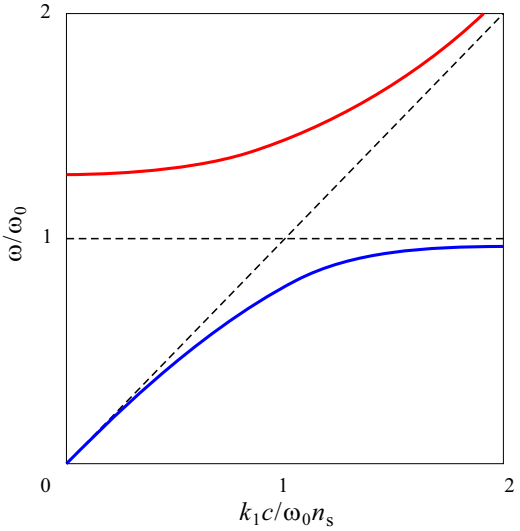


Fig. 4.7 Schematic drawing of the dispersion relation of exciton-polariton.

Equation (4.104b) represents the resonance at $\omega = \omega_0$, then we ignore k_2 in (4.104a) to get

$$\omega\sqrt{\frac{\omega - \omega_- \Delta_{\rm ex}}{\omega - \omega_0}} = \frac{ek_1}{\sqrt{\epsilon_{\rm s}}}, \tag{4.105}$$

which gives the dispersion relation of exciton-polariton.

Appendix 4B: Kramers-Kronig relation

Here we just show well-known Kramers-Kronig relation. Let us consider a complex function with a complex argument ω as

$$\chi(\omega) = \chi_1(\omega) + i\chi_2(\omega), \quad \chi_1,\ \chi_2 \in \mathbb{R}. \tag{4B.1}$$

$\chi(\omega)$ is analytic in the upper half of ω-plane and diminish faster than $1/|\omega|$ for large $|\omega|$. Then there hold relations between χ_1 and χ_2 as

$$\chi_1(\omega) = \frac{1}{\pi}\mathcal{P}\int_{-\infty}^{\infty}\frac{\chi_2(\omega')}{\omega' - \omega}d\omega', \quad \chi_2(\omega) = -\frac{1}{\pi}\mathcal{P}\int_{-\infty}^{\infty}\frac{\chi_1(\omega')}{\omega' - \omega}d\omega'. \tag{4B.2}$$

Here $\mathcal{P}$ represents the Cauthey's principal value. The above are the Kramers-Kronig relation.

Appendix 4C: Lattice vibration in semiconductors

Lattice vibration is a phenomenon in which an atom vibrates around it with kinetic energy while being localized at an equilibrium position as a time average position. This is an important subject in semiconductor physics, should be discussed using at least one whole chapter, but that is impossible due to the lecture time. Here we take a minimum look at very basics.

4C.1 Lattice vibration in one-dimensional system

Just as in the electron system, we introduce basic concepts in one-dimensional systems.

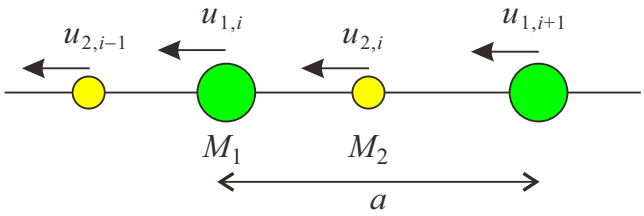


Fig. 4C.1 Schematic diagram of one-dimensional lattice vibration.

We consider a one-dimensional latteic with the unit cell of length a, which has two atoms with masses M_1 and M_2. The shifts of the atoms from the equilibrium positions are written as u_{1j}, u_{2j} (j: integer). The force working on the atoms is assumed to be harmonic oscillator-like, that is, the force proportional to the shift of the distance between neighboring atoms from the equilibrium value $a/2$. Let α be the coefficient for the force then we get the equation of motion as

$$M_1\frac{d^2u_{1,j}}{dt^2} = -\alpha(u_{1,j} - u_{2,j-1}) + \alpha(u_{2,j} + u_{2,j-1}) = \alpha[-2u_{1,j} + (u_{2,j} - u_{2,j-1})], \tag{4C.1a}$$

$$M_2\frac{d^2u_{2,j}}{dt^2} = \alpha[-2u_{2,j} + (u_{1,j} - u_{1,j-1})]. \tag{4C.1b}$$

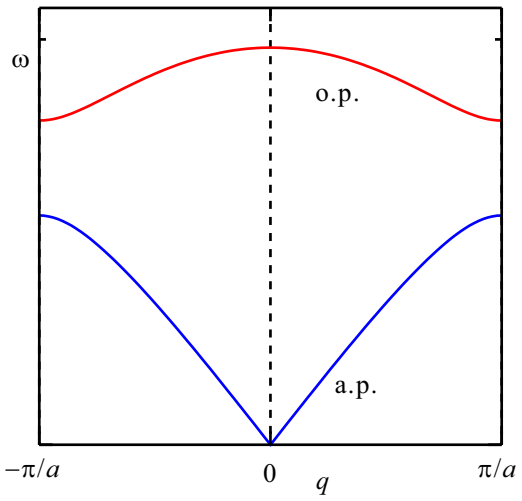


Fig. 4C.2 Dispersion relation of one dimensional lattice vibration $\omega(q)$ obtained from eq.(4C.5) is plotted for the case $M_1 = 2M_2$.

The equation (4C.1) remains unchanged with the parallel shift operation $j \to j + n$ (n is an integer) and the solution can be written in the form of Bloch function. Let us take x coordinate along the lattice direction and the equations for the wavenumber q are

$$\begin{cases} u_{1,j}(x_j) = e^{iqx_j} u_{1,q}, \\ u_{2,j}(x_j + a/2) = e^{iq(x_j + a/2)} u_{2,q}. \end{cases} \tag{4C.2}$$

Substituting the above into (4C.1) we obtain

$$\begin{cases} M_1 \dfrac{d^2 u_{1,q}}{dt^2} = 2\alpha(-u_{1,q} + \cos\dfrac{ja}{2} u_{2,q}), \\ M_2 \dfrac{d^2 u_{2,q}}{dt^2} = 2\alpha(-u_{2,q} + \cos\dfrac{ja}{2} u_{1,q}). \end{cases} \tag{4C.3}$$

In order to find the solution we assume $u_{(1,2),q} \propto \exp(i\omega t)$ to write down

$$\begin{pmatrix} 2\alpha - M_1\omega^2 & -2\alpha\cos\dfrac{qa}{2} \\ -2\alpha\cos\dfrac{qa}{2} & 2\alpha - M_2\omega^2 \end{pmatrix} \begin{pmatrix} u_{1,j} \\ u_{2,q} \end{pmatrix} \equiv \boldsymbol{A} \begin{pmatrix} u_{1,q} \\ u_{2,q} \end{pmatrix} = \vec{0}. \tag{4C.4}$$

For the equations to have non-trivial solution $\{u_{i,q}\}$ other than $\vec{0}$, $|\boldsymbol{A}| = 0$ leads to

$$\frac{\omega_\pm^2}{\alpha} = \left(\frac{1}{M_1} + \frac{1}{M_2}\right) \pm \sqrt{\left(\frac{1}{M_1} + \frac{1}{M_2}\right)^2 - 4\frac{\sin^2(qa/2)}{M_1 M_2}}. \tag{4C.5}$$

We consider non-negative ω, and eq.(4C.5) has two modes, the dispersion relations of which are shown in Fig. 4C.2.

The following description of wording does not depend on the dimension.

The modes with linear dispersion around $q \approx 0$ are called **acoustic modes**, those with finite ω and $d\omega/dq = 0$ for $q = 0$ are called **optical modes**. The naming acoustic mode comes from the property that the group velocity does not depend on the frequency just like sound in the air or electromagnetic wave in the vacuum. The naming optical mode comes from the interaction with photons as the small wavenumber and the large energy. The quantized particles of them are called **acoustic phonon** and **optical phonon** respectively.

4C.2 Lattice vibration in zinc-blende crystals

We consider zinc-blende (ZB) crystals as an example of three-dimensional crystal which has two species of atoms in the unit cell. The Bravais lattice is fcc but the ZB crystalline structure can be considered as an overlapp of two "fcc crystals", in which one atom is placed at the lattice point of fcc lattice.

(to be continued in the next lecture note.)

References

[1] 櫛田孝司「光物性物理学」(朝倉書店，2009).

[2] 佐藤勝昭「光と磁気」(朝倉書店，1988).

[3] S. Sugano ed., “Magneto-optics” (Springer, 1999).

[4] A. K. Zvezdin and V. A. Kotov, “Modern Magnetooptics and Magnetooptical Materials” (IOPP, 1997).

[5] 青野 修「電磁気学の単位系」(丸善，1990).

[6] 高梨弘毅「磁気工学入門」(共立出版，2008)

[7] 2018 年の SI 単位改定については，沢山解説が出ているが，改定自身というより単位系全般の話を非常に易しく解説したものとして，和田純夫, 大上雅史, 根本和昭「単位がわかると物理がわかる」(ベレ出版, 2002).

[8] 浜口智尋「半導体物理」 (朝倉書店, 2001)．英語版は C. Hamaguchi, “Basic Semiconductor Physics” (Springer, 2017).

[9] G.W. Fehrenbach, W. Schafer, and R. G. Ulbrich, J. Luminescence, **30**, 154 (1985).

[10] M. D. Sturge, Phys. Rev. **127**, 768 (1962).

[11] J. I. Jang, Y. Sun, B. Watkins, and J. B. Ketterson, Phys. Rev. B **74**, 235204 (2006).

Lecture note on **Physics of Semiconductors (6)**

19th May (2021) Shingo Katsumoto, Institute for Solid State Physics, University of Tokyo

Chapter 5 Semi-classical treatment of electrical transport

The electric tranport is a response to external perturbations as important as the optical response. Treatment of non-equilibrity to some extent is inevitable for the discusstion of transport. A big difference between the electric transport and the optical response is, however, in the former the characteristic energy scale is much smaller than that in the latter ($\sim E_{\rm g}$). In this chapter, we have a brief look at very basic part of the transport in the linear response regime, in semi-classical treatment. We will go into the quantum transport in the later chapters.

5.1 Classical transport phenomena

Among transport phenomena interests of physicist mainly lies in quantum transport such as the quantum Hall effect. In earth-flooding semiconductor devices, however, dominant is the classical transport [*1] .

The reason that the classical theories are applicable to transport in semiconductors at room temperatures mainly lies in the low density of carriers. In bulk transport, for example, the Fermi level $E_{\rm F}$ lies in band-gap, that is, there is no density of states around $E_{\rm F}$. When we are looking at the energy distribution of electrons, what we actually see is the tail of the Fermi distribution function, which can be approximated with Maxwellian.

Heavy doping changes semiconductors into disordered metals, or spatial modulation of materials which shift the positions of Fermi levels above the conduction band edges provide low-dimensional metallic systems. Even in many of such systems, classical approximations hold around room temperatures. The Fermi degeneration temperature for a system with density n and particle mass m is

$$T_{\rm F} = \frac{\hbar^2}{2mk_{\rm B}}(3\pi^2 n)^{2/3} \quad :3{\rm D}, \qquad \frac{\hbar^2}{16\pi mk_{\rm B}}n \quad :2{\rm D}. \tag{5.1}$$

Substitution of typical values for semiconductors give, *e.g.* for a two-dimensional electron system, which has comparatively large Fermi energy, about 70K for $T_{\rm F}$. That is, the distribution of kinetic energy is still described by a Maxwellian. Furthermore, the width of distribution is as large as the Fermi energy making the quantum mechanical interference effect obscure. In this chapter we thus concentrate on the phenomena, which can be described within classical theories for electron kinematics in solids.

5.1.1 Transport phenomena and transport coefficient

"Transport" here means transportation of some physical quantity in real space. In the treatment of such a problem, we often map the problem onto a set of particles and the transport is transfer of the particles in the model. For example, consider a stretched string and some local shift from the stretched line. The shift is transmitted on the string as a wave but we can also treat the shift as a particle, which brings some potential energy. We may consider, then, the transport of the

[*1] In many devices quantum confinement is working and low-dimensional systems are realized though the transport can be understood within classical theories.

shift. In solids, we actually have various elementary excitations such as phonons, spin-waves (magnons), etc. In electric conduction, which is a representative transport in semiconductors, the physical quantity is charge and a particle bringing it is called a "carrier". Examples are, of course, electrons and holes, many-body states of band electrons actually.

Needless to say, we first need to exclude trivial motion of the center of mass due to arbitrary selection of inertial system. We thus assume that the center of mass for the system under consideration sits still in equilibrium without perturbation. Transport is a response flow of some physical quantity to an external perturbation. [*2]. In the case of linear response, like electric current for voltage in Ohm law, the coefficient is called **transport coefficient**.

We often have strongly non-linear response in electric transport in semiconductors though we begin with linear response. As a typical example, in the electric current response to the field, thre linear response between the current density $\boldsymbol{j}$ and the field $\mathbf{E}$ is written as

$$\boldsymbol{j} = \boldsymbol{\sigma}\mathbf{E}, \quad \mathbf{E} = \boldsymbol{\rho}\boldsymbol{j} = \boldsymbol{\sigma}^{-1}\boldsymbol{j}, \tag{5.2}$$

where $\boldsymbol{\sigma}$ is the **conductivity tensor**, $\boldsymbol{\rho}$ is the **resistivity tensor**. These two are inverse tensor to each other.

5.1.2 Boltzmann equation

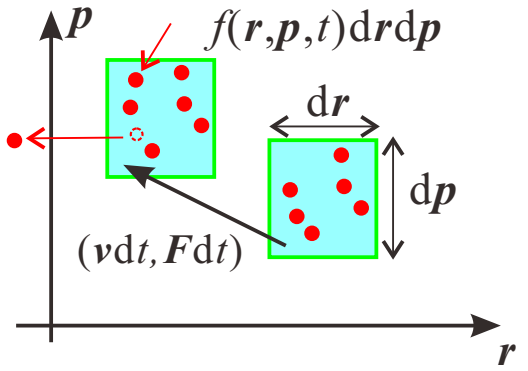


Fig. 5.1 Illustration of time evolution for particles in an infinitesimal volume $d\boldsymbol{r}d\boldsymbol{p}$ in an infinitesimal time dt with a scattering.

Let us consider a distribution function $f(\boldsymbol{r}, \boldsymbol{p}, t)$ in a six-dimensional space of spatial coordinate $\boldsymbol{r}$ and momentum $\boldsymbol{p}$, *i.e.*, a phase space. The meaning of f is that the ratio of particles in the volume $d\boldsymbol{r}d\boldsymbol{p}$ around the point $(\boldsymbol{r}, \boldsymbol{p})$ in the whole system is $f(\boldsymbol{r}, \boldsymbol{p}, t)d\boldsymbol{r}d\boldsymbol{p}$.

In the absence of scattering, the classical equation of motion is described as

$$\mathrm{d}\boldsymbol{r}/\mathrm{d}t = \boldsymbol{v} = \boldsymbol{p}/m^*, \qquad \mathrm{d}\boldsymbol{p}/\mathrm{d}t = \boldsymbol{F}, \tag{5.3}$$

with $\boldsymbol{F}$ as the force working on the particle. Kinematic states of particles in $\mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{p}$ are the same in the first order and so are the time evolution during dt, giving

$$f(\boldsymbol{r} + \boldsymbol{v}dt, \boldsymbol{p} + \boldsymbol{F}dt, t + dt) = f(\boldsymbol{r}, \boldsymbol{p}, t).$$

Some scatterings bring shifts in f as illustrated in Fig. 5.1. We write the coefficient in the shifts as $(\partial f/\partial t)_c$, that is,

$$f(\boldsymbol{r} + (\boldsymbol{p}/m^*)dt, \boldsymbol{p} + \boldsymbol{F}dt, t + dt) + (\partial f/\partial t)_c dt = f(\boldsymbol{r}, \boldsymbol{p}, t).$$

Expanding f in the left hand side to the first order, we get

Boltzmann equation

$$\frac{\partial f}{\partial t} + \frac{\boldsymbol{p}}{m^*} \cdot \frac{\partial f}{\partial \boldsymbol{r}} + \boldsymbol{F} \cdot \frac{\partial f}{\partial \boldsymbol{p}} = -\left(\frac{\partial f}{\partial t}\right)_c \tag{5.4}$$

Equation (5.4) is called **Boltzmann equation**, and the right hand side is called the collision term.

[*2] This definition cannot include supercurrent or diamagnetic current at edge states of quantum Hall effect. But we usually include them into transport phenomena taking the reference of coordinate to crystal lattices.

The collision term depends on the scattering mechanism and the nature of scattering centers and is generally difficult for us to calculate. The simplest approximation of this term is the constant relaxation time approximation, in which we consider a relaxation time τ independent of energy and put

Constant relaxation time approximation

$$-\left(\frac{\partial f}{\partial t}\right)_c = -\frac{f - f_0}{\tau}, \tag{5.5}$$

where f_0 is the equilibrium distribution function for $\boldsymbol{F} = \boldsymbol{0}$, τ, the relaxation time, is the time for recovery from non-equilibrium states. In spatially uniform systems, $\partial f/\partial \boldsymbol{r} = \boldsymbol{0}$, and the approximation (5.5) can be generalized to the one with energy or momentum dependence in τ.

Below, to avoid trivial failure in pure classical pictures, we use some quantum mechanical relation like $\boldsymbol{p} = \hbar\boldsymbol{k}$ or quantum statistics.

5.1.3 Drift current, diffusion current

As currents we here consider electric currents. Net particle flow appears when the distribution function f gets some anisotropy in $\boldsymbol{p}$ space. Hence we need to consider perturbations in (5.4) other than anisotropy or non-uniformity in $\boldsymbol{p}$. The candidates are then $\boldsymbol{F}(= -e\mathbf{E})$, and $\partial/\partial\boldsymbol{r}$. The former perturbation, *e.g.* acceleration by external electric field, brings about non-uniformity of distribution function $f(\boldsymbol{r}, \hbar\boldsymbol{k}, t)$ in $\boldsymbol{k}$-space resulting in the flow of carriers in the real space. That kind of flow is called **drift current**. The latter is non-uniformity of the distribution in the real space and also causes carrier transport, which is called **diffusion current**.

First let us consider a steady uniform electron system under uniform electric field $\mathbf{E}$. From this assumption, $\partial f/\partial t = 0$ (steady) and $\partial f/\partial \boldsymbol{r} = 0$ (uniform). We further assume τ only depends on $\boldsymbol{p}$. Then eq.(5.4) becomes

$$-e\mathbf{E}\cdot\frac{\partial f}{\partial \boldsymbol{p}} = -\frac{f - f_0}{\tau(\boldsymbol{p})} \quad \therefore f(\boldsymbol{p}) = f_0(\boldsymbol{p}) + e\tau(\boldsymbol{p})\mathbf{E}\cdot\frac{\partial f}{\partial \boldsymbol{p}}.$$

In the next step of approximation, we take $\mathbf{E}$ as a small perturbation. Hence, the 1st order expansion is obtained with replacing f in the right hand side with f_0 as

$$f(\boldsymbol{p}) \simeq f_0(\boldsymbol{p}) + e\tau(\boldsymbol{p})\mathbf{E}\cdot(\partial f_0/\partial \boldsymbol{p}). \tag{5.6}$$

Higher order terms can be obtained by successive replacements. Now eq.(5.6) can be viewed as the first order expansion of $f(\boldsymbol{p}) \simeq f_0(\boldsymbol{p} + e\tau(\boldsymbol{p})\mathbf{E}$ with $\mathbf{E}$, which means this $f(\boldsymbol{p})$ is the one shifted by $-e\tau(\boldsymbol{p})\mathbf{E}$ in $\boldsymbol{p}$ space form $f_0(\boldsymbol{p})$. If τ is constant for $\boldsymbol{p}$, the shift is uniform as illustrated in Fig. 5.2.

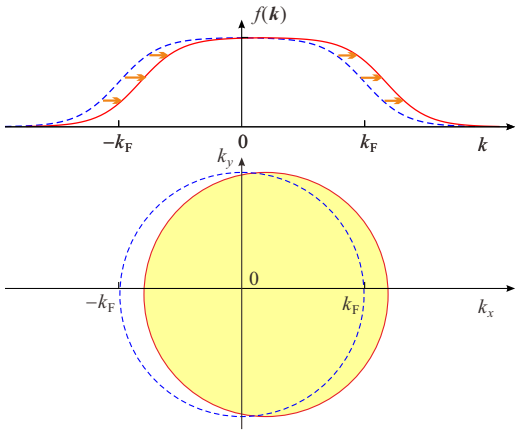


Fig. 5.2 Schematic view for constant shift of Fermi sphere of electrons under acceleration by external electric field in the space of wavenumber. The distribution $f(\boldsymbol{k})$ shifts from the equilibrium position (indicated by broken line) by a wavevector indicated by small arrows. The upper shows the shift in the distribution and the lower shows the shift of Fermi sphere in two-dimensional systems. In realistic systems, the shifts are much smaller than that illustrated here.

We need to integrate $\boldsymbol{v}(\boldsymbol{k})f(\boldsymbol{k})$ in $\boldsymbol{k}$ space to obtain the current. Without loosing generality we take $\mathbf{E} = (\mathcal{E}_x, 0, 0)$ and erasing integrals of odd functions we obtain

$$\int \frac{d^3k}{(2\pi)^3}\boldsymbol{v}(\boldsymbol{k})\left(f_0 + e\tau\mathbf{E}\cdot\frac{\partial f_0}{\hbar\partial\boldsymbol{k}}\right) = \int \frac{d^3k}{(2\pi)^3}\frac{\hbar k_x}{m}e\tau\mathcal{E}_x\frac{\partial f_0}{\hbar\partial k_x} = \frac{e\mathcal{E}_x}{m}\int \mathscr{D}(E)\tau(E)\frac{\hbar^2k_x^2}{m}\frac{\partial f_0}{\partial E}dE, \tag{5.7}$$

where we assume τ depends only on energy. $\hbar^2k_x^2/2m$, the kinetic energy along x-direction is $E/3$ from the equipartition condition.

For a metallic Fermi-degenerated system, $\partial f_0/\partial E$ can be approximated as $-\delta(E - E_\mathrm{F})$ in (5.7). For a three-dimensional system the density of states is $\mathscr{D}(E) = A\sqrt{E}$ with a coefficient A, then (5.7) is

$$\langle v_x\rangle = -A\frac{e\mathcal{E}_x}{m}\frac{2\tau(E_\mathrm{F})}{3}E_\mathrm{F}^{3/2},$$

while the particle density is calculated as

$$n = \int_0^{E_\mathrm{F}} \mathscr{D}(E)dE = A\frac{2}{3}E_\mathrm{F}^{3/2}.$$

Putting together the above expressions we obtain the expression for $\sigma = j/E = -e\langle v_x\rangle/\mathcal{E}_x$ as

Drude conductivity

$$\sigma = \frac{e^2n\tau(E_\mathrm{F})}{m} \tag{5.8}$$

which is well known **Drude conductivity**.

When the temperature is high, or the particle density is low and the Maxwellian approximation holds, from $f_0 \approx A\exp(-E/k_\mathrm{B}T)$,

$$-\frac{\partial f_0}{\partial E} = -\frac{A}{k_\mathrm{B}T}\exp\left[-\frac{E}{k_\mathrm{B}T}\right] = -\frac{f_0}{k_\mathrm{B}T} = -\frac{f_0}{(2\langle E\rangle/3n)}$$

is obtained, in the last equation of which we have used averaged kinetic energy $k_\mathrm{B}T/2$ for single kinetic degree of freedom. The electric conductivity is again given in the Drude form as

$$\sigma = e^2\int \tau(E)\mathscr{D}(E)\frac{2E}{3m}\frac{3nf_0}{2\langle E\rangle}dE = \frac{ne^2\langle\tau\rangle_E}{m}. \tag{5.9}$$

Here, $\langle E\rangle_E$ represents the average with weight $E^{3/2}$:

$$\langle\tau\rangle_E = \frac{\langle\tau E\rangle}{\langle E\rangle} = \int_0^\infty \tau(E)E^{3/2}f_0dE \Big/ \int_0^\infty E^{3/2}f_0dE. \tag{5.10}$$

We then proceed to the diffusion current caused by non-uniformity of f in real space. In Boltzmann equation (5.4), $\boldsymbol{F}$ is set to zero and constant relaxation time approximation (5.5) is applied to the space distribution of $f = f_0 + f_1$ as

$$\boldsymbol{v}\cdot\boldsymbol{\nabla}f = -f_1/\tau, \quad \text{take to the first oder of } f_1 \quad f_1 = -\tau\boldsymbol{v}\cdot\boldsymbol{\nabla}f_0. \tag{5.11}$$

When a constant diffusion current $\boldsymbol{J}$ is flowing through a spatial volume V, it is written as

$$\boldsymbol{J} = (-e)\int_V \tau\boldsymbol{v}(\boldsymbol{v}\cdot\boldsymbol{\nabla}f_0)d\boldsymbol{r}.$$

The direction of $\boldsymbol{\nabla}f_0$ is assumed to be constant and along x-axis then the components in $\boldsymbol{v}$ other than v_x vanish with integration since they are odd functions. $\langle v_x^2\rangle = \langle v^2\rangle/3$ and we further assume that the temperature is uniform and constant, no spatial variation in $\langle v^2\rangle$, then the current density is

$$j_x \text{ (current density)} = -e\int_{\text{unit volume}} \tau v_x^2\frac{\partial f_0}{\partial x}d\boldsymbol{r} = -e\left\langle\frac{\tau v^2}{3}\right\rangle\frac{\partial n}{\partial x}.$$

That is,

$$\boldsymbol{j} = (-e)D\boldsymbol{\nabla}n, \quad D = \langle \tau v^2/3 \rangle. \tag{5.12}$$

Here D is **diffusion constant**and within constant relaxation time approximation,

Einstein relation

$$D = \frac{\tau}{3}\langle v^2 \rangle = \frac{\tau k_{\mathrm{B}} T}{m^*} = \frac{\mu}{e} k_{\mathrm{B}} T \tag{5.13}$$

Equation (5.13) is called **Einstein relation**. μ in the right end is the **mobility**, defined in (5.19), which appears later.

5.1.4 Hall effect

The drift current under magnetic field (flux density $\boldsymbol{B}$) can be calculated with substituting Lorentz force into $\boldsymbol{F}$ in (5.4). The straightforward but a bit long calculation is summarized in Appendix A. Here we consider the situation shown in Fig. 5.3, that is, the sample has a finite length along y-axis and infinitely elongated along x-axis and the electric field $\mathbf{E} = (\mathcal{E}_x, 0, 0)$ is applied. j_y brings the carriers and accumulates them to the edges. The charges at the edges form electric field $\mathbf{E}_{\mathrm{int}} = (0, \mathcal{E}_y, 0)$ and in the ultimate steady state $j_y = 0$.

This phenomenon, which generates an electric field vertical both to the current and the magnetic field is the **Hall effect**. The linear response coefficient

$$R_{\mathrm{H}} = \frac{\mathcal{E}_y}{J_x B_z} \tag{5.14}$$

is called **Hall coefficient**. Hall field $\mathcal{E}_y$ is obtained as follows. From $j_y = 0$,

$$\mathcal{E}_y = -(A_t/A_l)\mathcal{E}_x. \tag{5.15}$$

Substituting the above and (5A.11b) into (5.14), we obtain the conductivity tensor defined in $\boldsymbol{j} = \hat{\sigma}\mathbf{E}$ as

$$\sigma_{xx} = \frac{ne^2}{m^*}A_l = \frac{ne^2}{m^*}\left\langle \frac{\tau}{1+(\omega_c\tau)^2} \right\rangle_E, \ \sigma_{xy} = \frac{ne^2}{m^*}\left\langle \frac{\omega_c\tau^2}{1+(\omega_c\tau)^2} \right\rangle_E, \tag{5.16}$$

$$R_{\mathrm{H}} = -\frac{1}{ne}\frac{A_t}{\omega_c(A_l^2 + A_t^2)}. \tag{5.17}$$

In weak fields, from $\omega_c\tau \ll 1$,

$$R_{\mathrm{H}} = -\frac{1}{ne}\frac{\langle\tau^2\rangle_E}{\langle\tau\rangle_E^2} = \frac{1}{n(-e)}\frac{\Gamma(2s+5/2)\Gamma(5/2)}{(\Gamma(s+5/2))^2} = \frac{r_{\mathrm{H}}}{n(-e)}\left(= \frac{1}{n(-e)}\right). \tag{5.18}$$

Knowing s, we obtain the carrier concentration as well as the sign of charge from the Hall measurement (for holes $-e$ is replaced with e). r_{H}, which is called Hall factor, takes in many cases values around 1 depending on the scattering mechanism at high temperatures (see Tab. 5.1). Within constant relaxation time approximation ($s = 0$) or when the system is Fermi-degenerated, $r_{\mathrm{H}} = 1$. When $s = 0$, as eq.(5A.9) tells, eq.(5.18) holds giving the expression shown in the last parentheses.

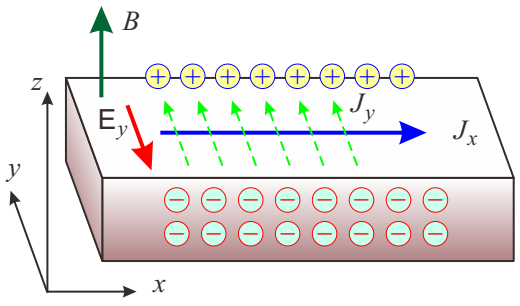


Fig. 5.3 Magnetic field is applied along z-axis. Current along x-axis generates y-component J_y through the Lorentz force. The y-component in current results in charge accumulation at the sample edges, which creates Hall electric field along y-axis. In steady state, J_y is canceled by the Hall field and the total current is along x-axis.

Let v be the average velocity gained by the electrons from the electric field $\mathcal{E}$, the **mobility**is defined as $v/|\mathcal{E}|$, and in the relaxation time approximation, written as

$$\mu = \frac{v}{|\mathcal{E}|} = \frac{nev}{ne|\mathcal{E}|} = \frac{j}{ne|\mathcal{E}|} = \frac{\sigma}{ne} = \sigma|R_{\mathrm{H}}| = \frac{e\tau}{m^*}. \tag{5.19}$$

Scattering mechanism	E exponent	T exponent	Hall factor
Acoustic phonon	$-1/2$	$-3/2$	1.18
Ionized impurity (weak screening)	+3/2	+3/2	1.93
Ionized impurity (strong screening)	+1/2	+1/2	1.18
Neutral impurity	0		1.00
Piezoelectric phonon	$+1/2$		1.10

Tab. 5.1 Hall factors for various scattering mechanism. E, T-exponents are for scattering time. See *e.g.* [1].

5.1.5 Various scatterings

We have considered the Boltzmann equation by relaxation time approximation, but various mechanisms such as scattering with phonons and other degrees of freedom in solids contribute to relaxation. We consider relaxation time for each relaxation mechanism, and index each relaxation time τ_α with index α. Then the frequencies of the relaxations ($\propto \tau_\alpha^{-1}$) is summed up to give the total relaxation. This gives the Matthiessen's rule

$$\tau^{-1} = \sum_\alpha \tau_\alpha^{-1}. \tag{5.20}$$

In the relaxation time approximation of classical transport, the carrier scatterings are taken into account through the averaged scattering time and the Matthiessen's rule(5.20) into the total relaxation time. Therefore we can infer the scattering mechanism dominating the present transport by tuning, for example, a parameter which gives different effects on different scattering times. Scattering of band electrons (holes) have many origins as shown in Fig. 5.4. In this section, representative scatterings and their characteristics are listed.

Phonon scattering: Quantization of lattice vibration gives phonons. The phonons are classified into acoustic phonons, which have the dispersion $E(k_{\mathrm{p}}) \to 0$ for wavenumber $k_{\mathrm{p}} \to 0$, and optical phonons, which have finite $E(k_{\mathrm{p}} \to 0)$. In a plain expression, the difference comes from whether the oscillations of neighboring atoms are in phase or out of phase. For the band electrons the lattice vibration is distortion in the lattice potential and causes scattering. The scattering of electrons causes rebounding of nuclei resulting in the phonon scattering. Such phonon scattering is, from the electron side, inelastic associated with the energy gain/loss.

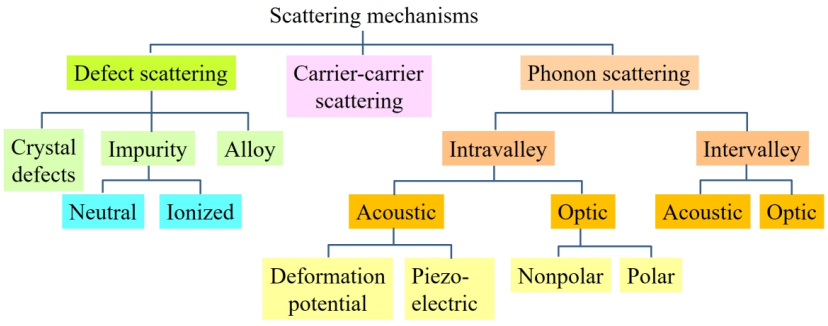


Fig. 5.4 Classification of scatterings mostly by the origins.

The relaxation time due to the acoustic phonon has energy dependence as $\tau(E) = a_{\rm ph}E^{-1/2}$. The averaged scattering time with energy-weight $\langle\tau_{\rm ph}\rangle_E$ is

$$\langle\tau_{\rm ph}\rangle_E = a_{\rm ph}(k_{\rm B}T)^{-1/2}\frac{\Gamma(2)}{\Gamma(5/2)} = \frac{8\sqrt{\pi}a_{\rm ph}}{3\sqrt{k_{\rm B}T}}. \tag{5.21}$$

In high temperature approximation, the energy distribution of phonons gives $a_{\rm ph} \propto (k_{\rm B}T)^{-1}$, then the mobility limited by the acoustic phonons $\mu_{\rm ph}$ has the temperature dependence as

$$\mu_{\rm ph} \propto \langle\tau\rangle_E \propto (k_{\rm B}T)^{-3/2}. \tag{5.22}$$

On the other hand, optical phonons have large energies around $k \approx 0$ and do not affect the transport for weak electric field. In hot electron transport, in which the electrons are very far from equilibrium by the effect of strong electric field, the optical phonon scatterings are very important.

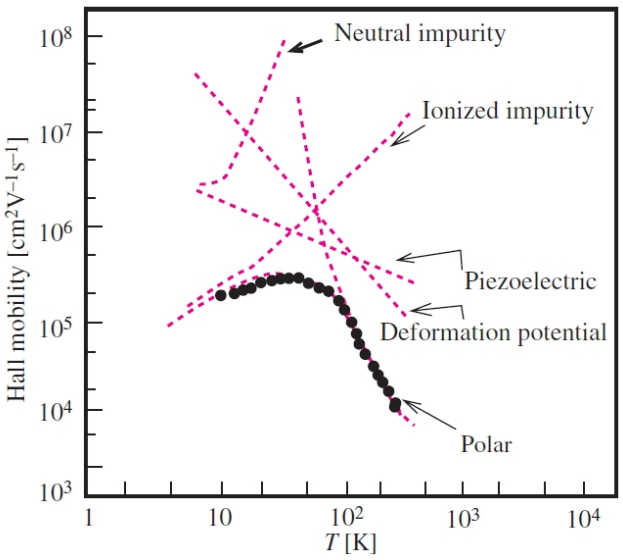


Fig. 5.5 Hall mobility in GaAs(experiments, points) and fitting by putting various scattering mechanisms with temperature dependence into (5.20). Red broken lines indicate temperature dependences of various scattering mechanisms[2].

Ionized impurity scattering : Impurity atoms in solids often emit electrons to become positive or trapped negative ions, forming a Coulomb potential for band electrons and causing scattering. In most cases, such potentials are screened by surrounding charge carriers and have the Yukawa-type ($e^{-r/L_{\rm D}}/r$) distance (r) dependence rather than the Coulomb-type $1/r$. When the ionized impurities have magnetic moments due to the electron spins, they also cause magnetic impurity scattering the the internal freedom causes peculiar effects like the Kondo effect. If there is no internal freedom the scattering is simple potential scattering and elastic.

Scattering by Yukawa potential of carriers with Maxwell distribution, contributes to the scattering time as

$$\tau_{\rm ion} \propto T^{3/2}, \quad \mu_{\rm ion} \propto \frac{T^{3/2}}{\ln(1+x) - \frac{x}{1+x}}, \quad x \equiv \frac{24m^*\lambda k_{\rm B}T}{\hbar^2} \tag{5.23}$$

for weak screening, *i.e.* a long $L_{\rm D}$. For strong screening with a very short $L_{\rm D}$, the scattering is δ-function like and the contribution is

$$\tau_{\rm ion} \propto T^{1/2}, \quad \mu_{\rm ion} \propto T^{1/2}. \tag{5.24}$$

The mobility in GaAs obtained from Hall and conductivity measurements, and the result of fitting by considering various types of scatterings included in eq.(5.20) are shown in Fig. 5.5. The broken lines show temperature dependences of the scatterings. We see all of these limit the mobility.

5.2 Thermal transport and electric transport

In the Boltzmann equation (5.4), the second and the third term in left hand side representing non-uniformity in the phase space, correspond to drift current and diffusion current respectively. In this subsection we treat the thermoelectric effect, in which coexistence of the both types of currents should be considered. A temperature gradient in solids causes a heat current (or thermal flux). Here we consider heat transport by charge carriers, *i.e.* electrons and holes though lattice vibrations (phonons) also carry heat in solids. Below, we do not consider Joule heating for a while.

5.2.1 Thermal conductivity

Thermal flux density along x-direction with carrier concentration n is defined as

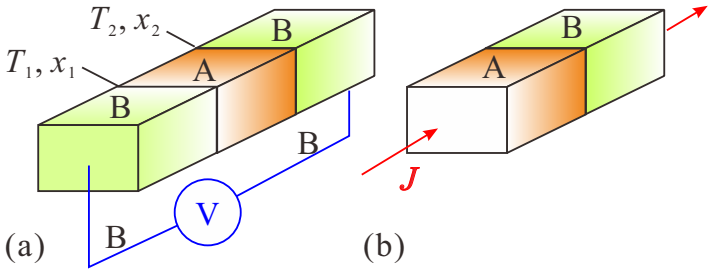
$$j_{qx} = \langle n v_x (E-\mu) \rangle = \int_0^\infty v_x (E-\mu) f(E) \mathscr{D}(E) dE. \tag{5.25}$$

Then thermal conductivity κ_n under temperature gradient ∇T is defined as

$$\kappa_n = -\frac{j_{qx}}{\partial T/\partial x}. \tag{5.26}$$

In vector format $\boldsymbol{j}_q = -\hat{\kappa} \nabla T$.

5.2.2 Thermoelectric effects



The heat flux in (5.26) should also lead to some electric effect. Such complex effects of temperature gradient and electric response are called **thermoelectric effects**.

Let the temperatures at edges of a conductor A T_1 and T_2 respectively. Two conductors of another material B with the same lengths are connected to the edges. Other ends of conductors (material B) are connected to a voltmeter with infinite input impedance ((a) in the left figure). In the steady state, there is no net current and the electric current driven by heat flow should be compensated by the voltage $V_{\rm AB}$ measured at the voltmeter. This is called **Seebeck effect**, and the ratio of the voltage to the temperature difference ($\Delta T = T_1 - T_2$)

$$S_{\rm AB} = \frac{V_{\rm AB}}{\Delta T} \tag{5.27}$$

is called **Seebeck coefficient**. On the other hand as in (b), when there is a junction of A and B set at a uniform temperature, a current J causes heat fluxes $Q_{\rm A}$ and $Q_{\rm B}$. In a steady state there is no charge accumulation and J is uniform, that means $Q_{\rm A}$ and $Q_{\rm B}$ are different reflecting difference in the thermal transport coefficients. The difference results in heating at the interface. This is called **Peltier effect** and the ratio of the heating speed to J,

$$\Pi_{\rm AB} = \frac{Q_{\rm AB}}{J} \tag{5.28}$$

is called **Peltier coefficient**. If we apply a current J to a BAB type junction as in (a), the same current flows with inverted directions through the two interfaces. Hence if a heating occurs at one interface, a cooling of the same amount of heat occurs at the other end.

In a uniform conductor with a current J and a temperature gradient (assume along x-direction) $\partial T/\partial x$, cooling or heating occurs. Heat creation per unit length $\partial Q/\partial x$ is proportional to the product of J and $\partial T/\partial x$. This is **Thomson effect** and the coefficient

$$\tau = \frac{\partial Q/\partial x}{J(\partial T/\partial x)} \tag{5.29}$$

is called **Thomson coefficient**.

Among the above three kinds of coefficient, **Kelvin (Thomson) relations**

$$\Pi_{\rm AB} = S_{\rm AB} T, \quad \tau_{\rm A} - \tau_{\rm B} = T \frac{dS_{\rm AB}}{dT} \tag{5.30}$$

hold (Appendix B). From the relations we can define material specific (combination free) Seebeck coefficient as

$$S_{\mathrm{A}}(T) \equiv \int_0^T \frac{\tau_{\mathrm{A}}(T')}{T'} dT'. \tag{5.31}$$

The relation with the coefficient in (5.27) is

$$S_{\mathrm{AB}} = S_{\mathrm{A}} - S_{\mathrm{B}}. \tag{5.32}$$

In the measurement of Seebeck effect, we need to connect the sample and the voltmeter with leads, which also have Seebeck coefficient. Hence the measured voltage is the difference between the Seebeck effects of the sample and the leads. Equation (5.32) indicates the fact. **Thermocouple** works as a sensor for temperature difference ΔT with knowledge of Seebeck coefficients for the two components.

5.2.3 Boltzmann equation and thermoelectric coefficients

Let us look for the relation between the thermoelectric coefficients and the distribution function with Boltzmann equation under relaxation time approximation (5.4), (5.5). In a steady state $\partial f/\partial t = 0$ we rewrite the equation as

$$\boldsymbol{v} \cdot \nabla f + \frac{\boldsymbol{F}}{m} \nabla_v f = -\frac{f - f_0}{\tau(E)}. \tag{5.33}$$

We take the approximation that the shifts from equilibrium are small and replace f in the left hand side with f_0.

∇f_0 due to temperature gradient ∇T is written as

$$\nabla f_0 = \nabla T \frac{\partial f_0}{\partial T}.$$

In f_0, E and T always appear in the expression $-(E - E_{\mathrm{F}})/k_{\mathrm{B}}T$, which we write a here for short description. Then

$$\frac{\partial f_0}{\partial T} = \frac{\partial f_0}{\partial E}\frac{\partial E}{\partial a}\frac{\partial a}{\partial T} = \frac{\partial f_0}{\partial E}(-k_{\mathrm{B}}T)\frac{E - E_{\mathrm{F}}}{k_{\mathrm{B}}T^2} = \frac{\partial f_0}{\partial E}\frac{E_{\mathrm{F}} - E}{T},$$

$$\text{therefore} \quad \nabla f_0 = \nabla T \frac{E_{\mathrm{F}} - E}{T}\frac{\partial f_0}{\partial E}. \tag{5.34a}$$

$$\text{And} \quad \nabla_v f_0 = \nabla_v E \frac{\partial f_0}{\partial E} = m\boldsymbol{v}\frac{\partial f_0}{\partial E}. \tag{5.34b}$$

When the electric field $\mathbf{E}$ and the temperature gradient ∇T coexist, (5.33) can be written with (5.34) as

$$f = f_0 - \tau(E)\boldsymbol{v} \cdot \left[-e\mathbf{E} + \frac{E_{\mathrm{F}} - E}{T}\nabla T\right]\frac{\partial f_0}{\partial E}. \tag{5.35}$$

We take $\mathbf{E} = (\mathcal{E}_x, 0, 0)$ and the current along x-direction is

$$j_x = -e\langle n v_x\rangle = -e\int_0^\infty v_x f(E)\mathscr{D}(E)dE = e\int_0^\infty v_x^2 \tau\left[-e\mathcal{E}_x + \frac{E_{\mathrm{F}} - E}{T}\frac{\partial T}{\partial x}\right]\frac{\partial f_0}{\partial E}\mathscr{D}(E)dE.$$

The Seebeck coefficient is obtained with putting $j_x = 0$ as

$$\begin{aligned} S = \frac{\mathcal{E}_x}{\partial T/\partial x} &= \int_0^\infty v_x^2 \tau \frac{E_{\mathrm{F}} - E}{eT}\frac{\partial f_0}{\partial E}\mathscr{D}(E)dE \Big/ \int_0^\infty v_x^2 \tau \frac{\partial f_0}{\partial E}\mathscr{D}(E)dE \\ &= \frac{1}{eT}\left[E_{\mathrm{F}} - \int_0^\infty \tau E^2 \frac{\partial f_0}{\partial E}\mathscr{D}(E)dE \Big/ \int_0^\infty \tau E \frac{\partial f_0}{\partial E}\mathscr{D}(E)dE\right]. \end{aligned} \tag{5.36}$$

Here v_x^2 is replaced with $2E/3m$.

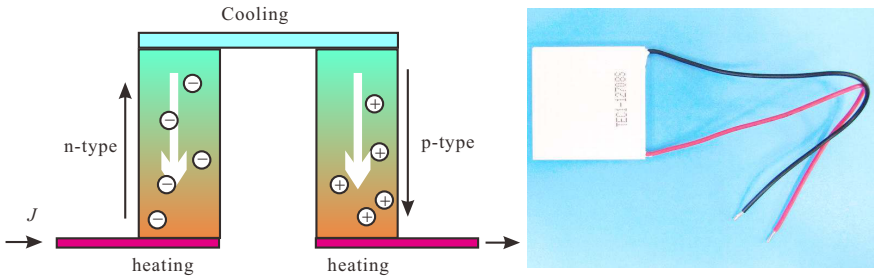


Fig. 5.6 Left panel: Schematic of Peltier device. p-type semiconductors and n-type semiconductors are placed staggered (in the figure just a single pair) along the current path. While electric current meanders heat flows one way. Right panel: Photo of a Peltier device. From Akizuki-denshi web site.

In eq.(5.35) inside the parentheses at right hand side, the first term represents the drift current while the second the diffusion current caused by the temperature distribution. The canceling of these term results in the Seebeck effect, therefore the Seebeck effect is the result of diffusion current which causes charge non-uniformity inside the sample. The non-uniformity creates electric field, of which the drift current cancels the diffusion current.

In Maxwellian approximation, $\partial f_0/\partial E = -f_0/k_{\rm B}T$, and we further assume the energy dependence of the relaxation time as $\tau \propto E^s$, then

$$S = -\frac{1}{eT}\left[\frac{\langle \tau E\rangle_E}{\langle \tau\rangle_E} - E_{\rm F}\right] = -\frac{1}{eT}\left[\left(\frac{5}{2}+s\right)k_{\rm B}T - E_{\rm F}\right]. \tag{5.37}$$

This equation tells that if we can measure the temperature dependence of S, we obtain $E_{\rm F}$ and s. The above calculation is for electrons and for holes $-e$ is replaced with $+e$, hence measurement of S also gives the sign of carriers. This result for Maxwellian approximation does not depend on the carrier concentration, which can be understood as follows. The Einstein relation (5.13) connects the diffusion constant and the mobility, which are material constants for diffusion and drift currents respectively. Hence these constants disappear from the balancing equation leaving the temperature. The carrier concentration also included as the first order in both currents and dropped. In the case of Hall coefficient, the drift current by external field comes into one side and the carrier concentration remains in the expression.

5.2.4 Peltier device

Peltier and Thomson coefficients can also be obtained from the Kelvin relations. Peltier coefficient also changes its sign with that of carriers. In a material with junctions to n and p-type semiconductors, a current flow through this structure thus causes heating at one junction and cooling at the other resulting in a heat flow. Such a device is called **Peltier device**.

Peltier devices once were frequently used in combination with cooling fans for cooling CPUs in PCs. They have long been used where we need cooling without noises such as refrigerators in bedrooms.

Appendex 4C: Lattice vibration in semiconductors (continued)

Continuing from the last time, let us briefly look at the lattice vibration of sphalerite-type crystals as an example of three-dimensional crystal lattice vibration.

4C.2 Lattice vibration in zinc-blende crystals

We consider zinc-blende (ZB) crystals as an example of three-dimensional crystal which has two species of atoms in the unit cell. The Bravais lattice is fcc but the ZB crystalline structure can be considered as an overlapp of two "fcc crystals", in which one atom is placed at the lattice point of fcc lattice. We consider two such fcc-crystals with different atoms with common lattice constant a. We obtain a ZB crystal by placing these two with the shift of $a(1/4, 1/4, 1/4)$.

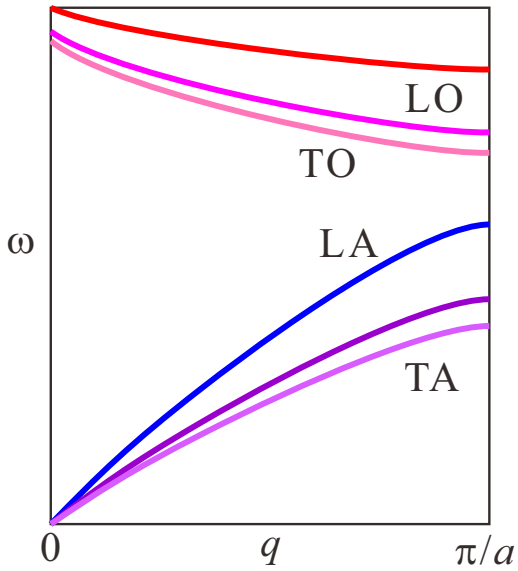


Fig. 4C.3 Schematic diagram of lattice vibration modes (dispersion relation) in a zinc-blende crystal.

Let $\boldsymbol{u}_{\alpha,\boldsymbol{R}}$ be the atomic shift vectors. Here α is the index of the two sublattice, $\boldsymbol{R}$ is the lattice point. The lattice kinetic energy can be written as

$$E_{\rm K} = \sum_{\alpha,\boldsymbol{R}} \frac{1}{2} M_\alpha \dot{\boldsymbol{u}}^2_{\alpha,\boldsymbol{R}}. \tag{4C.6}$$

On the other hand, with expanding the ptential $V(\boldsymbol{r})$ to the second order, the potential energy is written as

$$E_{\rm P} = \sum_{\alpha\alpha',\boldsymbol{R}\boldsymbol{R}',jj'} u^j_{\alpha,\boldsymbol{R}} u^{j'}_{\alpha',\boldsymbol{R}'} \frac{\partial^2 V}{\partial u^j_{\alpha,\boldsymbol{R}} \partial u^{j'}_{\alpha',\boldsymbol{R}'}}, \quad j = x, y, z. \tag{4C.7}$$

The equation of motion can be obtained by defining the Lagrangian $\mathcal{L} \equiv E_{\rm K} - E_{\rm P}$, and general coordinate $q_k \equiv u^j_{\alpha,\boldsymbol{R}}$ from the Lagrange equation;

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_k}\right) - \frac{\partial \mathcal{L}}{\partial q_k} = 0.$$

In the present case from (4C.6), (4C.7) we get

$$M_\alpha \ddot{u}^j_{\alpha,\boldsymbol{R}} = -\sum_{\alpha',\boldsymbol{R}',j'} \frac{\partial^2 V}{\partial u^j_{\alpha,\boldsymbol{R}} \partial u^{j'}_{\alpha',\boldsymbol{R}'}} u^{j'}_{\alpha',\boldsymbol{R}'} \equiv -\sum_{\alpha',\boldsymbol{R}'} \boldsymbol{C}_{\alpha\alpha',\boldsymbol{R}\boldsymbol{R}'} \boldsymbol{u}_{\alpha',\boldsymbol{R}'}. \tag{4C.8}$$

Tensor $\boldsymbol{C}$ corresponds to "force constant" and just depends on the combination $\alpha\alpha'$, and on the relative position of unit cell $\boldsymbol{R}'' = \boldsymbol{R}' - \boldsymbol{R}$. Then with $\boldsymbol{C}_{\alpha\alpha',\boldsymbol{R}\boldsymbol{R}} = \boldsymbol{C}_{\alpha\alpha'}(\boldsymbol{R}'')$ we can write

$$M_\alpha \ddot{u}^j_{\alpha,\boldsymbol{R}} = -\sum_{\alpha,\boldsymbol{R}''} \boldsymbol{C}_{\alpha\alpha'}(\boldsymbol{R}'') \boldsymbol{u}_{\alpha',\boldsymbol{R}+\boldsymbol{R}''}. \tag{4C.9}$$

This equation is invariant againg the shifts among the lattice points $\boldsymbol{R} \to \boldsymbol{R}'$. Then we can write the solution in the form of Bloch function

$$\boldsymbol{u}_{\alpha,\boldsymbol{R}}(t) = \boldsymbol{u}_\alpha(\boldsymbol{q}, t) \exp(i\boldsymbol{q} \cdot \boldsymbol{R}_\alpha). \tag{4C.10}$$

As the time dependence, we consider the oscillation with angular frequency of ω, and assume

$$\boldsymbol{u}_{\alpha,\boldsymbol{r}}(t) = \frac{1}{\sqrt{M_\alpha}} \boldsymbol{u}_\alpha(\boldsymbol{q}, \omega) \exp[i(\boldsymbol{q} \cdot \boldsymbol{R}_\alpha - \omega t)]. \tag{4C.11}$$

Substituting this to (4C.9) results in

$$\omega^2 \boldsymbol{u}_\alpha(\boldsymbol{q}, \omega) = \sum_{\alpha'} \left[\frac{1}{\sqrt{M_\alpha M_{\alpha'}}} \sum_{\boldsymbol{R}} C^{jj'}_{\alpha\alpha'}(\boldsymbol{R}) \exp(i\boldsymbol{q} \cdot \boldsymbol{R}) \right] \boldsymbol{u}_{\alpha'}(\boldsymbol{q}, \omega) \equiv \sum_{\alpha'} \boldsymbol{D}_{\alpha\alpha'}(\boldsymbol{q}) \boldsymbol{u}_{\alpha'}(\boldsymbol{q}, \omega). \tag{4C.12}$$

For (4C.12) to have solutions other than the trivial $\vec{0}$,

$$|D^{jj'}_{\alpha\alpha'}(\boldsymbol{q}) - \omega^2\delta_{\alpha\alpha'}\delta_{jj'}| = 0. \tag{4C.13}$$

The dispersion relations can be obtained by solving this numerically. The 6th order equation gives 6 modes, in which 3 acoustic modes and 3 optical modes exist. The each 3 are separated into 2 transverse modes and 1 longitudinal mode. The namings are, then, TA, LA, TO, LO.

Appendix 5A: Galvanomagnetic effect

We consider the response of drift current to magnetic flux $\mathbf{B}$. In the Boltzmann equation (5.4), $\mathbf{F}$ is taken as $\mathbf{F} = -e(\mathbf{E} + \boldsymbol{v} \times \mathbf{B})$ and the relaxation approximation (5.5) is applied. With $f_1 \equiv f - f_0$,

$$-\frac{e}{\hbar}(\mathbf{E} + \boldsymbol{v} \times \mathbf{B}) \cdot \frac{\partial f}{\partial \boldsymbol{k}} = -\frac{f_1}{\tau} \quad (\boldsymbol{p} = \hbar\boldsymbol{k}). \tag{5A.1}$$

In the first term of the left hand side, f in $\partial f/\partial \boldsymbol{k}$ is replaced with f_0. Form $dE = \boldsymbol{v} \cdot d\boldsymbol{p}$, the second term is $\partial f_0/\partial \boldsymbol{k} = \hbar(\partial f_0/\partial E)\boldsymbol{v}$ and the term of f_0 is orthogonal with $\boldsymbol{v} \times \mathbf{B}$ and vanishes (magnetic field driven force is orthogonal with $\boldsymbol{v}$ and does not give work). In the second term we take terms to f_1 and obtain

$$-e\boldsymbol{v} \cdot \mathbf{E}\frac{\partial f_0}{\partial E} - \frac{e}{\hbar}(\boldsymbol{v} \times \mathbf{B}) \cdot \frac{\partial f_1}{\partial \boldsymbol{k}} = -\frac{f_1}{\tau}. \tag{5A.2}$$

Here we introduce a vector $\mathbf{E}_a$ with the physical dimension of electric field satisfying

$$f_1 = e\tau(\boldsymbol{v} \cdot \mathbf{E}_a)\frac{\partial f_0}{\partial E}. \tag{5A.3}$$

This is from the concept that the Lorentz force shifts the Fermi sphere as in Fig. 5.2 and the origin of the shift is represented as an electric field. Then the equation is

$$-\boldsymbol{v} \cdot \mathbf{E} = -\boldsymbol{v} \cdot \mathbf{E}_a + \frac{e\tau}{m}(\boldsymbol{v} \times \mathbf{B}) \cdot \mathbf{E}_a, \quad \therefore \mathbf{E} = \mathbf{E}_a - \frac{e\tau}{m^*}\mathbf{B} \times \mathbf{E}_a. \tag{5A.4}$$

The solution to eq.(5A.4) is given as follows.

$$\mathbf{E}_a = \frac{1}{1 + \omega_c^2\tau^2}\left[\mathbf{E} + \frac{e\tau}{m^*}\mathbf{B} \times \mathbf{E} + \left(\frac{e\tau}{m^*}\right)^2 (\mathbf{B} \cdot \mathbf{E})\mathbf{B}\right], \tag{5A.5}$$

$$\omega_c = \frac{e|\mathbf{B}|}{m^*}. \tag{5A.6}$$

ω_c is the **cyclotron frequency**. Then f_1 is given as follows.

$$f_1 = \frac{e\tau\mathbf{E}}{1 + \omega_c^2\tau^2} \cdot \left[\boldsymbol{v} + \frac{e\tau}{m^*}\boldsymbol{v} \times \mathbf{B} + \left(\frac{e\tau}{m^*}\right)^2 (\mathbf{B} \cdot \boldsymbol{v})\mathbf{B}\right]\frac{\partial f_0}{\partial E}. \tag{5A.7}$$

We take the case $\mathbf{B} = (0, 0, \mathcal{B}_z)$, $\mathbf{E} = (\mathcal{E}_x, \mathcal{E}_y, 0)$. From $v_z = 0$ and eq.(5A.7), f_1 is calculated as

$$f_1 = e\frac{\partial f_0}{\partial E}\left[v_x\left(\frac{\tau}{1 + (\omega_c\tau)^2}\mathcal{E}_x - \frac{\omega_c\tau^2}{1 + (\omega_c\tau)^2}\mathcal{E}_y\right) + v_y\left(\frac{\omega_c\tau^2}{1 + (\omega_c\tau)^2}\mathcal{E}_x + \frac{\tau}{1 + (\omega_c\tau)^2}\mathcal{E}_y\right)\right]. \tag{5A.8}$$

For example, to obtain $j_x = -en\langle v_x\rangle$ from this equation take the expectation value of v_x with $f = f_0 + f_1$. The expectation value for f_0 is zero and odd components in v is dropped from the integration over $\boldsymbol{k}$. Then

$$j_x = 2\int(-e)v_x f(\boldsymbol{k})\frac{d\boldsymbol{k}}{(2\pi)^3} = -\frac{e^2}{4\pi^3}\int\frac{\tau v_x^2}{1 + (\omega_c\tau)^2}(\mathcal{E}_x - (\omega_c\tau)\mathcal{E}_y)\frac{\partial f_0}{\partial E}d\boldsymbol{k}. \tag{5A.9}$$

The integrand in (5A.9) is the same as that in equilibrium other than v_x^2 and is a function of kinetic energy E. For a general function $\xi(E)$, the principle of energy equipartition gives

$$\int v_x^2 \xi(E) d\boldsymbol{k} = \frac{2}{3m^*} \int E\xi(E) d\boldsymbol{k}. \tag{5A.10}$$

With the Maxwellian approximation $f_0 = A\exp(-E/k_{\rm B}T)$, and density of states $\mathscr{D}(E) = A_D E^{1/2}$, eq.(5A.10) leads to

$$\frac{\partial f_0}{\partial E} = \frac{f_0}{-k_{\rm B}T}, \quad n = A_D \int_0^\infty f_0 E^{1/2} dE = \frac{2A_D}{3k_{\rm B}T} \int_0^\infty E^{3/2} f_0 dE.$$

These being substituted into (5A.9) and we obtain

$$j_x = \frac{ne^2}{m^*} \left[\left\langle \frac{\tau}{1+(\omega_c\tau)^2} \right\rangle_E \mathcal{E}_x - \left\langle \frac{\omega_c\tau^2}{1+(\omega_c\tau)^2} \right\rangle_E \mathcal{E}_y \right], \tag{5A.11a}$$

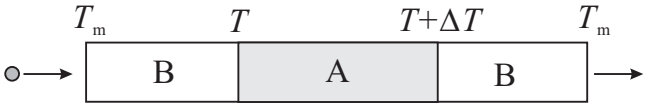
$$\equiv (ne^2/m^*)(A_l\mathcal{E}_x - A_t\mathcal{E}_y) \ \text{ (definitions of } A_l \text{ and } A_t). \tag{5A.11b}$$

$\langle \cdots \rangle_E$ is defined in eq.(5.10). j_y is obtained in the same way and the conductivity tensor in xy-plane is expressed as

$$\boldsymbol{j} = \frac{ne^2}{m^*} \begin{pmatrix} A_l & -A_t \\ A_t & A_l \end{pmatrix} \mathbf{E}. \tag{5A.12}$$

Appendix 5B: Kelvin relations

We consider two species of metals A, B and a junction BAB as shown in the right figure. The temperature at the two edges is kept to $T_{\rm m}$ and a unit charge moves from one edge to the other quasi-statically. As indicated in the figure, temperatures at the two junctions are T and $T + \Delta T$. The voltage between the two edges is $V_{\rm AB}$.



From the requirement of quasi-static assumption, we apply the first and the second laws of thermodynamics to obtain the conditions,

$$V_{\rm BA} + \Pi_{\rm BA}(T) - \Pi_{\rm BA}(T+\Delta T) + (\tau_{\rm B} - \tau_{\rm A})\Delta T = 0$$

$$\frac{\Pi_{\rm BA}(T)}{T} - \frac{\Pi_{\rm BA}(T+\Delta T)}{T+\Delta T} + \frac{\tau_{\rm B} - \tau_{\rm A}}{T}\Delta T = 0.$$

In the differential formula with $\Delta T \to 0$,

$$\frac{dV_{\rm BA}}{dT} - \frac{d\Pi_{\rm BA}}{dT} + \tau_{\rm B} - \tau_{\rm A} = 0, \quad \frac{d}{dT}\left(\frac{\Pi_{\rm BA}}{T}\right) = \frac{\tau_{\rm B} - \tau_{\rm A}}{T}.$$

From the second equation

$$\tau_{\rm B} - \tau_{\rm A} = T\frac{d}{dT}\left(\frac{\Pi_{\rm BA}}{T}\right) = \frac{d\Pi_{\rm BA}}{dT} - \frac{\Pi_{\rm BA}}{T},$$

and we reach

$$\therefore S_{\rm AB} = \frac{\Pi_{\rm AB}}{T}, \quad \frac{dS_{\rm AB}}{dT} = \frac{\tau_{\rm A} - \tau_{\rm B}}{T} \tag{5B.1}$$

with exchange of A and B.

References

[1] M. Lundstrom “Fundamentals of carrier transport” 2nd ed. (Cambridge, 2000).

[2] K. Fletcher, P.N. Butcher, J. Phys. C **5**, 212- 224 (1972).

Chapter 6 Homo·hetero junctions

So far we have seen the bulk properties of uniform semiconductors. Henceforth we go into the rich physical phenomenon in spatially structures semiconductors, the actions as defices.

6.1 Electrical and optical characteristics of homo pn junctions

The pn junction is one of the first semiconductor devices for electric circuits. For the detailed history of the device, see *e.g.* [1] (though in Japanese, out-of-print).

6.1.1 Thermal equilibrium

A pn junction, as it expresses, is a junction of a p-type semiconductor and an n-type semiconductor. Here we consider homojunctions, in which the same species of semiconductor is used for p- and n-layers. In such a junction, the electron density is high in the n-layer and the hole density in the p-layer. Hence there should be diffusion pressures which drive electrons to the p-layer and holes to the n-layer for increase of entropy S. On the other hand, such diffusions charge up the p-layer to negative and the n-layer to positive creating charge double layer at the junction (charge **depletion layer**). This electro-magnetically enhances the internal energy U. In thermal equilibrium, the double layer width is determined from the condition for free energy ($U - TS$) minimum.

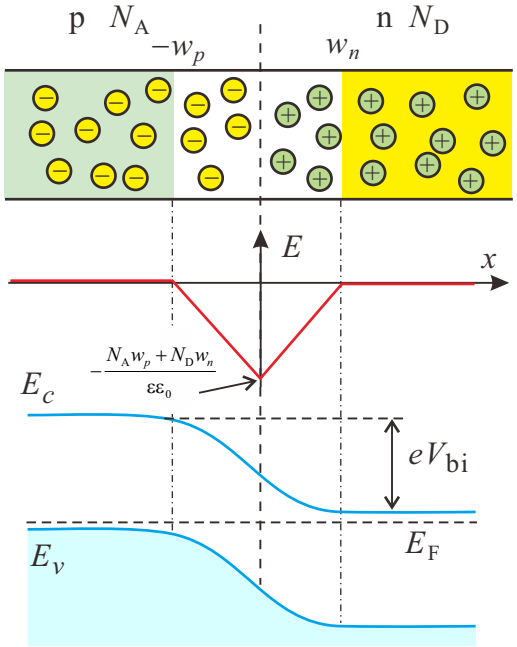


Fig. 6.1 (a) Schematic of an abrupt pn-junction. (b) Electric field $E(x)$ in depletion layer. x-direction is taken as positive for the field. (c) Band diagram.

We take a simple model of an abrupt junction (Fig. 6.1), and $p \sim n \sim n_i$ in the depletion layer. We write the **built-in voltage** due to the pn structure at the interface across the depletion layer $V_{\rm bi}$. In the process that an electron moves from the n-layer to the p-layer, the energy increases by $eV_{\rm bi}$. In the n-layer the electron density $n_n \sim N_{\rm D}$, and in the p-layer the semiconductor equation tells $n_p \sim n_i^2/N_{\rm A}$. We consider a general case that N_1 and N_2 electrons are respectively distributed in two boxes with site number N. The number of cases is $W = {}_N C_{N_1} {}_N C_{N_2}$. Here only particle exchanges are considered hence $dN_1 = -dN_2$. Under assumption $N \gg N_{1,2}$, $d(\ln W) \approx \ln(N_2/N_1) dN_1$ (**mixing entropy** of gases). Applying this to the pn-junction with $dN_1 = -1$, $N_1 = n_n$, $N_2 = n_p$, condition $d(U - TS)/dn_n = 0$

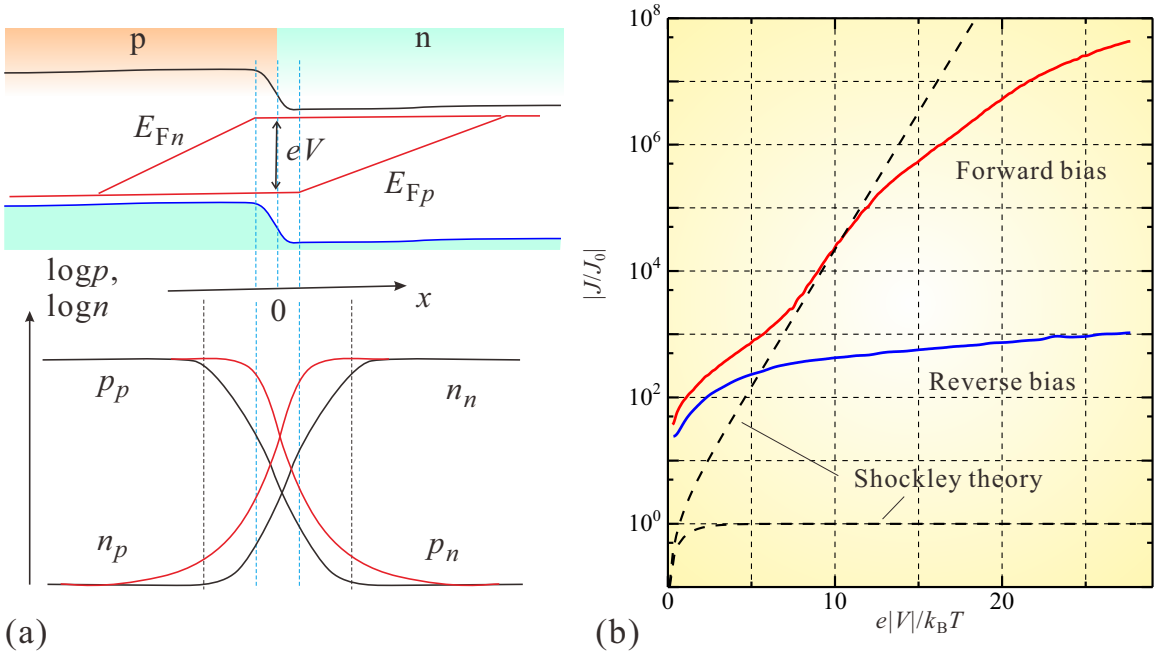


Fig. 6.2 (a) Upper panel: quasi-Fermi levels in a pn-junction under external forward voltage V. Lower panel: Spatial variation of carrier densities. (b) Broken line: I-V characteristics of Shockley theory (eq.(6.11)). Normalized with J_0, which is the coefficient in eq.(6.11). Solid line: Realistic I-V characteristics, in which series resistance, recombination inside the depletion layer, tunneling through localized states are taken into account. The inset is a linear plot of (6.11).

gives

$$eV_{\rm bi} = k_{\rm B}T \ln \frac{n_n}{n_p} \sim k_{\rm B}T \ln \frac{N_{\rm D}N_{\rm A}}{n_i^2} = E_{\rm g} - k_{\rm B}T \ln \frac{N_c N_v}{N_{\rm D}N_{\rm A}}. \tag{6.1}$$

($n_n \sim N_{\rm D}$, $p_p \sim N_{\rm A}$).

In equilibrium it is also required that the chemical potential (Fermi energy) is constant through the junction, independent of the spatial coordinate. Far inside p, n-layers apart from the junction, the band structure should recover the bulk states. Hence the band diagram in Fig. 6.1(c) is drawn. Let the depletion layer widths in p and n-layers w_p, w_n respectively, then $E(x)$ is given as

$$-\epsilon\epsilon_0 E(x) = N_{\rm A}(2x + w_p) + N_{\rm D}w_n \ \ (x < 0), \quad N_{\rm A}w_p + N_{\rm D}(w_n - 2x) \ \ (x \geq 0), \tag{6.2}$$

where ϵ is the dielectric constant. Then $V_{\rm bi}$ is calculated as

$$V_{\rm bi} = \int_{-w_p}^{w_n} (-E(x))dx = \frac{e}{\epsilon\epsilon_0}(N_{\rm D} + N_{\rm A})w_n w_p = \frac{e}{\epsilon\epsilon_0}(N_{\rm D} + N_{\rm A})\frac{N_{\rm D}}{N_{\rm A}}w_n^2 \quad \because w_n N_{\rm D} = w_p N_{\rm A}. \tag{6.3}$$

From eqs.(6.1) and (6.3), we obtain the relation between the doping concentrations and the depletion layer width.

6.1.2 Current-Voltage characteristics

In equilibrium of a pn-junction, the net current is zero as a result of balance between the entropy and the internal energy. An externally applied voltage pushes off the balance and a current flows as a result. When the energy cost is lowered by the voltage, the diffusion current causes **injection of minority carriers**. Minority carrier injection is an action that increases density of minority carrier dynamically. The minority carrier injection breaks the semiconductor equation locally. Even in such circumstances, by introducing **quasi-Fermi level**, we can treat electrons and holes as in quasi-equilibriums and apply the Boltzmann equation to obtain carrier fluxes. The semiconductor equation (law of mass

action) can also be recovered in a bit modified manner. The goal here is to give the net current as a function of external voltage.

We model the effect of external voltage V as follows. All the voltage drops outside the depletion layer are ignored and V is applied inside it. Far from the junction, the current is carried by majority carriers, which have high concentration and the gradient in the chemical potential in such regions is ignorable. Around the depletion layer, imbalance between the internal energy cost and the increase of entropy causes a flow of carriers. V is applied against $V_{\rm bi}$ lowering the barrier for diffusion currents, then the majority carriers flows into the counter layers increasing the minority carrier densities at the depletion layer edges. The injected minority carriers diffuse into the bulk, recombine with majority carriers and disappear. The diffusion-annihilation process forms a exponential decay in the steady minority carrier density distribution.

In the above model, we assume that local thermal equilibrium is attained in each thin layer parallel to yz plane through the carrier-carrier interaction and the particles can be exchanged between neighboring layers. Quasi-Fermi levels, which depends on x-coordinate, for electrons ($\mu_e(x)$) and holes ($\mu_h(x)$) are introduced as follows,

$$n(x) = N_c \exp[-(E_c(x) - \mu_e(x))/k_{\rm B}T], \quad p(x) = N_v \exp[-(\mu_h(x) - E_v(x))/k_{\rm B}T], \tag{6.4a}$$

$$i.e., \;\; \mu_e(x) = E_c(x) + k_{\rm B}T \ln \frac{n(x)}{N_c}, \quad \mu_h(x) = E_v(x) - k_{\rm B}T \ln \frac{p(x)}{N_v}. \tag{6.4b}$$

The diffusion of minority carriers (densities n_p, p_n) is described by the following diffusion equations.

$$D_e \frac{d^2 n_p}{dx^2} = \frac{n_p - n_{p0}}{\tau_e} - G(x), \quad D_h \frac{d^2 p_n}{dx^2} = \frac{p_n - p_{n0}}{\tau_h} - G(x), \tag{6.5}$$

where $G(x)$ represents minority carrier creation *e.g.* by light illumination and in the dark $G(x) = 0$. n_{p0}, p_{n0} are minority carrier concentrations in the bulk regions, $D_{e,h}$, $\tau_{e,h}$ are the diffusion constant and the lifetime respectively (e for electrons, h for holes). Then **minority carrier diffusion lengths** for electrons and holes are

$$L_e = \sqrt{D_e \tau_e}, \quad L_h = \sqrt{D_h \tau_h}. \tag{6.6}$$

The solution for (6.5) (p_n for $x > w_n$, n_p for $x < -w_p$) which satisfies the boundary condition $n_p \to n_{p0}$ ($x \to -\infty$) and $p_n \to p_{n0}$ ($x \to \infty$), is obtained as

$$n_p(x) = \delta n_0 \exp\left(\frac{x + w_p}{L_e}\right) + n_{p0}, \quad p_n(x) = \delta p_0 \exp\left(-\frac{x - w_n}{L_h}\right) + p_{n0}, \tag{6.7}$$

where δn_0, δp_0 are concentrations of injected minority carriers at the edges of the depletion layer. From the definition (6.4b), in the region of diffusion and with ignoring n_{p0}, p_{n0} in (6.7), the quasi-Fermi levels linearly depend on the distances as

$$\mu_e(x) = E_c + k_{\rm B}T \left[\frac{x + w_p}{L_e} + \ln \frac{\delta n_0}{N_c}\right], \quad \mu_h(x) = E_v - k_{\rm B}T \left[\frac{x - w_n}{L_h} + \ln \frac{\delta p_0}{N_v}\right]. \tag{6.8}$$

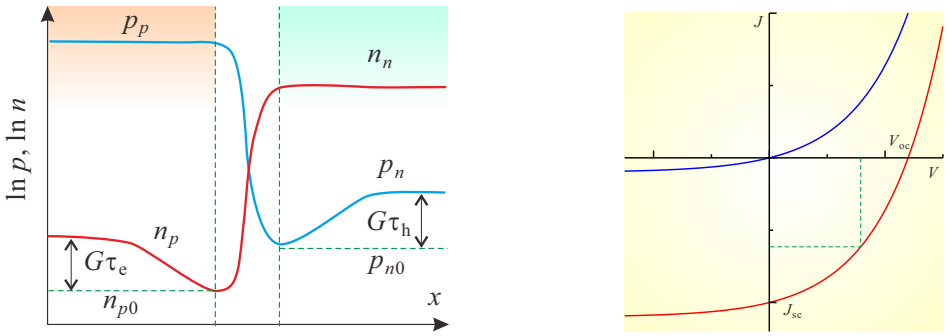
These should join the bulk values $E_{\rm F}^{(p),(n)}$ at $x \to \pm\infty$ respectively and $E_{\rm F}^{(p),(n)}$ differ by eV, *i.e.*, $E_{\rm F}^{(p)} - E_{\rm F}^{(n)} = eV$. Therefore they are schematically drawn as in Fig. 6.2(a).

We ignore electron-hale recombination inside the depletion layer and assume the currents are limited by the diffusion of minority carriers. Then the net current density is the sum of minority carrier diffusion currents at the two edges of the depletion layer. From eq.(6.7) and eq.Fig. 6.2(a),

$$\delta n_0 + n_{p0} = n(-w_p) = n_{p0} \exp \frac{eV}{k_{\rm B}T}, \quad \delta p_0 + p_{n0} = p(w_n) = p_{n0} \exp \frac{eV}{k_{\rm B}T}. \tag{6.9}$$

The electron diffusion current density at $x = -w_p$ in the process (6.5) is thus

$$j_e = eD_e \left.\frac{dn_p}{dx}\right|_{-w_p} = \frac{eD_e \delta n_0}{L_e} = \frac{eD_e}{L_e} n_{p0} \left[\exp \frac{eV}{k_{\rm B}T} - 1\right]. \tag{6.10}$$



(a) (b)

Fig. 6.3 (a) Carrier density distribution around a pn-junction under photo-generation of minority carriers G. Depletion layer edges are indicated by perpendicular broken lines. Bias is taken as shortage $V = 0$. (b) Schematic I-V characteristics in the dark and under illumination.

The hole current can be calculated in the same way and the net current is given as

$$j = e\left[\frac{D_e}{L_e}n_{p0} + \frac{D_h}{L_h}p_{n0}\right]\left[\exp\frac{eV}{k_{\rm B}T} - 1\right] \approx en_i^2\left[\frac{D_e}{L_eN_{\rm A}} + \frac{D_h}{L_hN_{\rm D}}\right]\left[\exp\frac{eV}{k_{\rm B}T} - 1\right]. \tag{6.11}$$

Equation (6.11) is the very basics of the Schottky theory of pn-junction. Though the model grabs the essence, real pn junctions are much more complicated. Important modifications are series resistance, recombination in depletion layer and tunneling conductance through localized level inside energy gap (parallel Ohmic resistance). With these modifications, a realistic characteristics shown in Fig. 6.2(b) differs considerably from the Shockley theory.

6.1.3 Photo-response of pn-junctions

Let us take the simplest model for a pn-junction under illumination assuming majority carrier generation $G(x)$ does not depend on x (a constant G) in the diffusion equation (6.5). Just as before, the solution for $n_p(x)$ and $p_n(x)$ which satisfies the boundary condition $n_p \to n_{n0} + G\tau_e$ for $x \to -\infty$, and $p_n \to p_{n0} + G\tau_h$ for $x \to \infty$ is

$$n_p(x) = n_{p0} + G\tau_e + \left[n_{p0}\left(\exp\left(\frac{eV}{k_{\rm B}T}\right) - 1\right) - G\tau_e\right]\exp\left(\frac{x + w_p}{L_e}\right), \tag{6.12a}$$

$$p_n(x) = p_{n0} + G\tau_h + \left[p_{n0}\left(\exp\left(\frac{eV}{k_{\rm B}T}\right) - 1\right) - G\tau_h\right]\exp\left(-\frac{x - w_n}{L_h}\right). \tag{6.12b}$$

The solution for $V = 0$ is schematically drawn in Fig. 6.3(a).

From the solution, the net current density is given as

$$j = j_0\left[\exp\frac{eV}{k_{\rm B}T} - 1\right] - eG(L_e + L_h), \tag{6.13}$$

where j_0 is the coefficient in front of the parentheses in (6.11). Equation (6.13) is a simple negative shift of (6.11) by $j_{\rm sc} \equiv G(\tau_e + \tau_h)$. Figure 6.3(b) shows the characteristics. Real solar cells are more complicated but the common is the negative shift of the current characteristics with illumination. The parameters which characterize each device are the negative shift at short-circuit condition $|J_{\rm SC}|$ (**short circuit current**) and the voltage at open-circuit condition $V_{\rm OC}$ (**open circuit voltage**). These depend, of course, on the strength and the spectrum of illumination.

In the characteristics shown in Fig. 6.3(b), the cell pumps out an electric energy under the bias condition in the fourth quadrant. Current J and voltage V give power $W = |JV|$. In the forth quadrant $|J| \le |J_{\rm SC}|$, $|V| \le |V_{\rm OC}|$ then $W \le |J_{\rm SC}V_{\rm OC}|$. Then $J_{\max}$, $V_{\max}$ which give the maximum power is determined and

$$FF \equiv \frac{J_{\max}V_{\max}}{J_{\rm SC}V_{\rm OC}} \le 1 \tag{6.14}$$

is called **filling factor** (FF). The better the squareness of the I-V characteristics, the higher the FF. $J_{\rm SC}$, $V_{\rm OC}$, and FF are useful parameters for discussing phenomenology of solar cells, modeling equivalent circuits. In the ideal characteristics (6.13),

$$|J_{\rm SC}| = eG(L_e + L_h), \quad V_{\rm OC} = \frac{k_{\rm B}T}{e} \ln\left[\frac{eG(\tau_e + \tau_h)}{j_0} + 1\right]. \tag{6.15}$$

The above is the basics of photoelectric conversion and applied to *e.g.* solar cells. For the solar cells see the article by the present author [2] (in Japanese).

6.2 *pn*-junction transistors

From left, John Bardeen, William Shockley, Walter Brattain. At AT&T Bell Laboratories, in 1948.

Today, we see two kinds of semiconductor devices invented by a genius named William Shockley. The style of research and development which he began, as well as his devices, has been changing the human life. The above expression is not exaggeration, I believe. I have read a short commentary, which tells "the researchers in Bell Labs. were doing basic research on the surface states of Ge with putting tips on the surfaces and accidentally found the transistor action". But this is far from real situation. Walter Brattain and John Bardeen, who were the direct finders, were doing research aiming at construction of "solid state amplifier" under the team leader Shockley. They did not expect such an easy finding probably but they realized the amplification certainly because they were doing such objective research.

The experiment was done a little before the Christmas of 1947 (said to be 12/16. The application for patent was 12/23) Shockley was out of the labs for a journey. He was thus not so glad hearing the success. Also the transistor (the term is a combination of transfer and resistor) which Brattain and Bardeen accidentally found was called "point contact type", unstable, had low reproducibility. It should have serious obstacles for commercial viability. Their finding might have stimulated Shockley's fight as an inventor, he was absorbed in thought as a theorist aiming at realization of "reproducible device for amplification" and finally got the brilliant inspiration of junction transistor, on the new year's eve allegedly. The theory for the junction transistor established 1/23 in the next year. The experimental realization was a year later. The event was the glorious dawn of the semiconductor physics, in which artificial structures in solids utilize the structural sensitivity of semiconductors and create new functions, new stages of physics[3].

6.2.1 Junction transistor: structure

Figure 6.4 shows basic structure of **junction transistor** (Bipolar Junction Transistor, BJT, at times just "bipolar transistor"), in which two pn-junctions are placed close to each other. npn and pnp are possible types of junctions. An ohmic contact to the central layer is required for the device to have three terminals. The terminals at the two ends are called **Collector** (C), **Emitter** (E) respectively and the central one is called **Base** (B). In the very beginning, the structure was fabricated with alloying metals which work as dopants to both sides of the base material. The naming "Base" came from the fact though lithography and thermal diffusion, ion implantation and epitaxy soon became the dominant methods. As we will see for the transistor action, the base should be very thin. Thinner than the minority carrier diffusion length.

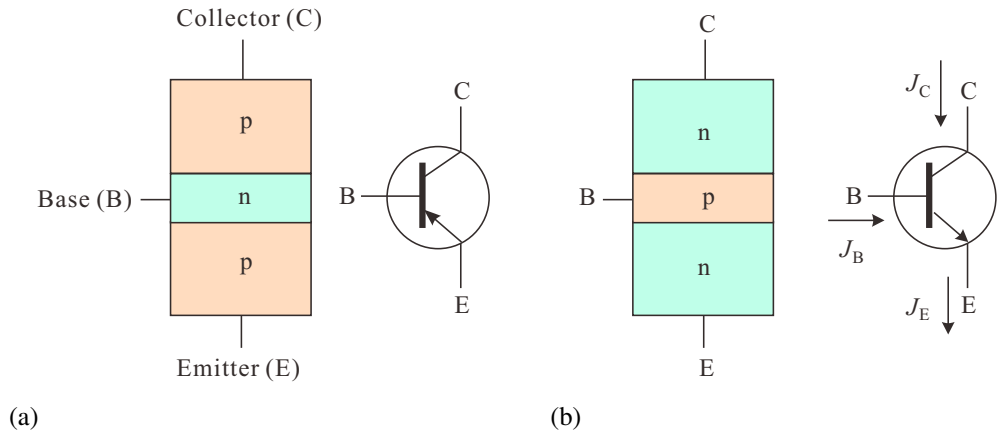


Fig. 6.4 (a) Schematic structure of pnp transistor. Circuit symbol and the names of terminals. (b) Schematic structure of npn transistor and the circuit symbol.

Circuit symbols of transistors are shown in Fig. 6.4, which represent connections of two electrodes to the base graphically. Circles are often omitted. pnp and npn are distinguished with the direction of arrow, which indicates direction of electric current when minority carriers are injected into the base electrode. Below we consider npn-type and define the directions of the current as in the figure.

6.2.2 Current-amplification of junction transistors

In the first experiment a constant voltage source is connected to B-C and collector current $J_{\rm C}$ is measured. Inside the structure B-C is nothing but a pn diode and the result is a well known rectification characteristics ($J_{\rm E} = 0$ in Fig. 6.5(a)). Now we connect a constant current source between E and B, and apply finite currents through E. Because B-E is also a pn junction, the forward bias is positive for B. As shown in Fig. 6.5(a) $V_{\rm BC} - J_{\rm C}$ curve shifts parallely to negative. The amount of shift is almost $J_{\rm E}$.

It should be noted that the characteristics is close to that of a solar cell shown in Fig.2.3(b). The similarity is not a coincidence, rather, the physical situation is almost the same. While In a solar cell, the minority carriers are directly created by photon irradiation, in a transistor, the minority carriers are injected through the pn junction between E and B to the other junction between B and C.

The phenomenon occurring in the junctions are summarised as follows. Here we only describe the phenomenon in conduction band while that in valence band can be discussed in parallel. In an npn junction, a reverse bias voltage to B(p)-C(n) suppresses the diffusion current from the n-layer to the p-layer. The (reverse) diffusion of *electrons* from the p-layer to the n-layer is not enhanced by the reverse bias because all the electrons reach from the p-layer to the junction are swung to the n-layer and it is already saturated at zero-bias. Under the reverse (or zero) bias condition of $V_{\rm BC}$, let the other pn-junction (E-B) be under a forward bias condition. This is possible because an Ohmic contact is attached to the base electrode, hence $V_{\rm EB}$ and $V_{\rm BC}$ can be controlled independently. The forward bias lowers the barrier by the built-in potential in E-B junction and the electrons (majority in the n-layer) diffuse into the base layer and the minority carrier concentration increases in B. This is the phenomenon called **minority carrier injection**, which decays over the **minority carrier diffusion length** through the recombination with majority carriers (holes). Note that the continuity in current is hold. The flow by injected electrons is not driven by the electric field but by the density gradient. So the flux is perpendicular to the junction plane almost ignoring the base Ohmic electrode (the recombination current goes to the electrode). When the B-layer is much thinner than the minority carrier diffusion length, most of the injected carriers reach the other junction enhancing the *reverse current*. In Fig. 6.6(a), this appears as the enhancement of the reverse current, the amount of which is determined that of injected minority carriers. Hence the current does not depend on $V_{\rm BC}$ as long

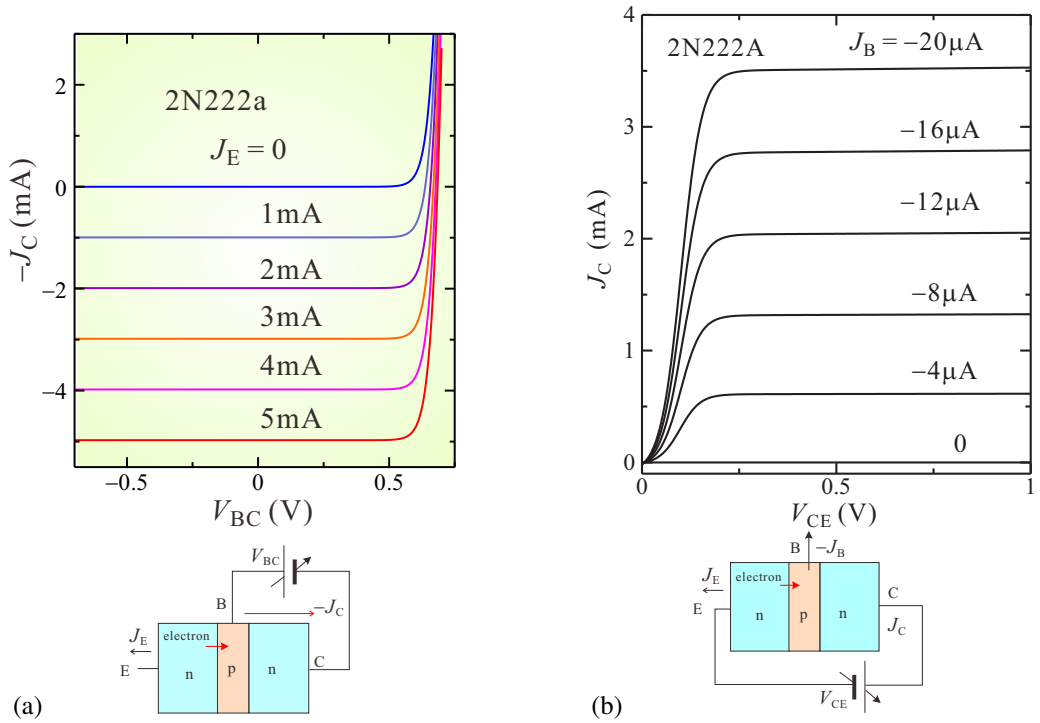


Fig. 6.5 (a) J_C (upside down for convenience) as a function of V_{BC} with J_E as a parameter in the circuit shown in the lower panel. With increasing J_E *i.e.*, injecting electrons from E to B, the characteristics resembles to that of an illuminated solar cell. (b) Application of collector-emitter voltage V_{CE} with floating B, almost no current flows due to the reverse bias in C-B. The biasing B with some currents J_C appears according to J_B showing saturation for V_{CE}.

as there is no forward current.

Now an amplification circuit can be composed as follows. Let the electrodes C-E be voltage biased as in Fig. 6.6(c). The amount of minority carrier injection into B-layer is determined by V_{BE}. Hence in this circuit J_C strongly depends on V_{BE} as shown in Fig. 6.6(a). However, the relation is too non-linear for the use of the device in a voltage-input circuit.

Some of the injected minority carriers recombine with majority carriers and some portion flows out to B-electrode. The base current J_B depends on V_{BE} in the same functional form only but the coefficient as J_C because the pn-junction is the same. J_C is thus proportional J_B, that is,

$$J_C = h_{FE} J_B. \tag{6.16}$$

The good linearity is confirmed in the measurement as shown in Fig. 6.6(b). h_{FE} is called **current amplification factor**. And it is often said that "a bipolar transistor works as a current amplification device" from this face. This is in practice, true as long as we use it as a black box device in electric circuits. However in physical mechanism, as discussed above, there is no such causality that a small current drives a larger current. The following expression may be closer to reality: a small current here is just a monitor for voltage to control a large current.

In the usage of a BJT in a circuit, care should be taken that (because it is a "current amplification device") the input voltage bias should be set to a low differential resistance region. Particularly in high frequency circuits, the impedance matching should be taken to the characteristic impedance of the transmission line. One simple "rule" for transistor circuits is that when a transistor is working as an amplifier, the base-emitter bias voltage should be around the quasi-threshold voltage (though as we saw there is no threshold voltage in pn-junctions, in ordinary circuit scale, the I-V curve seems as if it has).

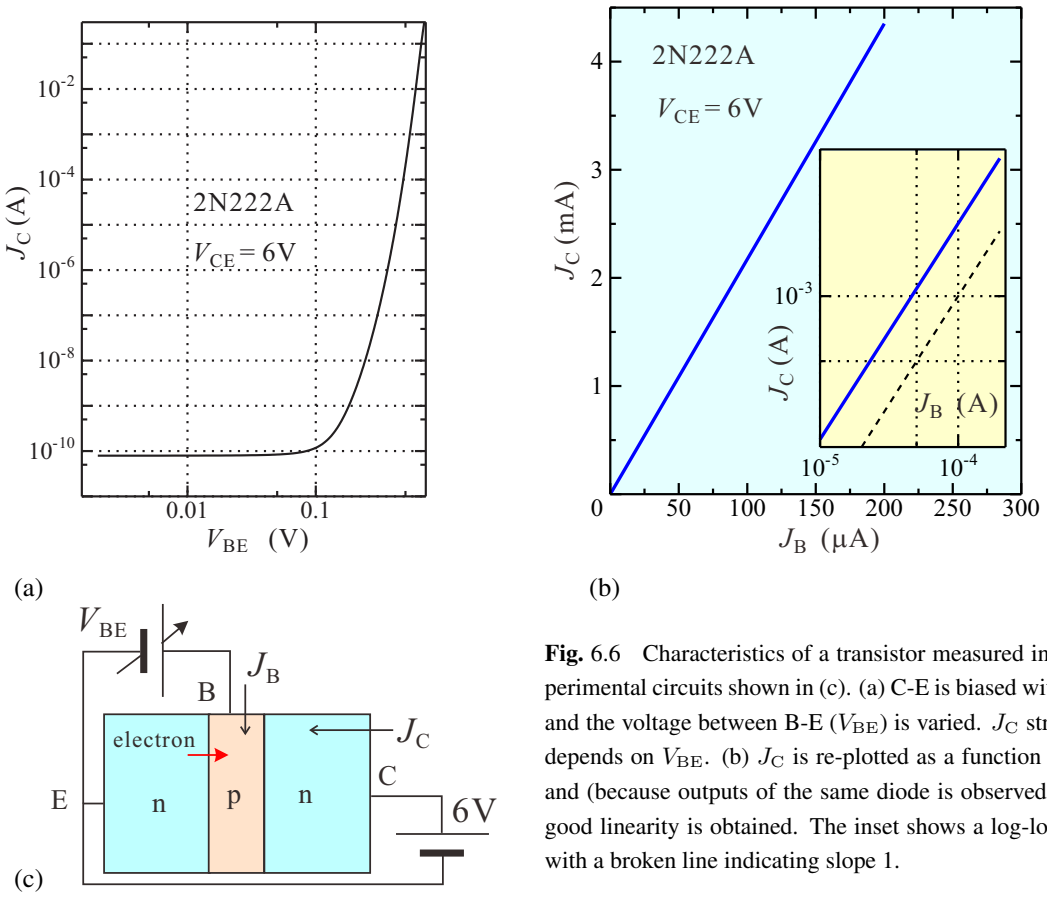


Fig. 6.6 Characteristics of a transistor measured in a experimental circuits shown in (c). (a) C-E is biased with 6 V and the voltage between B-E ($V_{\rm BE}$) is varied. $J_{\rm C}$ strongly depends on $V_{\rm BE}$. (b) $J_{\rm C}$ is re-plotted as a function of $J_{\rm B}$ and (because outputs of the same diode is observed) very good linearity is obtained. The inset shows a log-log plot with a broken line indicating slope 1.

6.3 Field effect transistors I

Field Effect Transistors (FETs) are now used much more widely in circuits than BJTs. And the idea of FET was born even long before that of BJT [*1], but for the realization of FET requires technologies even higher than those for BJT and the realization was later than that for BJT. In these 20 years, Metal-Oxide-Semiconductor (**MOS**) type FETs are mainly used but the first FET was realized for Junction FET (JFET), which utilize pn junctions.

6.3.1 pn-junction and depletion layer

For understanding the device action of JFET, the relation between the reverse bias voltage and the depletion layer is important. We consider a pn-junction shown in Fig. 6.7, with x-dependent potential $\phi(x)$. The Poisson equation is given as

$$\frac{d^2\phi}{dx^2} = -aq(x) \quad (a \equiv (\epsilon\epsilon_0)^{-1}). \tag{6.17}$$

In the space-charge region (depletion layer) we assume abrupt concentration distribution of dopants and sharp cutting of the end of depletion layer. Then

$$\begin{cases} q = -eN_A & (-w_p \le x \le 0), \\ q = eN_D & (0 \le x \le w_n). \end{cases} \tag{6.18}$$

[*1] Shockley wrote a patent on FET before BJT though many similar ideas had existed before that. We cannot say the patent is as unique as that of BJT.

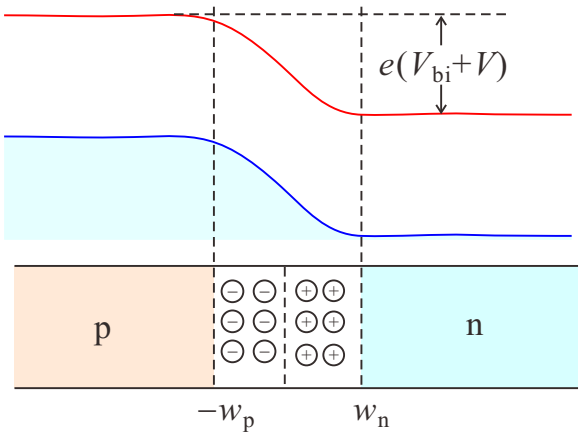


Fig. 6.7 Simple model of a pn junction

Let's take the asymptotic condition as $\phi(-\infty)=0$. When there is external reverse bias voltage V, the boundary condition at the edges of depletion layer is

$$\begin{aligned} \phi(-w_p)&=0, \qquad \left.\frac{d\phi}{dx}\right|_{-w_p}=0, \\ \phi(w_n)&=V+V_{\rm bi}, \quad \left.\frac{d\phi}{dx}\right|_{w_n}=0. \end{aligned} \tag{6.19}$$

Integration of the above gives

$$\phi(x)=\begin{cases} (aeN_A/2)(x+w_p)^2 & (-w_p\le x\le 0), \\ V+V_{\rm bi}-(aeN_D/2)(x-w_n)^2 & (0\le x\le w_n). \end{cases} \tag{6.20}$$

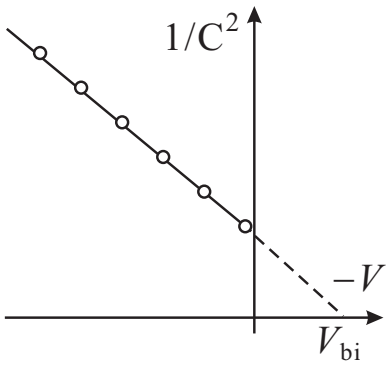
From the condition for the connection at $x=0$

$$\lim_{x\to+0}\phi=\lim_{x\to-0}\phi, \qquad \lim_{x\to+0}(d\phi/dx)=\lim_{x\to-0}(d\phi/dx), \tag{6.21}$$

the widths of depletion layer w_p, w_n are given as follows.

$$w_p=\left[\frac{2\epsilon_0\epsilon(V+V_{\rm bi})}{eN_A}\cdot\frac{N_D}{N_D+N_A}\right]^{1/2}, \quad w_n=\left[\frac{2\epsilon_0\epsilon(V+V_{\rm bi})}{eN_D}\cdot\frac{N_A}{N_D+N_A}\right]^{1/2} \tag{6.22}$$

$$w_d=w_p+w_n=\left[\frac{2\epsilon_0\epsilon(V+V_{\rm bi})}{e}\cdot\frac{N_A+N_D}{N_AN_D}\right]^{1/2}. \tag{6.23}$$



The charge accumulated in the depletion layer on n-side is $Q=eN_Dw_d$ per unit area giving the effective capacitance (differential capacitance) as

$$\frac{dQ}{dV}=eN_D\sqrt{\frac{2\epsilon\epsilon_0}{eN_D}}\frac{1}{2\sqrt{V+V_{\rm bi}}}=\sqrt{\frac{\epsilon\epsilon_0eN_D}{2}}(V+V_{\rm bi})^{-1/2}. \tag{6.24}$$

In a p^+n-structure, that is, $N_A\gg N_D$,

$$w_d\approx\left[\frac{2\epsilon\epsilon_0(V+V_{\rm bi})}{eN_D}\right]^{1/2}\approx w_n. \tag{6.25}$$

This means the depletion layer expands in proportional to the square root of the reverse bias voltage plus the built-in potential.

This relationship is frequently used for characterization of pn-junctions. For example, differential capacitance $C(V)$ can be measured with applying high frequency voltage source with a small amplitude and through the phase shift. We plot the data as shown in the left figure (for the convenience, the horizontal axis is taken to $-V$), $1/C^2$ versus $-V$. If $N_{\rm D}$ is spatially uniform, the data points should be aligned on a line. (6.24) is

valid only for $V > 0$ and $C \to \infty$ cannot be realized. But with extrapolation from $V > 0$ the point $1/C^2 = 0$ can be specified and we obtain $V_{\rm bi}$ from this.

When N_D is not uniform spatially or some deep level traps exist, we obtain information of the spatial distribution from differentiating the plot. Application of pulses in V and analysis of transient response under light illumination or related techniques can bring much of the information inside the semiconductor[4].

6.3.2 Junction Field Effect Transistors

Figure 6.8 shows a schematic drawing of the JFET structure in a cross sectional view. It is for an n-channel, which has two electrodes on the both edges. They are called **Source** (S) and **Drain** (D) respectively. The channel is sandwiched by p^+ layers called **Gates** (G).

The principle of device action is very simple as can be seen in Fig. 6.8. Applying reverse bias to the gates causes expansion of white-colored depletion layer according to eq.(6.23). This makes the conduction channel narrower and enhances the channel resistance up to infinity for pinch-off. Thus the current through the device is controlled by the gate voltage. This is apparently a voltage-controlled device and the input impedance is typically resistance of pn-junction in reverse bias condition. So it is classified into high input impedance device.

A characteristic feature here is that a large source-drain current causes a significant voltage drop across the device, resulting in gradient of effective reverse bias voltage for the channel-controlling depletion layer. Let us see a simple model. As before in the model for pn-junctions, we assume the boundaries between depletion layers and conduction channel are abrupt. Let the gate length L, the thickness of JFET $2w_t$. We take the channel direction along y-axis. The depletion layer with w_d is

$$w_d(y) = \sqrt{\frac{2\epsilon\epsilon_0 V(y)}{eN_D}}, \tag{6.26}$$

where $V(y)$ is local voltage at position y between the channel and the gate. $V(y)$ can be obtained by subtracting voltage along the channel $V_{\rm ch}$ due to the source-drain current from the sum of the built-in potential $V_{\rm bi}$ and the reverse bias gate voltage $V_{\rm g}$.

$$V(y) = V_{\rm g} + V_{\rm bi} - V_{\rm ch}(y).$$

We have no injection of minority carrier and only consider the drift current of majority carriers. The electric field along y-direction is dV/dy. Let the channel depth W and the drift current through the channel is

$$J_{ch} = eN_D\mu_n \frac{dV}{dy} \cdot 2(w_t - w_d)W. \tag{6.27}$$

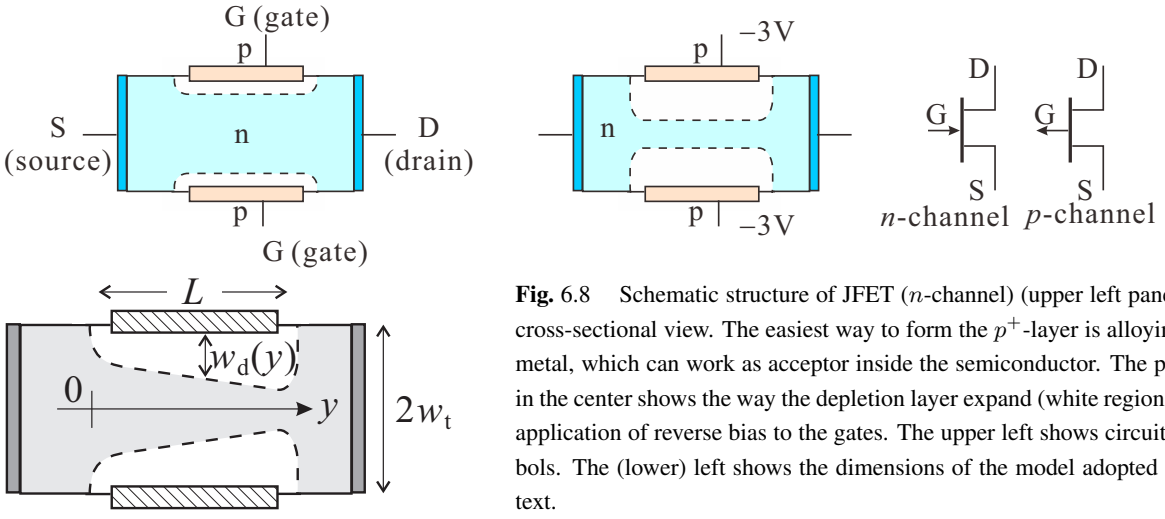


Fig. 6.8 Schematic structure of JFET (n-channel) (upper left panel). A cross-sectional view. The easiest way to form the p^+-layer is alloying the metal, which can work as acceptor inside the semiconductor. The picture in the center shows the way the depletion layer expand (white region) with application of reverse bias to the gates. The upper left shows circuit symbols. The (lower) left shows the dimensions of the model adopted in the text.

In steady state there is no charging up and $J_{\rm ch}$ is uniform through the channel thus integration over the channel should be $J_{\rm ch}L$.

$$J_{\rm ch}L = \int_0^L J_{\rm ch}dy = 2eN_D\mu_n W\int_0^L (w_t - w_d)\frac{dV}{dy}dy = 2w_t eN_D\mu_n W\int_{V_0}^{V_L}\left(1-\frac{w_d}{w_t}\right)dV. \tag{6.28}$$

Let the critical voltage $V_{\rm c}$ at which the channel is pinched ($w_d = w_t$) and $J_{ch} = 0$ then $V_c = eN_D w_t^2/2\epsilon\epsilon_0$. Hence from $w_d/w_t = \sqrt{V/V_c}$, $J_{\rm ch}$ in this model is obtained as

$$J_{\rm ch} = \frac{2N_D e\mu_n W w_t}{L}\left[V_L - V_0 + \frac{2}{3\sqrt{V_c}}(V(V_0)^{3/2} - V(V_L)^{3/2})\right]. \tag{6.29}$$

In eq.(6.29), at small voltages, the first linear term in V_L is dominant and $J_{\rm ch}$ increases linearly. With increasing the voltage, the last $V_L^{3/2}$ term grows and at last the current begins decreasing, which means negative differential resistance. In actual device, this does not occur and $J_{\rm ch}$ simply saturates with increasing V. The model contains various shortages, *e.g.*, the equipotential lines are straight and along x-axis. Improved models can reproduce the saturation but they are inevitably complicated. There are also empirical analytical formulas well fit to the experiments but they have no physical reasoning.

Appendix 6A: Analysis of pn junction transistor

Let us have a brief look at the simplest analysis of charrier statistics in bipolar transistors.

6A.1 Current-voltage characteristics

Figure 6A.1 illustrates the bias conditions and the carrier concentrations in an npn-type transistor. We take the x-axis along the device current direction, and the depletion layer edge at the emitter side of the base is set to $x = 0$. The electron (minority carrier) concentration at $x = 0$ is

$$n_p(0) = n_{p0}\exp\frac{eV_{\rm BE}}{k_{\rm B}T}. \tag{6A.1}$$

They diffuse the base region and reach the depletion edge at the other side $x = W_{\rm B}$. From there the electrons are immediately swept out to the collector by the electric field in the depletion layer. Hence the electron concentration in the vicinity of $W_{\rm B}$ should be very small.

$$n_p(W_{\rm B}) = n_{p0}\exp\frac{-eV_{\rm BC}}{k_{\rm B}T} \approx 0. \tag{6A.2}$$

Providing that $W_{\rm B}$ is much shorter than the minority carrier diffusion length, we can ignore the carrier recombination and the diffusion current in the base is constant. Equation (5.12) tells the current is proportional to dn_p/dx. Hence n_p varies

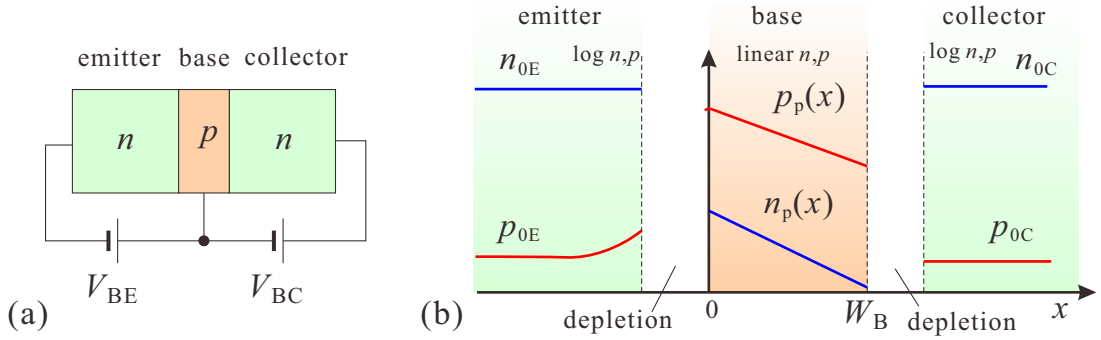


Fig. 6A.1 (a) Biasing condition of the npn transistor under consideration. (b) Schematic diagram of carrier concentrations in a npn type transistor. In the base the ordinate is in linear scale while logarithmic in other regions.

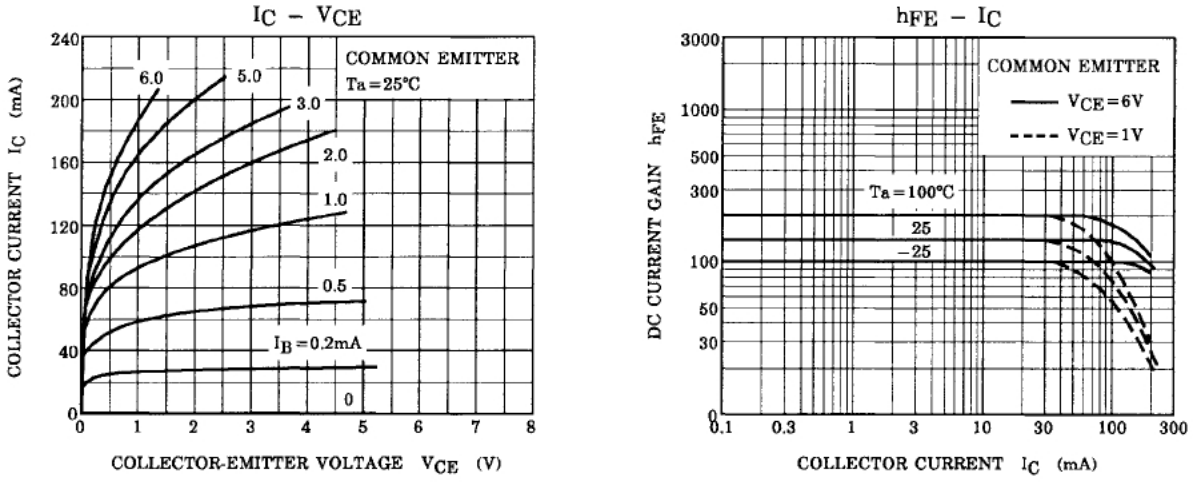


Fig. 6A.2 Characteristics of transistor 2SC1815 for small signal amplification (from the datasheet). Left panel: Collector-emitter voltage $V_{\rm CE}$ dependence of collector current $J_{\rm C}$ for various base current $J_{\rm B}$.

linearly against x as illustrated in the figure (the concentration is in the linear scale only in the base region). Hence from (6A.2) the diffusion current density in the base is

$$j_{\rm De} = -D_e \frac{dn_p}{dx} \approx eD_e \frac{n_p(0)}{W_{\rm B}}. \tag{6A.3}$$

This is the major part of the collector current and the collector current is with A as the cross setion of the device,

$$J_{\rm C} = eAD_e \frac{n_p(0)}{W_{\rm B}}. \tag{6A.4}$$

From the semiconductor equation $n_{p0} \approx n_i^2/N_{\rm A}$,

$$J_{\rm C} \approx \frac{eAD_e n_{p0}}{W_{\rm B}} \exp\frac{eV_{\rm BE}}{k_{\rm B}T} \approx \frac{eAD_e n_i^2}{W_{\rm B}N_{\rm A}} \exp\frac{eV_{\rm BE}}{k_{\rm B}T} \equiv J_{\rm S} \exp\frac{eV_{\rm BE}}{k_{\rm B}T}. \tag{6A.5}$$

$J_{\rm S} = eAD_e n_i^2/W_{\rm B}N_{\rm A}$ is the coefficient which is inversely proportional to $W_{\rm B}N_{\rm A}$.

On the other hand, the base $\rightarrow$ emitter is in forward biasing condiction while base $\rightarrow$ collector is reverse biased. Hence most of the base current flows to the emitter, which is determined by the hole diffution current in the emitter. The calcution is along the same line as the above collector current and the diffusion part of the base current is

$$J_{\rm Bh} = \frac{eAD_h}{L_h} p_{n\rm E}(0) = \frac{eAD_h}{L_h} p_{n\rm E0} \exp\frac{eV_{\rm BE}}{k_{\rm B}T} = \frac{eAD_h}{L_h} \frac{n_i^2}{N_{\rm D}} \exp\frac{eV_{\rm BE}}{k_{\rm B}T}. \tag{6A.6}$$

In the base, minority carrier concentration is enhanced and the recombination current may give some contribution. The total charge of the minority carriers is $Q_e = -en_p(0)W_{\rm B}A/2$. Let τ_b be the minority carrier life time and the recombination current is

$$J_{\rm Br} = \frac{Q_e}{\tau_b} = \frac{en_p(0)AW_{\rm B}}{2\tau_b} \exp\frac{eV_{\rm BE}}{k_{\rm B}T}. \tag{6A.7}$$

Therefore the base current is written as the sum of the above as

$$J_{\rm B} = eA\left(\frac{D_h}{L_h}\frac{n_i^2}{N_{\rm D}} + \frac{n_{p0}W_{\rm B}}{2\tau_b}\right) \exp\frac{eV_{\rm BE}}{k_{\rm B}T}. \tag{6A.8}$$

Then from (6A.5) and (6A.8), the current gain is obtained as

$$h_{\rm FE} = \left(\frac{D_h}{D_e}\frac{W_{\rm B}}{L_h}\frac{N_{\rm A}}{N_{\rm D}} + \frac{W_{\rm B}^2}{2\tau_b D_e}\right)^{-1}. \tag{6A.9}$$

6A.2 Effect of depletion layer width

Figure 6A.2 shows the characteristics of a transistor numbered 2SC1815 (Toshiba, Co. Ltd.). The right panel shows $h_{\rm FE}$ as a function of $J_{\rm C}$. $h_{\rm FE}$ is almost constant in the low $J_{\rm C}$ region indicating good linearity. On the other hand, the left panel shows $J_{\rm C}$ as a functionof $V_{\rm CE}$ with $J_{\rm B}$ as a parameter. In this panel, in the region $V_{\rm CE} \approx 0$, the base-collector is forward biased and not in the region of current amplification. Even in the current amplificatio region, $J_{\rm C}$ increases with $V_{\rm CE}$. This is called the Early effect caused by the widening of the depletion layer thus by the thinning of the base width $W_{\rm B}$.

Let ΔW be the variation in the width of base width and the collector current is given as

$$J_{\rm C} = eAD_e \frac{n_p(0)}{W_{\rm B} - \Delta W} \approx eAD_e \frac{n_p(0)}{W_{\rm B}} \left(1 + \frac{\Delta W}{W}\right) \equiv J_{\rm C0} \left(1 + \frac{\Delta W}{W}\right). \tag{6A.10}$$

ΔW grows rapidly with $V_{\rm CE}$ as in (6.23) when $V_{\rm CE}$ is small while the rate lowers with $V_{\rm CE}$. In Fig. 6A.2, such tendency is apparent. In Fig. 6.5(b), the Early effect is small and the increase in $J_{\rm C}$ can be approximated to be linear in $V_{\rm CE}$.

付録 6B：Deep level transient spectroscopy (DLTS)

Here I would like to give qualitative explanation on the basic principles of Deep Level Transient Spectroscopy (DLTS). For details, see *e.g.* ref. [4]. We consider modification to effective capacitance (6.24), which depends on the reverse bias voltage V. Let $N_{\rm D}$ be the shallow donor concentration, $N_{\rm P}$ the one for a deep donor. In the region where this deep donor responds to change in the bias voltage, the voltage-differential capacitance is expressed as a function of reverse voltage V as

$$w_d(V) \approx \left[\frac{2\epsilon\epsilon_0(V + V_{\rm bi})}{e(N_{\rm D} + N_{\rm P})}\right]^{1/2} \approx w_n, \tag{6B.1}$$

$$C(V) = \sqrt{\frac{\epsilon\epsilon_0 e(N_{\rm D} + N_{\rm P})}{2}} (V + V_{\rm bi})^{-1/2}. \tag{6B.2}$$

For simplicity, we consider the situation that the reverse bias V_p is applied and kept for sufficiently long time for electrons to escape from the depletion layer including the deep levels [*2]. Now V is abruptly lowered to $V_0 < V_p$ and the carriers are captured by the donor levels within $w(V_0) < x \leq w(V_p)$. Shallow donors have high capture rate and can respond within ms without delay, deep levels, on the other hand, the capture rate strongly depends on temperature and with decreasing temperature, the average time for capture often elongates from ms to s, min, hour and sometimes day. Then if we open up a fixed time window and observe the time evolution of C, the time dependence is observed in the time window at some temperature range and in low or high temperature regions the effect of deep levels does not observed.

Such a process is illustrated in Fig. 6B.1(a). We take $t = 0$ at the time the reverse bias is changed:$V_p \to V_0$ and measure the difference in the differential capacitances at t_1 and t_2: $\Delta C = |C(t_1) - C(t_2)|$ as a function of temperature T.

We now assume existence of two species of deep donors, which have temperature dependent capture cross sections shown in the upper panel of Fig. 6B.1(b). ΔC should show two peaks in the temperature dependence. Analysis of the data gives the concentration and capture cross section of each deep level, and combination with photo-response, in some

[*2] At low temperatures the capture/emission rates of deep levels become very small and it is not rare that we need days for the emission. So this condition is, in general, hard to be fulfilled. But the consideration of this does not give significant change and thus we adopt the assumption.

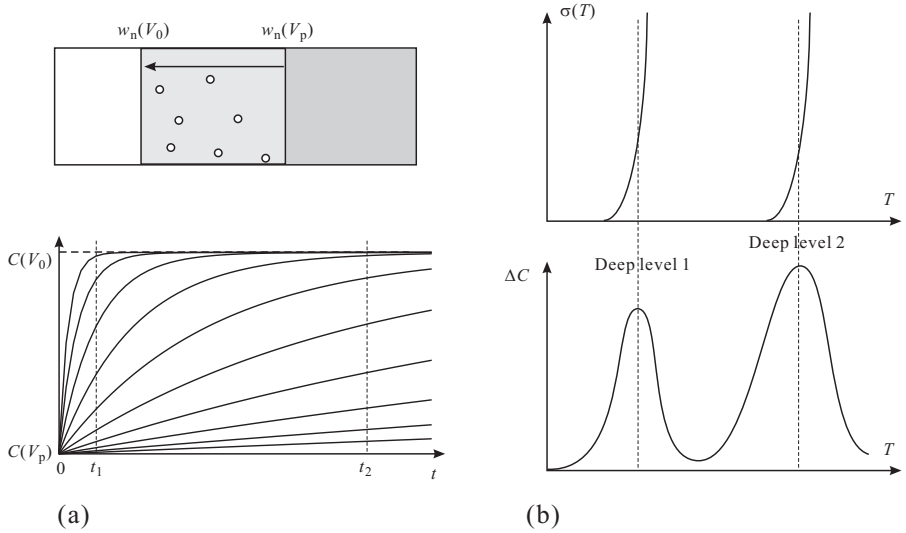


Fig. 6B.1 (a) Upper panel: Illustration that the change in the reverse bias $V_p \rightarrow V_0$ makes shallow levels and a part of deep levels ready for catching carriers. Lower panel: With progress in capture of carriers, differential capacitance $C(V)$ shows transient response. (b) Upper panel: two deep levels exist and assumed temperature dependences of the capture cross section σ are illustrated. Lower panel: shows how the DLTS signal appears from the temperature dependence $\sigma(T)$.

cases identification of deep levels or at least energy positions can be measured[4]. With variation of V_0 and V_p, depth profile of deep levels can be obtained also.

References

[1] 菊池誠「半導体の理論と応用 (中)」(裳華房, 1963).

[2] 勝本信吾 物性研究 電子版 Vol.3, No.3, 033209 (2014 年 8 月号)
http://mercury.yukawa.kyoto-u.ac.jp/~bussei.kenkyu/pdf/03/3/9999-033209.pdf

[3] Jon Gertner, “The Idea Factory: Bell Labs and the Great Age of American Innovation”, (Penguin Press, 2012).

[4] 国府田隆夫, 柊元宏 「光物性測定技術」(東大出版会, 1983).

[5] M. Jaros, “Deep levels in Semiconductors” (CRC Press, 1982).

Next we see FETs without pn-junction. For transistor action, they utilize phenomena on the surfaces or interfaces. In homo-type pn-junctions the uniformity of space is broken by impurity doping. They do not use interfaces or surfaces. This was important for Shockley and co-workers to realize "stable and reproducible" devices because for the semiconductor technologies in those days control of surfaces or interfaces was too difficult for commercial production. Even the high quality crystalline growth and the accurate doping technique, which are indispensable for the realization of pn-junctions, were surprisingly high technique. However the great strides in semiconductor technologies caught the control techniques of surfaces and interfaces in incredibly short time. Naturally there were movement to utilize them for device actions and they overwhelmed bulk shortly. We have a look for these representative modern devices here. But the limit of miniaturization inevitably requires three dimensionality nowadays and we do not know what happens next.

6.3.3 Schottky barrier (junction)

Here we consider junctions between semiconductors and metals. Simple guiding principles are

1. Rigid band approximation,
2. Recovery of bulk states away from the junction,
3. In equilibrium $E_{\rm F}$ (μ) is constant over the space.

On semiconductor surfaces, there usually are **surface states** with high density of states. Metal-semiconductor junctions are strongly affected by those states. Here, however, we first look what Anderson's rule tells about the interface[**?**]. The baseline of rigid bands can be taken to an edge of "band", in which electrons can freely travel between the metal and the semiconductor. It is usually impossible to find such an energy band inside insulators and semiconductors, which have very different energy bands. Then such a "band" can be found as the vacuum levels. Then the excitation energy required is so called **work function**. Let the work functions in the semiconductor and the metal $e\phi_{\rm S}$ and $e\phi_{\rm M}$ respectively. Generally $e\phi_{\rm M} \neq e\phi_{\rm S}$. On the other hand, from the guiding principle 2., the bulk $E_{\rm F}$'s in the metal and in the semiconductor away from the junction should be the same. And $E_{\rm F}$ should be constant throughout the system.

The following procedure, of course, is not real physical process but just a virtual process inside human brain, for construction of consistent band alignment. The final result, however, may be realized in the model of junctions though there still remain many idealizations and reality should be much more complex.

We assume $e\phi_{\rm M}$ is larger than $e\phi_{\rm S}$, the semiconductor is doped to n-type and the donor concentration is $N_{\rm D}$. We make the vacuum levels in the both sides fit to each other and extrapolate the bulk band structures to the interface to obtain the band alignment shown in Fig. 6.9(a). Here the Fermi level in the semiconductor places higher than that in the metal causing flow of carriers from the semiconductor to the metal. The carrier flow generates charge accumulation at the interface creating an electric field perpendicular to the junction plane. The metallic side is also charged up but it has much higher charge concentration, which screens the electric field within the screening length less than a lattice constant making the band bending negligible in this side. Let the accumulated charge in the metal side per unit area $-Q$, in the semiconductor side ($x > 0$, interface at $x = 0$), the electric field at x is $(eN_Dx - Q)/\epsilon\epsilon_0$ and the potential difference between 0 and x_d is

$$\phi(x_d) = \int_0^{x_d} (eN_Dx - Q)/\epsilon\epsilon_0 dx = \frac{1}{\epsilon\epsilon_0}\left(\frac{eN_D}{2}x_d^2 - Qx_d\right). \tag{6.30}$$

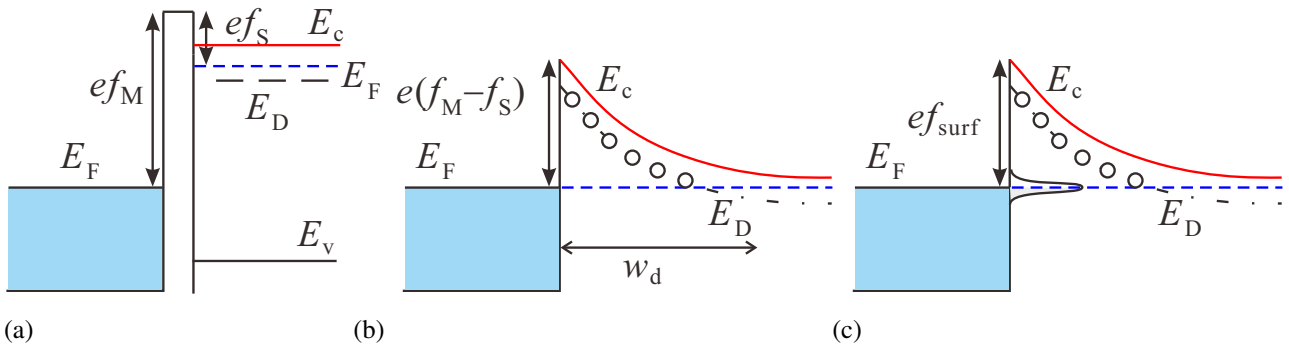


Fig. 6.9 (a) Virtual band alignment, in which a metal and a semiconductor are connected as the vacuum levels for them agree. (b) Band bending effect to make E_F constant throughout the junction is superposed to the alignment in (a). The situation corresponds to an ideal interface without surface states at the semiconductor side. (c) Illustration of Fermi level pinning by surface states. The surface potential ϕ_{surf} is determined by the position of the dominant surface states from the band edge E_c. This usually has nothing to do with the difference between the work function.

Let the space charge (depletion) layer width be w_d. The condition that electric field outside the depletion layer should be zero, gives $w_d = Q/eN_D$. On the other hand, the condition $e\phi(w_d) = \phi_M - \phi_S$ also gives Q as

$$Q = \sqrt{2\epsilon\epsilon_0 N_D e(\phi_M - \phi_S)}, \quad \therefore w_d = \sqrt{\frac{2\epsilon\epsilon_0(\phi_M - \phi_S)}{eN_D}} \equiv \sqrt{\frac{2\epsilon\epsilon_0 V_s}{eN_D}}. \tag{6.31}$$

Here we write $eV_s \equiv \phi_M - \phi_S$. Now we can illustrate the band structure for electrons (holes for p-type) around the metal-semiconductor interface as in Fig. 6.9(b), showing a potential barrier, which is called **Schottky barrier**.

An external voltage V is mostly bared in the semiconductor side, and the height of the barrier changes to $e(V_s - V)$ while the height from the metal side remains as eV_s. To be more accurate, we need to consider the kinetic energy distribution in the semiconductor and count the number of electrons which go over the barrier. But here for simplicity we assume the kinetic energy of electrons in the semiconductor is a constant. Then the equation for thermal electron emission from metallic surface can be applied to obtain

$$J = AT^2\left[\exp\left(\frac{e(V - V_s)}{k_B T}\right) - \exp\left(\frac{-eV_s}{k_B T}\right)\right] = eAT^2 \exp\left(\frac{-eV_s}{k_B T}\right)\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right]. \tag{6.32}$$

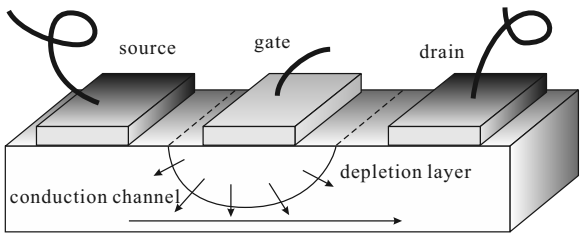
Here A is the **Richardson coefficient**. The first term is current from the semiconductor side, the second is that from the metal side. The current-voltage characteristics is similar to that of a pn-junction with the Schttkey barrier height corresponding to the built-in potential.

In the above the surface of semiconductor is too much idealized for it to have no surface states. However in real metal-semiconductor junctions, current-voltage characteristics are similar to eq.(6.32). One big difference is in eq.(6.32), the barrier height should change with changing the metal species but in reality, the barrier height is almost constant for semiconductor species and independent of metals. This is due to the **surface states** on the semiconductors. The surface states have narrow energy widths, very high density of states pinning the Fermi level to the center of them. Hence the band bending exists even before the connection to metals and the alignment is accomplished between the metal E_F and the surface states. This is called **pinning of Fermi level** by the surface states.

Once the Fermi level is pinned by the surface states, the band bending is determined by semiconductor species. Hence when n-type Schottky barrier can be formed for a semiconductor for example, p-type is not available for the same semiconductor. The other way around. Actually, for GaAs, p-type Schottky barrier is not available while for InP, n-type Schottky barrier is difficult. This makes it difficult to obtain complementary devices which utilize Schottky barriers. In the case of metal-oxide-semiconductor (MOS) devices, an **inversion layer** formed by *e.g.*, pushing down a band of a p-type semiconductor and turning it to an n-type channel, can be used for complementary device. This is, however, impossible for Schottky devices.

6.3.4 MES-FET

Among III-V semiconductors, GaAs is frequently used for electric devices as well as for optical devices. But it is difficult to form good quality oxide layers on the surfaces, hence no MOS type device for GaAs is available. Instead, MEtal-Semiconductor FET (MES-FET) structure has been frequently adopted. GaAs has light electron mass, high mobilities. And the effective capacitance of Schottky diode can be small. Hence GaAs MESFETs are often used for high-frequency application.



As shown in the left figure, the structure of MES-FET is simple. The conduction channel thickness is controlled with the reverse bias voltage (**gate voltage**) through that of depletion layer. The device action, characteristics are similar to those for JFET. Schottky junctions have larger leak current in gate characteristics, only single carrier type is available and complementary circuits cannot be composed with them. These properties are great obstacles for large scale integration. MES-FETs are still widely used as high frequency devices for *e.g.*, microwave.

6.3.5 MOS structure

As named, a thin oxide film for insulation is inserted between a metal and a semiconductor in a Metal-Oxide-Semiconductor (MOS) structure. Needless to say, most frequently used Si has SiO_2 as the oxide layer, which is very stable and has good insulation characteristics. An SiO_2 film can be easily formed with thermal oxidation onto a Si. Both p-type and n-type channels can be controlled and Complementary MOS (CMOS) circuits are easily realized. Also with low gate leakage current, high on-conductance, off-resistance, the power consumption in logic circuits jumped down with the CMOS circuits hence increased degree of integration. Now CMOS is doubtlessly the king of semiconductor circuits. A few decades ago high speed logic circuits were mainly composed with Emitter Coulpled Logic (ECL) of BJT but the requirement of large scale integration and the increase of cut-off frequency in CMOS circuit have made drastic change and now, even so called supercomputers are using CMOS circuit in CPU.

MOSFET structure also resembles to JFET and the essential difference to MESFET is the existence of thin oxide layer between the semiconductor and the gate metal. In a **depletion type** device, the conduction channel is pinched by depletion layer while in a enhancement type device, the band is pushed down with gate electric field to form conduction channel. An oxide layer bears much higher voltage than a Schottky barrier, hence with a strong bending, *e.g.*, formation of an n-type two-dimensional conduction channel below a p-type semiconductor surface (**inversion layer**).

Si-MOS structures are now used not only in integrated circuits but also for power devices. Recently however, SiC is collecting wider interest for power devices because of the lower ON-resistances. And for high-frequency power devices,

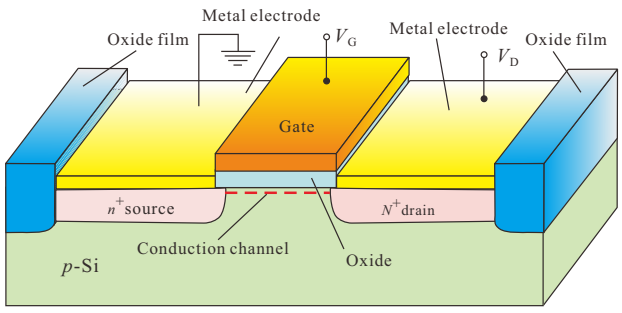
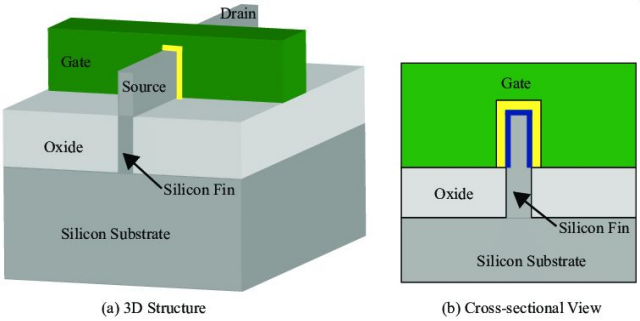


Fig. 6.10 Schematic view of a MOSFET device. In fabrication holes are opened on thermally oxidized films with lithography. The dopants are diffused through the holes. The structure like this often appears due to the process.

the weight of developement is shifting to GaN-related materials.



(a) 3D Structure (b) Cross-sectional View

6.3.6 FinFET

In the beginning of 21st century, in semiconductor integrated circuits, fierce competition for improving the degree of integration continued, centered on CMOS. Compared to other logic circuit schemes, CMOS is overwhelmingly advantageous in terms of power consumption. Even for such CMOS scheme, what limits the degree of integration is in-chip heating due to power consumption. In order to solve this, device driving at low voltage has come to be required. A way is to replace SiO_2 with some other insulating thin film that has higher dielectric constant. With this, ON/OFF action of the channels would be available. The letter κ is often used as the symbol of the dielectric constant and such dielectrics are called "high-κ" or simply "high-k" materials. From various restrictions, now hafnium silicate, hafnium oxide, zirconia are used for such high-κ materials.

Furthermore, the FinFET, in which the channel shape is changed from the planer type to the fish-fin shaped, has been now widely used. As illustrated in the figure, in a FinFET, a thin channel is covered with the gate and the depletion layer/ the inversion layer grow over the channel from both sides of the "fin", resulting in faster switching rate (less than 1 ps) and higher ON-conductance than those of planer structure. Also the device density can be higher. Now they are the main structure for the logic LSIs.

6.4 Heterojunction

The two materials on the sides of a junction have similar properties, lattice structures, etc. to each other in semiconductor **heterojunction** in comparison with Schottky junctions or with MOS structures. As a result, in semiconductor heterojunctions, sharp changes in the effective potential can be realized, the quantum coherence of electrons is kept over the junctions. Therefore they can be used for the building block of the devices which utilize quantum effects such as electron tunneling. And with heterojunctions, one even can create new periodic structures in solids and modify the band structure. This is called **band engineering**.

Because this lecture is for the physics in semiconductors, we begin with how to treat such hetero-interface physically.

6.4.1 Effective mass approximation at hetero-interfaces

As in the textbook[1] written by myself or that by Bastard[2], let us consider hte application of effective mass approximation for a hetero-interface under simplest case.

6.4.1.1 Hetero-interface

We consider the situation in which semiconductors A and B (A : $z < 0$, B : $z > 0$) are connected at $z = 0$ (xy-plane). In each region, the Bloch theorem is appled to write

$$\psi^{(\mathrm{A})}(\boldsymbol{r}) = \sum_l f_l^{(\mathrm{A})}(\boldsymbol{r}) u_{l\boldsymbol{k}}^{(\mathrm{A})}(\boldsymbol{r}), \quad \psi^{(\mathrm{B})}(\boldsymbol{r}) = \sum_l f_l^{(\mathrm{B})}(\boldsymbol{r}) u_{l\boldsymbol{k}}^{(\mathrm{B})}(\boldsymbol{r}), \tag{6.33}$$

where l is the band index, $u_{l\boldsymbol{k}}^{(\mathrm{A,B})}$ are functions with the lattice periodicity. For simplicity, the lattice periodic part of the Bloch functions and the band dispersions are the same other than the positions of band bottoms and tops.

$$u_{l\boldsymbol{k}}^{(\mathrm{A})}(\boldsymbol{r}) = u_{l\boldsymbol{k}}^{(\mathrm{B})}(\boldsymbol{r}), \quad \partial\epsilon_l^{(\mathrm{A})}/\partial\boldsymbol{k} = \partial\epsilon_l^{(\mathrm{B})}/\partial\boldsymbol{k}.$$

With this simplification, the continuity condition of wavefunction at $z = 0$ gives

$$f_l^{(\mathrm{A})}(\boldsymbol{r}_{xy}, 0) = f_l^{(\mathrm{B})}(\boldsymbol{r}_{xy}, 0),$$

where $\boldsymbol{r}_{xy}$ is a vector in the xy-plane. For the freedom of $\boldsymbol{r}_{xy}$, the Bloch theorem tells

$$f_l^{(\mathrm{A,B})} = \frac{1}{\sqrt{S}} \exp(i\boldsymbol{k}_{xy} \cdot \boldsymbol{x}) \chi_l^{(\mathrm{A,B})}(z),$$

where $1/\sqrt{S}$ is the partial normalization factor of plane wave in xy-plane, $\chi_l(z)$ is the envelopefunction along z-direction.

For the freedom along z-direction, we consier the $k \cdot p$ perturbation. That is, first we obtain the lattice periodic function and the discrete levels for $k = 0$ and the wavefunctions for $k \neq 0$ are obtained by the hybridization of these wavefunctions caused by the perturbation Hamiltonian, which is proportional to $k \cdot p$. We write down the equation for $\boldsymbol{\chi} = \{\chi_j\}$ as

$$\mathscr{D}^{(0)}\left(z, -i\hbar\frac{\partial}{\partial z}\right)\boldsymbol{\chi} = \epsilon\boldsymbol{\chi}, \tag{6.34}$$

where the $N \times N$ matrix of operators $\mathscr{D}^{(0)}$ is

$$\mathscr{D}_{lm}^{(0)}\left(z, \frac{\partial}{\partial z}\right) = \left[\epsilon_l(z) + \frac{\hbar^2 k_{xy}^2}{2m_0} - \frac{\hbar^2}{2m_0}\frac{\partial^2}{\partial z^2}\right]\delta_{lm} + \frac{\hbar\boldsymbol{k}_{xy}}{m_0} \cdot \langle l|\boldsymbol{p}_{xy}|m\rangle - \frac{i\hbar}{m_0}\langle l|p_z|m\rangle\frac{\partial}{\partial z} \tag{6.35}$$

with

$$\epsilon_l(z) = \epsilon_l^{(\mathrm{A})} \quad (z < 0), \quad \epsilon_l^{(\mathrm{B})} \quad (z \geq 0). \tag{6.36}$$

Here we write $|u_{m0}\rangle$ as $|m\rangle$, etc.

Emphasizing "band-discontinuity potential," we write

$$V_l(z) \equiv \begin{cases} 0 & z < 0 \quad (z \in \mathrm{A}) \\ \epsilon_l^{(\mathrm{B})} - \epsilon_l^{(\mathrm{A})} & z \geq 0 \quad (z \in \mathrm{B}). \end{cases} \tag{6.37}$$

Then we reach the simultaneous equation of $\{\chi_l\}$ as [*1]

$$\sum_{m=1}^{N}\left\{\left[\epsilon_{m0}^{(\mathrm{A})} + V_m(z) + \frac{\hbar^2 k_{xy}^2}{2m_0} - \frac{\hbar^2}{2m_0}\frac{\partial^2}{\partial z^2}\right]\delta_{lm} - \frac{i\hbar}{m_0}\langle l|\hat{p}_z|m\rangle\frac{\partial}{\partial z} + \frac{\hbar\boldsymbol{k}_{xy}}{m_0} \cdot \langle l|\hat{\boldsymbol{p}}_{xy}|m\rangle\right\}\chi_m = \epsilon\chi_l. \tag{6.38}$$

Let us consider the continuity condition of the envelope function χ_l of band l. Because we have assumed that u_l is common for A and B, χ_l should be continuous at the interface. On the other hand, the integration of (6.38) over the interface and the continuity of χ_l leads to the condition

$$\mathscr{A}^{(\mathrm{A})}\chi^{(\mathrm{A})}(z_0 = 0) = \mathscr{A}^{(\mathrm{B})}\chi^{(\mathrm{B})}(0), \tag{6.39}$$

where

$$\mathscr{A}_{lm} = -\frac{\hbar^2}{2m_0}\left[\delta_{lm}\frac{\partial}{\partial z} + \frac{2i}{\hbar}\langle l|p_z|m\rangle\right]. \tag{6.40}$$

It is now clear that the band-hybridizing term $\langle l|p_z|m\rangle$ from the $k \cdot p$ perturbation breaks the simple continuity of derivative of the envelope function.

[*1] If we go up to the second order in k, we have many other terms, which makes the equation very complicated. We thus have omitted them.

6.4.1.2 Joint of envelope function

Next we do not equate u nor the band dispertion (effective mass) but only the single band is considered. The effective mass equation is a second-order differential equation, and the general boundary connection conditions are as follows.

$$\begin{pmatrix} \chi^{(\mathrm{A})}(0) \\ \nabla_{\mathrm{A}}\chi^{(\mathrm{A})}(0) \end{pmatrix} = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix} \begin{pmatrix} \chi^{(\mathrm{B})}(0) \\ \nabla_{\mathrm{B}}\chi^{(\mathrm{B})}(0) \end{pmatrix}, \tag{6.41}$$

where, taking a as the common lattice constant,

$$\nabla_{\mathrm{A,B}} = \frac{m_0}{m_{\mathrm{A,B}}} \frac{\partial}{a\partial z}. \tag{6.42}$$

$T_{\mathrm{BA}} = \{t_{ij}\}$ is called **interface matrix**.

The particle current density along z is determined by the envelope function as

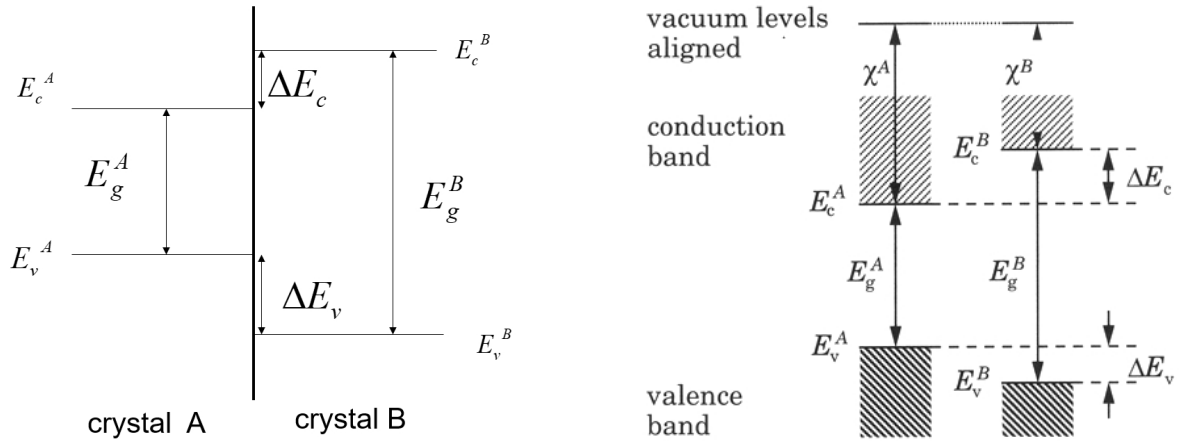
$$j(z) = \frac{\hbar}{2im^*} \left[\chi^*(z) \frac{\partial \chi}{\partial z} - \frac{\partial \chi}{\partial z} \chi(z) \right]. \tag{6.43}$$

From the particle-number conservation, $j(z)$ in A and B regions should be the same. The condition is equivalent to

$$\det T_{\mathrm{BA}} = 1. \tag{6.44}$$

Because this condition is fulfilled when T_{BA} is the unit matrix I, the simplest envelope function approximation is to put $T_{\mathrm{BA}} = I$. In this case, the envelope function can be treated just the same as the real wavefunction. In the case of GaAs-(Al,Ga)As interface, the interface matrix obtained for a one-dimensional tight-binding model indicates the envelope function approximation works well.

In such a case, we can consider the step function potential at the boundary with the height of **band discontinuity** which is determined by the combination of the materials. And the envelope function can be viewed as ordinary quantum wavefunction. On the above basis, we now can use methods to design quantum systems such as one-dimensional potential by thin film growth technique.



(a) (b)

Fig. 6.11 (a) Diagram displaying symbols for the band alignment parameters at a junction of crystals A and B. (b) Anderson model, in which the relative positions of bands are determined by the affinities from the vacuum level.

6.4.2 Anderson's model

Figure 6.11 illustrates a long-used Anderson model [*2] for heterojunctions[3]. In the model, as shown in Fig. 6.11(a), the bands in the bulk continue to the interface. The effect of charge transfer is taken into account as built-in potential just like the treatment of pn homo-junctions.

An important point in this model is the relative band position at the heterojunctions. In the Anderson model, as shown in Fig. 6.11(b), this is determined from the quantity called "affinity", which is the lowering in the energy of electrons with condensation into the crystal state. Then in this model, the connection of the bands is determined by the species of the crystals In the figure, the affinities of A and B are χ^{A} and χ^{B} respectively.

The model, in itself, has many problems, many of which are on the "affinity." Can the affinities be well-defined? Can we calculate them? Are they measurable? We do not have time to go into the problems and furthermore, the experiments have shown that such simple modeling does not work at the level of device designing, in which we need detailed information of band-discontinuity.

We will have a brief look at the junction types and summarize theoretical approachs to the band-discontinuities in Appendix 6C.

6.4.3 Classification of heterojunctions

Semiconductor heterojunctions are classified phenomenologically by the alignment of bands at the interface. Figure 6.12 shows three types of band alignment. (a) is most frequently found and called type-I. On the larger gap side, the conduction bottom is higher and the valence top is lower. In type-II, as shown in (b), the conduction bottom and the valence top shift to the same direction when an electron passes the interface. There is a common energy gap region for A and B in the case of Fig. 6.12 (b). When the missalignment is larger and the energy gap at the interface is closed as in (c), in Japan they call the alignment type-III and in other countries staggered type=II. For example, in ref. [4], the authors call (c) as type-II.

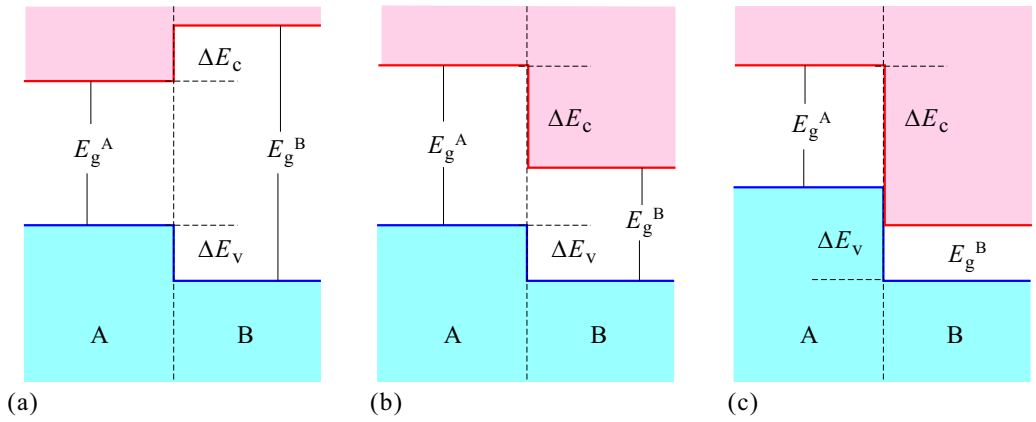


Fig. 6.12 (a) Type-I: To the wider gap semiconductors, the conduction bottom goes up, the valence top goes down. (b) Type-II: The conduction bottom and the valence top moves to the same direction when going through the junction. Also called broken-up or misaligned. (c) Type-III: (Special classification in Japan) Same as type-II but there is no overlapping in the energy gap. Instead there is an overlapp between the valence band on one side and the conduction band on the other side. Also called staggered.

[*2] This "Anderson" is a different person from novel laureate Philip W. Anderson in Bell lab. P. W. Anderson is famous for his "Anderson model" of impurities and R. L. Anderson was in IBM Watson. Bit confusing.

6.5 Formation of heterojunctions

We have already had a look on the epitaxial growth technique. Here I just mention about the lattice matching and energy variation.

6.5.1 Epitaxial growth

Most popular method to form heterojunctions of semiconductors is epitaxial growth already presented in the lecture by Prof. Akiyama. Epitaxial growth methods can be classified into liquid-phase epitaxy, vapour-phase epitaxy, and vacuum deposition. In liquid-phase epitaxy, precipitation onto crystal substrates from melts of ingredients is used. The growths occur in states close to equilibrium and high quality crystals can be obtained while it is hard to obtain sharp interfaces. When one needs sharp interfaces and precise control of layer thicknesses, usually the latter two methods of epitaxy are adopted.

An important point in the formation of heterojunction is the **lattice matching** in lattice constants and crystal systems. In Fig. 6.13, we plot representative compound semiconductors and elemental semiconductors on the plane of lattice constant and energy gap. Most of the plotted semiconductors have a common crystal system, FCC bravais lattice. Vertical gray bands indicate possible groups of lattice matched heterostructure growth though these combinations are not always available in practical growths. Besides these semiconductors, heterojunctions of GaN family are important for industrial demands. They usually have Wurtzite structure (hexagonal close-packed, HCP) and need high temperature treatments, the heterostructures thus are mostly composed within nitride families.

Even with considerable lattice mismatch, a misfit-dislocation free growth to a certain film thickness is possible. An estimation of the thickness given as a balance point of the strain energy concentrated on dislocations and that within whole grown film, is called **Matthews' critical thickness**[5]. Because actual crystal growths are carried out under some non-equilibrium condition, the total free energy not necessarily takes the minimum, the process is generally non-adiabatic. Hence the Matthews' thickness is just a rough estimation. In many cases we need to keep substrate temperatures high

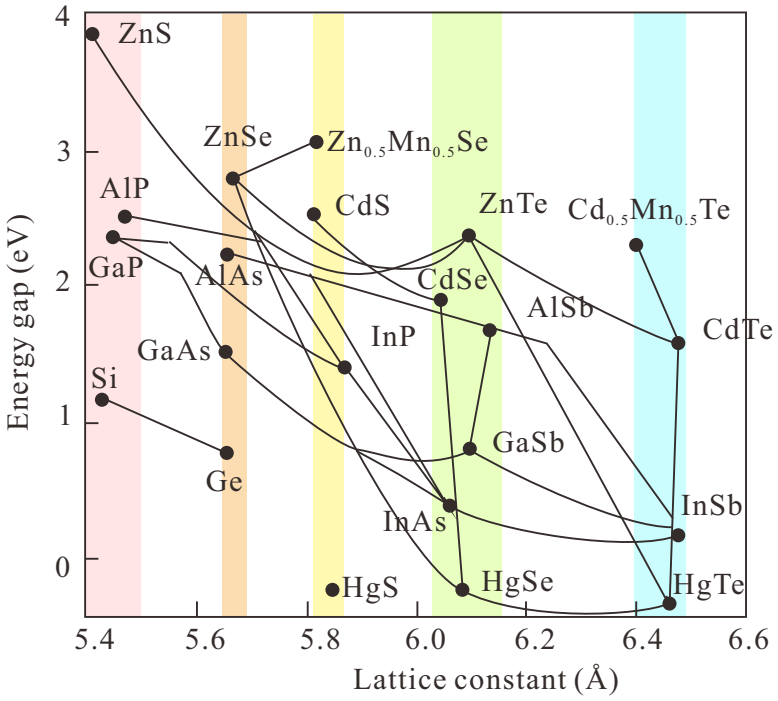


Fig. 6.13 Plots of the lattice constants and the energy gaps of II-VI, III-V compound semiconductors and IV elemental semiconductors. The lines connecting the points indicate possible mixed crystals. Vertical gray bands indicate possible groups of lattice matched heterostructure growth.

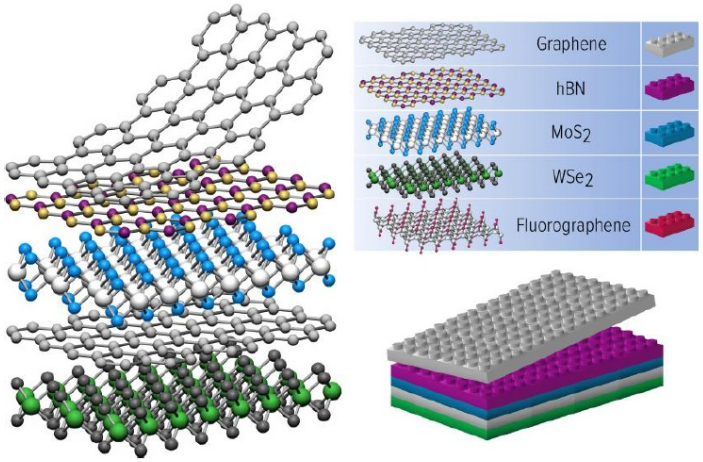


Fig. 6.14 Conceptual illustration of van der Waals heterostructure, which is produced by stacking various two-dimensional materials.

enough during growths and the difference in coefficients of thermal expansion in the two materials sometimes causes dislocations or strains. Many points should be taken into account in actual growths[6].

6.5.2 van der Waals heterostructure

Recently van der Waals heterostructure, which is formed in completely different way, is collecting attentions[7]. That is a mechanical stacking of two-dimensional materials like graphene as shown in Fig. 6.14 (graphene will be introduced later as a two-dimensional electron system without heterointerface). Sometimes epitaxial growth like CVD is adopted but in many cases mechanical stacking of exfoliated two dimensional materials creates high-quality heterostructure, which implies possible completely new formation method of heterostructure.

Chapter 7 Quantum structures (quantum wells, wires, dots)

So far, we have looked at low-dimensional systems such as graphene, which is a two-dimensional substance. These were, so to speak, natural low-dimensional systems. By using heterojunction, metal joining, and microfabrication technology, human hands have come to enter the design of material systems to a certain extent. In this chapter, we look at three typical quantum structures, two-dimensional, one-dimensional, and zero-dimensional systems.

7.1 Quantum well

A region with lower potential sandwiched with two heterojunctions to higher potential materials is **quantum well**. The readers should be familiar with it since introduction of elementary quantum mechanics. In other words, however, the semiconductor heterojunction technology has made the quantum well as a real substance from just an exercise for students.

7.1.1 Discrete quantum levels in a quantum well

Let the well width be L, the barrier height V_0. In $x \leq -L/2$, $L/2 \leq x$ (outside the well) Schrödinger equation is

$$\left[-\frac{\hbar^2 \mathrm{d}^2}{2m\mathrm{d}x^2} + V_0\right]\psi = E\psi. \tag{7.1}$$

Let us put $\kappa \equiv \sqrt{2m|E-V_0|}/\hbar$ and let $C_{1,2}$, $D_{1,2}$ be constants specific to the regions, the solution outside the well can be written as

$$\psi(x) = \begin{cases} C_1\exp(i\kappa x) + C_2\exp(-i\kappa x) & E \geq V_0, \\ D_1\exp(\kappa x) + D_2\exp(-\kappa x) & E < V_0. \end{cases} \tag{7.2}$$

In the case of $E < V_0$, the wavefunction should be localized around the well and zero for $x \to \pm\infty$, then

$$L/2 < x \text{ で } D_1^+ = 0, \qquad x < -L/2 \text{ で } D_2^- = 0.$$

Superscript $\pm$ distinguish the regions positive/negative of x. Inside the well, letting C_1, C_2 be constants, we write the wavefunction with plane waves as

$$\psi = C_1\exp(ikx) + C_2\exp(-ikx), \quad k \equiv \frac{\sqrt{2mE}}{\hbar}, \tag{7.3}$$

where for simplicity, we assume the effective mass m is common for inside and outside the well. The boundary condition at $x = \pm L/2$ where the potential is discontinuous is now applied. Continuity and differentiability at the potential boundary $x = 0$ require

$$\text{Continuity} \begin{cases} C_1\exp(ikL/2) + C_2\exp(-ikL/2) = D_2^+\exp(-\kappa L/2), \\ C_1\exp(-ikL/2) + C_2\exp(ikL/2) = D_1^-\exp(-\kappa L/2), \end{cases}$$

$$\text{Differentiability} \begin{cases} ikC_1\exp(ikL/2) - ikC_2\exp(-ikL/2) = -\kappa D_2^+\exp(-\kappa L/2), \\ ikC_1\exp(-ikL/2) - ikC_2\exp(ikL/2) = \kappa D_1^-\exp(-\kappa L/2), \end{cases}$$

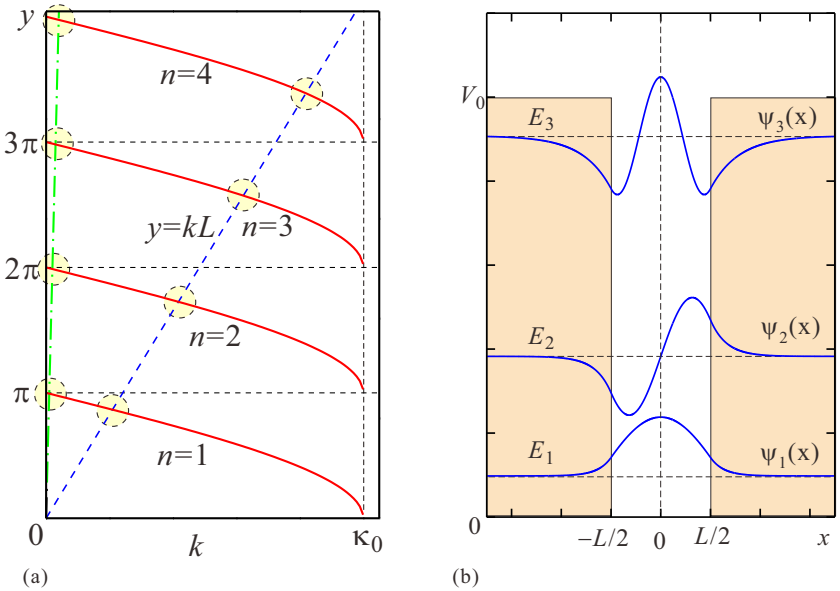


Fig. 7.1 (a) A plot for graphical solutions of k which satisfy eq.(7.4). The crossing points of the functions $-2\arctan(k/\kappa) + n\pi$ and kL give the solutions of (7.4). (b) Bound eigenstates for $n = 1, 2, 3$ under the condition $l = 8$. The baselines for the wavefunctions are the eigenenergies $E_{1,2,3}$ measured with V_0 (for $l = 8$ there are only three bound state solutions, which is different from the situation in the left figure).

respectively. Erasing the constants the following condition is obtained.

$$\exp(2ikL) = \left(\frac{\kappa - ik}{\kappa + ik}\right)^2 = \exp\left(-4i\arctan\frac{k}{\kappa}\right),$$
$$\therefore kL = -2\arctan\frac{k}{\sqrt{\kappa_0^2 - k^2}} + n\pi, \qquad \kappa_0^2 \equiv \frac{2mV_0}{\hbar^2}, \quad n = 1, 2, \cdots . \tag{7.4}$$

Let us take kL as a positive value without loosing generality because the solutions contain $-k$ equivalently, and we restrict the value of $\arctan(x)$ between 0 and $\pi/2$. As shown in Fig. 7.1(a), the crossing points of the curves and the line, $-2\arctan(k/\sqrt{\kappa_0^2 - k^2}) + n\pi$ and kL give the values of k, which satisfy (7.4). As easily guessed from the analogy with the case of infinite barriers, even numbers of n correspond to odd parity wavefunctions, while odd numbers correspond to even parities.

In Fig. 7.1(b), we show the form of wavefunctions for the bound states in the case of $l = 8$.

7.1.2 Optical absorption in quantum wells

We would like to have a short look at optical absorption in quantum wells. As usual we take z-axis vertical to the well plane. We write the envelope functions for electrons and holes as $\phi_e(z)$ and $\phi_h(z)$ respectively and then approximate the total wavefunction as

$$\left.\begin{aligned}\psi_e(\boldsymbol{r}) &= \phi_e(z)\exp(i\boldsymbol{k}_{xy}\cdot\boldsymbol{r}_{xy})u_c(\boldsymbol{r}),\\ \psi_h(\boldsymbol{r}) &= \phi_h(z)\exp(i\boldsymbol{k}_{xy}\cdot\boldsymbol{r}_{xy})u_v(\boldsymbol{r}).\end{aligned}\right\} \tag{7.5}$$

u_c, u_v are lattice periodic parts of the Bloch eigenfunction with $\boldsymbol{k} = 0$. Direct type inter-band optical absorption probabilities are proportional to

$$\langle u_c(\boldsymbol{r})|\boldsymbol{\nabla}|u_v(\boldsymbol{r})\rangle \int_{-\infty}^{\infty} dz\phi_e(z)^*\phi_h(z). \tag{7.6}$$

In the case of infinite height barriers, the envelope functions are written as $\sin(n\pi z/L)$, $\cos(l\pi z/L)$ ($n = 2, 4, \cdots$, $l = 1, 3, \cdots$) and the latter integration over z in (7.6) is finite only between electron envelope function and hole envelope

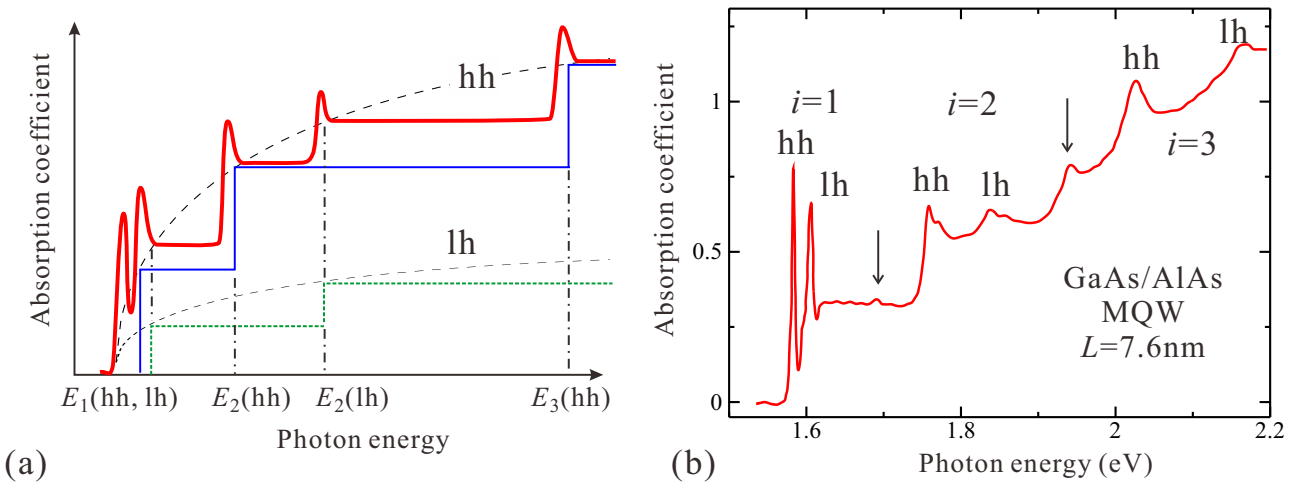


Fig. 7.2 (a) Illustration of theoretically proposed optical absorption spectrum, in which both the coupling density of states and the exciton density of states in the quantum well are taken into account. The approximation that the transition exists only between electrons and holes with the same quantum index. In the valence band of fcc semiconductors we have heavy and ligh holes and transitions with the two bands are considered in the figure. (b) Optical absorption spectrum of a AlAs/GaAs multiple (40 layers) quantum well with width 7.6 nm. The finite barrier height causes transitions between the levels with different quantum indices, which appear in exciton peaks.

function with the same quantum index (n or l in this case). For finite heights, this orthogonality breaks leaving parity selection rule but still elements between different quantum indices are small and we only consider the transition between the states with the same index. The energy associated with the transition is

$$E = E_{\rm g} + \Delta E_n^{(eh)} + \frac{\hbar^2}{2\mu} k_{xy}^2, \tag{7.7}$$

where $\Delta E_n^{(eh)}$ is the sum of the energies for electron and hole in n-th energy levels, $1/\mu = 1/m_e^* + 1/m_h^*$ is the reduced mass. The last term for two-dimensional kinetic energy indicates that there should be continuous absorption spectrum above $\Delta E_n^{(eh)}$ corresponding to the two-dimensional density of states.

From $E = (\hbar^2/2m^*)k^2$ and $n = \pi k^2/(2\pi)^2 = (E/4\pi)(2m^*/\hbar^2)$, the two-dimensional density of states can be written as

$$\frac{dn}{dE} = \frac{m^*}{2\pi\hbar^2} H(E) \quad (H(x) : \text{Heaviside function}). \tag{7.8}$$

This is constant for energy and with (7.7), we expect a staircase like optical absorption spectrum.

Formaion of excitons appears in optical absorption as peaks at energies lower than the fundamental absorption edge. Such peaks for excitons in quantum wells are illustrated in Fig. 7.2(a). Only the ground states ($n = 0$) of the excitons are considered. And coupling density of states between electrons and holes with different subband quantum indices is ignored assuming that the barrier is high enough. Figure 7.2(b) shows an experimental result on an AlGaAs/GaAs multiple quantum well with width 7.6 nm. The lineshape of the absorption spectrum can be understood as an overlap of staircase-like shape reflecting the two-dimensional density of states (7.8) and absorption by excitons indicated as hh or lh. Because the barrier height is finite in the experiment, peaks due to the transition between states with different quantum indices are also observed. The effect of low-dimensionality is observable in increases of binding energy of excitons, which results in wider separation of exciton peaks from absorption edges and the peaks persist up to higher temperatures.

Now we can see that the optical absorption spectra can provide experimental determination of band-discontinuities $\Delta E_{\rm c}$, $\Delta E_{\rm v}$. In the combination of GaAs-$Al_xGa_{1-x}As$, reseachers could not separate lh and hh peaks in very early experiments presumably due to low quality of samples. The result once led them to a wrong conclusion of $\Delta E_{\rm c} : \Delta E_{\rm v} = 85 : 15$ because $\Delta E_{\rm v}$ should be too small to accomodate the lh level. After the revised experiments, it was established

that $\Delta E_{\rm c} : \Delta E_{\rm v} = 57 : 43$ is a good empirical law.

Appendix 6C: Trials for simple theory to find band discontinuities

The capacity of computers has increased dramatically, and first-principles calculations that require a lot of computer resources as LAPW, can now be performed relatively easily. Even though, the researchers are still trying to construct a theory to obtain band discontinuity from a small number of experimental parameters using simple physical principles. I will introduce such researches so far, but it has been found that many cannot withstand the subsequent criticisms, experiments, and first-principles calculations.

6C.1 Common anion rule

This "common anion rule" is considered for compound semiconductors[8] which have finite inonicity. The claim is as follows. Because the valence band is mostly composed of p-orbitals of anion, $\Delta E_{\rm v} \approx 0$ for the semiconductors with a common anion. This is a surprizingly rough theory. The prediction is far from experiments and from other models.

6C.2 Pseudo-potential theory

The quantitiy "affinity" can be formally calculated from the first principles. In the era of Anderson's research, $\Delta E_{\rm c}$ are determined from the experiments, and the affinity is obtained from the fitting. Here we see a theory, which is constructed by Frensley and Krömer[9] aiming at finding band discontinuity from bulk parameters.

The calculation goes as follows. First with the self-consistent pseudo-potential method, the relative positions of bulk bands are calculated in the electrostatic potential inside the crystal[10]. Next from the electronegativity of consisting atoms and from the band structure, the electrostatic potential at the interface is calculated and the relative positions of bands are obtained. They claimed the agreement with exeriments[9].

6C.3 LCAO theory

W. A. Harrison applied his linear combination of atomic orbitals (LCAO) theory to the heterostructure in ref.[11, 12]. In Harrison's theory, LCAO forms the bands. Most of bands in semiconductors can be expressed with the combination of the single s-orbital and the three p-orbitals. The valence band top is composed of p-orbitals and expressed as

$$E_{\rm v} = \frac{\epsilon_p^c + \epsilon_p^a}{2} - \left[\left(\frac{\epsilon_p^c - \epsilon_p^a}{2} \right)^2 + V_{xx}^2 \right]^{1/2}, \tag{6C.1}$$

where $\epsilon_p^{c,a}$ are the energies of p-orbitals of cation and anion respectively on their own sites, V_{xx} is the matrix element between neighboring p-orbitals. According to the theory[11], V_{xx} is approximated as

$$V_{xx} = 2.16\hbar^2/md^2, \tag{6C.2}$$

where m is the electron mass and d is the bond length. The number 2.16 is obtained by fitting the results for Si and Ge to those of other band calculation[13].

In this method, the valence band discontinuities for many semiconductor can be easily calculated and often used for the estimation. It is said to give fairly good agreements with experiments in many cases though with many exceptions *e.g.*, in the case of GaAs-AlAs the Harrison theory gives $\Delta E_{\rm v}$=0.04 eV for about 0.5 eV in the experiments[14]. Actually the main difference in ref. (6C.1) from the common anion rule is just ϵ_p^c and still the approximation is rough.

6C.4 Interface dipole theory

This theory had been pur forward by Tersoff and Harrison[15, 16, 17]. First Tersoff criticized Harrison's LCAO theory that the theory is not reallistic in that the charge transfer between the semiconductors is ignored and no dipole exists at the interface[15]. Then they collaborated in constructing LCAO theories with electric dipoles[17].

In the Tersoff's original idea, on the surface of a semiconductor (an insulator), an energy level of "charge neutrality" can be defined in the band gap. The charge neutrality level is given at the point where the contributions of the valence orbitals and those of the conduction orbitals are balanced. In a metal-semiconductor Schottky junction, this charge neutrality level should be matched to the Fermi surface in the metal[*3]. The charge transfer through the interface is over a very short distance, only single lattice constant. The transfer length scale is very different from the space-charge created so as to match $E_{\rm F}$ with the bulk value. When two semiconductors are jointed, there is no charge transfer for matched charge neutrality levels. Otherwise the transfer occurs to match the charge neutrality levels. Hence, if we can calculate the position of charge neutrality level, the band offset can be derived from that.

	E_B	E_F(Au)[a]	E_F(Al)[a]
Si	0.36	0.32	0.40
Ge	0.18	0.07	0.18
AlAs	1.05	0.96	
GaAs	0.70	0.52	0.62
InAs	0.50	0.47	
GaSb	0.07	0.07	
GaP	0.81	0.94	1.17
InP	0.76	0.77	

[a]Reference 18.

(a)

	Experiments	Theory	Difference
AlAs/GaAs	0.19[b]	0.35	0.16
InAs/GaSb	0.51	0.43	−0.08
GaAs/InAs	0.17	0.20	0.03
Si/Ge	0.20	0.18	−0.02
GaAs/Ge	0.53	0.52	−0.01

[a]Reference 1.
[b]However, see text and Refs. 1, 19, and 20.

(b)

Tab. 6C.1 (a) The in-gap state E_B obtained by equating the contributions from the both bands to (6C.3) and the positions of $E_{\rm F}$ measured in the Schottky diodes with Au and Al as the electrodes.

For the calculation of the charge neutrality level (or, metal-induced gap states, MIGS), the contributions from the valence and the conduction to the real-space averaged Green function

$$G(\boldsymbol{R}, E) = \int d^3r \sum_{n\boldsymbol{k}} \frac{\psi^*_{n\boldsymbol{k}}(\boldsymbol{r})\psi_{n\boldsymbol{k}}(\boldsymbol{r}+\boldsymbol{R})}{E - E_{n\boldsymbol{k}}} = \sum_{n\boldsymbol{k}} \frac{e^{i\boldsymbol{k}\cdot\boldsymbol{R}}}{E - E_{n\boldsymbol{k}}} \tag{6C.3}$$

are equalized to give MIGS E_B. E_B obtained with this method and the positions of $E_{\rm F}$ obtained in Schottky diodes with Au and Al electrodes[**?**] are listed in Tab. 6C.1(a). Already in this table, the agreement is not very good. And after the publication, there occured many criticisms including the experiments. In conclusion, the theory is convenient in the discussion of chemical trend but it is hard to say that it can be used for device design.

6C.5 Example of first principles calculation

Wei and Zunger proceeded with the so-called first-principles calculation of the interface, and the low accuracy of the common anion law and even the simple LCAO theory is due to the roughness of the bulk band calculation rather than the effect of the interface dipole[19]. That is, the bulk contribution $\Delta E^{\rm b}_{\rm VBM}$ and the surface contribution $\Delta E^{\rm is}_{\rm VBM}$ to the energy difference $\Delta E_{\rm VBM}$ at the valence band maximum (VBM) are in the relation

$$\Delta E_{\rm VBM} = \Delta E^{\rm b}_{\rm VBM} + \Delta E^{\rm is}_{\rm VBM}. \tag{6C.4}$$

[*3] In the most of real Schottky junctions, there are defect levels with very high densities and the Fermi levels are pinned there. At the heterointerface with small defect densities, the situation is different.

According to their claim, $\Delta E^{\rm is}_{\rm VBM}$ is small and the problem in LCAO theory rather lies in the estimation of $\Delta E^{\rm b}_{\rm VBM}$. That is in Harrison theory, only s and p orbitals are considered but particularly the contribution from the d orbitals of cation is comparatively large and the most of disagreement with experiments can be explained with this (calculation was done by all-electron generalized linear augmented plane wave method[20]).

	Tight-binding[a]				All-electron (Present results)					
Systems	$\Delta E^{b}_{\rm VBM}$	$\Delta E^{\rm IS}_{\rm VMB}$	$\Delta E^{\rm tot}_{\rm VMB}$	$\Delta E^{\rm expt}_{\rm VMB}$	Average (with SO)	Average (no SO)	Using $1s$	Using $2s$	Using $3p_{1/2}$	δ_{pd}
CdTe-HgTe	0.00	0.09	0.09	0.35 ± 0.06[b]	0.37	0.39	0.377	0.388	0.400	0.34
CdTe-ZnTe	−0.07	0.00	−0.07	···	0.13	0.12	0.125	0.122	0.108	0.04
ZnTe-HgTe	0.07	0.09	0.16	···	0.26	0.29	0.277	0.286	0.289	0.30
AlAs-GaAs	0.01	0.15	0.16	0.45 ± 0.05[c]	0.42	0.41	0.41	0.40	···	0.31

Tab. 6C.2 $\Delta E_{\rm v}$ for semiconductors with Te and As as anion. The results of simple LCAO, experiments, and all-electron first principles calculation.

The calculated results are summarized in Tab. 6C.2. Now the LAPW method can be rather easily utilized in the form of convenient packages like HiLAPW or VASP though still consumes large calculation resources and the jobs are heavy). The method is only for periodic systems and in the case of heterointerface, the unit cell is taken large along vertical direction to the interfaceso as to contain two intefaces and the periodic boundary condition is applied. This is, in a sense, calculation of a superlattice band structure and can be used to check the staircase approximation of the heterointerface.

Appendix 6D: Recombination current and ideality factor

In the discussion of current-voltage characteristics of pn junctions in the text, we only considered the diffusion current. In reallistic pn-junctions, various other factors contribut the current. Here we have a brief look at the current caused by carrier recombination in the depletion layer at the junction interfaces.

First we consider direct gap semiconductors, in which the interband recombination rate is much higher than those in indirect ones. Let the interband recombination rate be R_e, this should be proportional to the carrier concentrations n and p. Thus R_e is proportional to the product pn. Let $R_{\rm rc}$ be the coefficient, then

$$R_e = R_{\rm rc} pn. \tag{6D.1}$$

R_e equals to the thermal activation rate $G_{\rm th}$ of the electron-hole pair in the dark and in equilibrium. Then the law of mass action gives

$$R_{\rm rc} = \frac{G_{\rm th}}{pn} = \frac{G_{\rm th}}{n_i^2}. \tag{6D.2}$$

When there is optical activation or minority carrier injection by the external current, the activation rate and the recombination rate are not balanced and the difference is the net recombination rate U. In n-type semiconductors, the variation in the hole concentration is the main factor. If we write $p_n = p_0 + \Delta p$, $n_n \approx N_{\rm D}$, then

$$U = R_e - G_{\rm th} = R_{\rm rc}(pn - n_i^2) \approx R_{\rm rc}\Delta p N_{\rm D} \equiv \frac{\Delta p}{\tau_p}, \tag{6D.3}$$

where we define the minority carrier lifetime as

$$\tau_p = \frac{1}{R_{\rm rc} N_{\rm D}}. \tag{6D.4}$$

Similarly the electron lifetime in p-type semiconductors are written as

$$\tau_n = \frac{1}{R_{\rm rc} N_{\rm A}}. \tag{6D.5}$$

In contrast, in the indirect gap semiconductors like Si or Ge, the carrier recombination is via the localized traps. In that case, the net recombinatin rate is, according to so called Shockley-Read-Hall statistics[21] written as

$$U = \frac{\sigma_n \sigma_p v_{\rm th} N_{\rm t}(pn - n_i^2)}{\sigma_n \left[n + n_i \exp \frac{E_{\rm t} - E_i}{k_{\rm B}T} \right] + \sigma_p \left[p + n_i \exp \frac{E_{\rm i} - E_t}{k_{\rm B}T} \right]}, \tag{6D.6}$$

where $N_{\rm t}$ is the trap density, σ_n, σ_p are the capture cross-sections for electrons and holes respectively, $E_{\rm t}$ is the trap level, E_i is the Femi level of the intrinsic semiconductors. And $v_{\rm the}$ is the thermal velocity of the minority carrier

$$v_{\rm th} = \sqrt{\frac{3k_{\rm B}T}{m^*}}. \tag{6D.7}$$

In eq.(6D.6), U takes the maximum at $E_{\rm t} \approx E_i$. Though actually $E_{\rm t}$ distribute over the band gap, the trap levels close to E_i contribute lartgely to U [*4], then as a coarse approximation, we consider only single species of traps and put $E_{\rm t} = E_i$ then

$$U = \frac{\sigma_n \sigma_p v_{\rm th} N_{\rm t}(pn - n_i^2)}{\sigma_n (n + n_i) + \sigma_p (p + n_i)}. \tag{6D.8}$$

Just like in the case of interband transition, we write U as $\Delta p/\tau_p$ or $\Delta n/\tau_n$, giving

$$\tau_p = \frac{1}{\sigma_p v_{\rm th} N_{\rm t}}, \quad \tau_n = \frac{1}{\sigma_n v_{\rm th} N_{\rm t}}. \tag{6D.9}$$

Now we use quasi-Fermi levels introduced in eq.(6.4) and from eq.(3.13) the np product is written as

$$np = n_i^2 \exp \frac{\mu_e - \mu_h}{k_{\rm B}T}. \tag{6D.10}$$

Substitutin the above into (6D.6), we obtain

$$U = \frac{\sigma_n \sigma_p v_{\rm th} N_{\rm t} n_i^2 \left[\exp \frac{eV}{k_{\rm B}T} - 1 \right]}{\sigma_n \left[n + n_i \exp \frac{E_{\rm t} - E_i}{k_{\rm B}T} \right] + \sigma_p \left[p + n_i \exp \frac{E_{\rm i} - E_t}{k_{\rm B}T} \right]}. \tag{6D.11}$$

Then again we put $E_{\rm t} = E_i$, and for further simplicity, we assume $\sigma_n = \sigma_p = \sigma$ to obtain

$$U = \frac{\sigma v_{\rm th} N_{\rm t} n_i^2 \left[\exp \frac{eV}{k_{\rm B}T} - 1 \right]}{n + p + 2n_i} = \frac{\sigma v_{\rm th} N_{\rm t} n_i^2 \left[\exp \frac{eV}{k_{\rm B}T} - 1 \right]}{n_i \left[\exp \frac{\mu_e - E_i}{k_{\rm B}T} + \exp \frac{E_i - \mu_h}{k_{\rm B}T} + 2 \right]}. \tag{6D.12}$$

Further, when μ_e, mu_h are position dependent, U takes the maximum in the case E_i places in the middle between μ_e and μ_h. Then (6D.12) reduces to

$$U \approx \frac{\sigma v_{\rm th} N_{\rm t} n_i^2 \left[\exp \frac{eV}{k_{\rm B}T} - 1 \right]}{2n_i \left[\exp \frac{eV}{2k_{\rm B}T} + 2 \right]} \approx \frac{1}{2} \sigma v_{\rm th} N_{\rm t} n_i \exp \frac{eV}{2k_{\rm B}T} \quad eV > k_{\rm B}T. \tag{6D.13}$$

Because the above is the maximum the estimation should be a bit large but the current density due to the recombination can be written as

$$j_{\rm rc} = \int_0^{w_d} qU dx \approx \frac{q w_d n_i}{2\tau} \exp \frac{eV}{2k_{\rm B}T}. \tag{6D.14}$$

In eq. (6D.14) in comparison with eq. (6.11), the voltage term in the exponential has an extra factor 1/2. To put it plainly, this is because the energy exchange when recombination occurs in the trap is half that of the case where the

[*4] This means that the lifetime of minority carriers is determined by a deep level, especially in indirect semiconductors. In "solar grade" Si, in conparison with those for LSI (impurity $^{-10} \sim 10^{-11}$), the purity can be a bit lower while the deep level concentrations should be reduced.

current is generated by overcoming the bandgap (6.11). In this way, different processes in which current flows generally have different voltage coefficients. Then in experiments, the forward current is written as

$$J_{\mathrm{F}} \propto \exp \frac{eV}{\eta k_{\mathrm{B}} T} \tag{6D.15}$$

and the factor η (**ideality factor**) is fit to the experiment. When η is close to 1, the diffusion current is dominant and the junction is close to the ideal case. When it is close to 2, the recombination current inside the depletion layer is dominant.

In the laboratories, η sometimes goes over 2 and still takes higher values. In the case of pn-junctions, the interface comes to the middle of depletion layer and there is some interdiffusion of dopants, thus the factor 2 is frequently obtained while it is usually close to 1 in the case of Schottky junctions.

References

[1] 勝本信吾「半導体量子輸送物性」(培風館，2014).

[2] G. Bastard, “Wave Mechanics Applied to Semiconductor Heterostructures” (Editions de Physique, France, 1990).

[3] R. L. Anderson, IBM J. Res. Dev. **4**, 283 (1960); Solid-State Electronics **5**, 341 (1962).

[4] C. Liu *et al.*, Phys. Rev. Lett. **100**, 236601 (2008).

[5] J.E. Matthews, A.E. Blakeslee, J. Crystal Growth **27**, 118 (1974).

[6] H. C. Casey, Jr., M. B. Panish, “Heterostructure Lasers” Part B (Academic Press, 1978).

[7] A. K. Geim and I. V. Grigorieva, Nature **499**, 419 (2013).

[8] A. G. Milnes and D. L. Feucht, “Heterojunctions and Metal Semiconductor Junctions” (Academic Press, 2012).

[9] W. R. Frensley and H. Kroemer, Phys. Rev. B **16**, 2642 (1977).

[10] M. L. Cohen and J. R. Chelikowsky, “Electronic Structure and Optical Properties of Semiconductors” (Springer, 1989).

[11] W. A. Harrison, J. Vac. Sci. Tech. **14**, 1016 (1977).

[12] W. A. Harrison *et al.*, Phys. Rev. B **18**, 4402 (1978).

[13] D. J. Chadi and M. L. Cohen, phys. stat. solidi (b) **68**, 405 (1975).

[14] 竹田 美和, 応用物理 **67**, 1077 (1998).

[15] J. Tersoff, Phys. Rev. B **30**, 4874 (1984).

[16] J. Tersoff, Phys. Rev. Lett. **56**, 2755 (1986).

[17] W. A. Harrison and J. Tersoff, J. Vac. Sci. Tech. B **4**, 1068 (1986).

[18] S. M. Sze and K. K. Ng, “Physics of Semiconductor Devices” 3rd ed. (Wiley, 2008).

[19] S.-H. Wei and A. Zunger, Phys. Rev. Lett. **59**, 144 (1987).

[20] S.-H. Wei, H. Krakauer, and M. Weinert, Phys. Rev. B **32**, 7792 (1985).

[21] W. Schockley and W. T. Read, Phys. Rev. **87**, 835 (1952); R. N. Hall, Phys. Rev. **87**, 387 (1952).

9th June (2021) Shingo Katsumoto, Institute for Solid State Physics, University of Tokyo

We have seen the effect of confinement with heterojunctions. When the barrier width is finite, the transport across the barrier with quantum tunneling. These can be viewed as elementary quantum mechanics though it is an important step that such phenomena can be observed in real systems in the very beginning of semiconductor quantum physics. Actually the heterojunction technique leads to the prosperity of the field and many novel devices have been created.

7.1.3 Excitons in two-dimensional systems

In the previous section, in the absorption spectrum of a quantum well, we observed peak structures around the absorption edges. They are from the excitons explained in Sec. 3.3.2 thought the lowering of the spatial dimension results in some quantitative differences from the excitons in the bulk. We would like to have a brief look at the excitons in both two and three dimensions.

Let us treat it as a problem of a hydrogen atom then we treat Schrödinger equation with a Coulomb-type central force potential $V_{\rm c}(\boldsymbol{r})$,

$$\left(-\frac{\hbar^2}{2m^*}\nabla^2 + V_{\rm c}(\boldsymbol{r})\right)\psi(\boldsymbol{r}) = E\psi(\boldsymbol{r}), \tag{7.9}$$

in lower dimensions. Here m^* is the electron-hole reduced mass. And we need to change the potential form as

$$V_{\rm c}^{\rm 2d}(\boldsymbol{r}) = -\frac{e^2}{4\pi\epsilon\epsilon_0|\boldsymbol{r}|}, \quad V_{\rm c}^{\rm 1d}(r) = -\frac{e^2}{4\pi\epsilon\epsilon_0(|z|+0.3r_0)}, \tag{7.10}$$

particularly for one-dimensional (along z-axis) systems. This is because simple transformation of eq.(7.9) into one-dimension causes anomalous behavior including divergence of binding energy. The potential form in eq.(7.10) is given as an empirical formula which well fits to a practical numerical calculation on confinement into a finite width quantum wire (a cylinder with radius r_0). Below, we rapidly see the solutions, which are nothing but hydrogen atom solutions. Under variable separation hypothesis, the solutions for eq.(7.9) can be written in the forms

$$\psi^{\rm 3d} = \rho^l e^{-\rho/2}R(\rho)Y_{l,m}(\theta,\varphi), \quad \psi^{\rm 2d} = \rho^{|m|}e^{-\rho/2}R(\rho)e^{im\varphi}, \quad \psi^{\rm 1d} = R(\zeta). \tag{7.11}$$

ρ and ζ are dimensionless variables, which correspond to radial variable and z variable respectively. The definitions are

$$\rho = \alpha r, \quad \zeta = \alpha(|z|+0.3r_0), \quad \alpha = \frac{\sqrt{-8m^*E}}{\hbar}. \tag{7.12}$$

$R(\rho)$, $R(\zeta)$ are the solutions of the following equations.

$$\begin{cases} \left(\rho\dfrac{\partial^2}{\partial\rho^2} + (p+1-\rho)\dfrac{\partial}{\partial\rho} + q\right)R(\rho) = 0 : & \text{3-, 2-dimensional,} \\ \left(\dfrac{\partial^2}{\partial\zeta^2} + \dfrac{\partial}{\partial\zeta} + \dfrac{\lambda}{\zeta}\right)R(\zeta) = 0, \quad \lambda \equiv \dfrac{e^2}{4\pi\epsilon_0\hbar}\sqrt{-\dfrac{m^*}{2E}} : & \text{1-dimensional,} \end{cases} \tag{7.13}$$

where p, q are

$$p = \begin{cases} 2l+1 & \text{(3-dimensional)} \\ 2|m| & \text{(2-dimensional)} \end{cases}, \quad q = \begin{cases} \lambda - l - 1 & \text{(3-dimensional)} \\ \lambda - |m| - 1/2 & \text{(2-dimensional)} \end{cases}, \tag{7.14}$$

where l is angular momentum quantum number and m is magnetic quantum number.

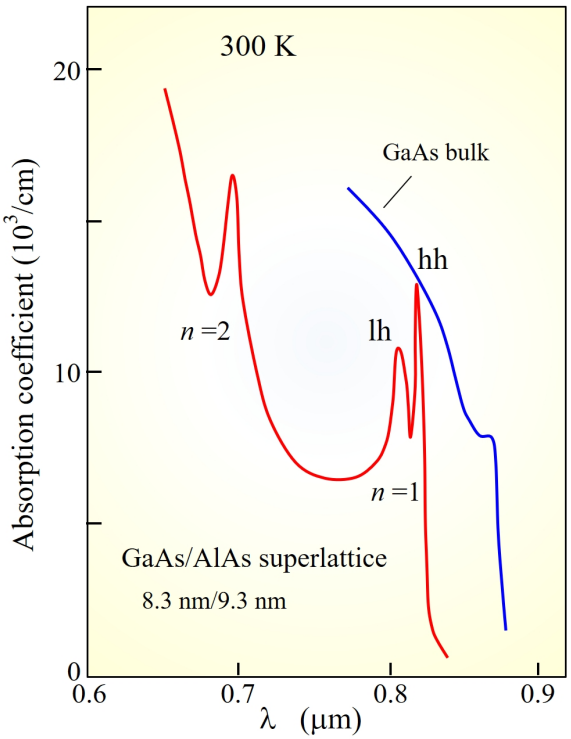


Fig. 7.3 Exciton absorption peaks appeared in the absorption spectrum of a GaAs(8.3 nm)/AlAs(9.3 nm) superlattice at room temperatre (red line). The coupling between the quantum well in the superlattice is weak. n in the figure is the subband index and different from the one of excitonic states index. The peak at the ground subband ($n = 1$) shows the splitting into lh and hh[1]. The $n = 2$ peak is considered as from hh reduced mass.

For three and two dimensional systems, $R(\rho)$ in eq.(7.13) is expanded as follows.

$$R(\rho) = \sum_{\nu} \beta_\nu \rho^\nu, \quad \beta_{\nu+1} = \beta_\nu \frac{\nu - q}{(\nu+1)(\nu+p+1)}. \tag{7.15}$$

For this $R(\rho)$ to be finite, this expansion should stop at a finite number, which condition requires $\nu_{\max} = q$. The main quantum number q then is defined as follows.

$$n \equiv \lambda = \nu_{\max} + l + 1 \ \text{(3-dimensional)}, \quad n \equiv \lambda - \frac{1}{2} = \nu_{\max} + |m| \ \text{(2-dimensional)}. \tag{7.16}$$

The exciton energy levels for three- and two-dimensional systems can be expressed as follows.

$$E_{bn}^{3d} = -\frac{E_0}{n^2} \qquad n = 1, 2, \cdots, \tag{7.17}$$

$$E_{bn}^{2d} = -\frac{E_0}{(n+1/2)^2} \qquad n = 0, 1, \cdots. \tag{7.18}$$

Here the energy unit E_0 is

$$E_0 = \frac{e^2}{8\pi\epsilon\epsilon_0 a_0^*}, \quad a_0^* = \frac{4\pi\epsilon\epsilon_0\hbar^2}{m^* e^2}, \tag{7.19}$$

where a_0^* is the effective Bohr radius. From eq.(7.16), we see that $n = 0$ is available for two-dimensional systems and the ground bound state energy is $-4E_0$. This means the binding energy is four times larger than that in three-dimensional systems where the ground state energy is $-E_0$. In the process of an exciton formation, spatial confinement increases the kinetic energy due to the uncertainty in momentum. In three-dimensional systems, the enhancement occurs for all three dimensions while in two dimensional systems, the confinement along the direction perpendicular to the plane has already been included into the shift of band edge and the binding energy is measured from the edge. Hence it is qualitatively easily understood that the exciton binding energy becomes larger with lowering the system dimension.

Generally radial wavefunction is expressed with Laguerre bi-polynomial and exponential functions. In three dimensional systems, $1s$ wavefunction is written as $\psi_{1s}^{3d} \propto \exp(-r/a_0^*)$. Similarly let $\psi_{1s}^{2d} \propto \exp(-r/a_0^{*2d})$, (7.13) $\curvearrowleft$ $l = m = 0$ and substitution into Schrödinger equation gives $a_0^{*2d} = a_0^*/2$. The spatial size of excitons in two-dimensional systems is half of that in three-dimensional systems in accordance with increment in the binding energy.

In Fig. 7.3, we show the absorption coefficient of a GaAs(8.3 nm)/AlAs(9.3 nm) superlattice (red line) and taht of a high purity bulk GaAs (blue line). In the bulk line, a shoulder structure at the absorption edge is observed. On the other hand, the exciton absorptions take clear peak structures. Furthermore, the peak at the absorption edge of 1st subband ($n = 1$) shows a clear splitting due to two effective masses lh and hh in the valence band, which result in the two reduced masses. The exciton peak at the second subband edge is also clearly observed and considered as from hh reduced mass. These observations are possible by the above enhancement in the binding energy due to the confinement.

7.2 Quantum barrier

"Upside down" of a quantum well potential gives a quantum barrier potential. In the quantum well problem, the focus was on the bound states inside the well while in quantum barriers we see characteristic tunneling phenomena in the upside-down states of **resonant scattering**.

7.2.1 Transfer matrix

Let us consider a region Q in a one-dimensional space and as shown in Fig. 7.5(a), and incoming wavefunction $A(k)$ with wavenumber k from the left hand side (LHS), outgoing wavefunction $A_2(k)$ to the right hand side (RHS), and $B_2(k)$, $B_1(k)$ for the other way around. Here we take the momentum k to be common for the momentum conservation. The suffices 1 and 2 indicates the boundaries 1 and 2.

Let us take for an example that a rectangular barrier with width L, and height V_0. We define $\kappa \equiv \sqrt{2mV_0}/\hbar$. Let the wavefunction inside the barrier be $V_i(\kappa) + W_i(\kappa)$. V, W correspond to $e^{-\kappa x}$, $e^{\kappa x}$ respectively and from the Schrödinger equation, $\partial V_i/\partial x = -\kappa V_i$, $\partial W_i/\partial x = \kappa W_i$. The suffix i indicates positions in real space, just as above, putting 1 and 2 to the left and the right edges of the barrier and

$$V_2 = V_1 e^{-\kappa L}, \qquad W_2 = W_1 e^{\kappa L}.$$

Now the boundary condition can be written as $\partial A_{1,2}/\partial x = ikA_{1,2}$, $\partial B_{1,2}/\partial x = -ikB_{1,2}$, hence,

$$A_1 + B_1 = V_1 + W_1, \qquad A_2 + B_2 = e^{-\kappa L}V_1 + e^{\kappa L}W_1, \tag{7.20}$$

$$ik(A_1 - B_1) = \kappa(-V_1 - W_1), \qquad ik(A_2 - B_2) = \kappa(-e^{-\kappa L}V_1 + e^{\kappa L}W_1). \tag{7.21}$$

For short expression, k, κ for $A \sim V$ are not shown.

First we erase V_1, W_1, then (A_2, B_2) and be expressed with (A_1, B_1). Because of the linearity, the solution can be written in a matrix form as

$$\begin{pmatrix} A_2 \\ B_2 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \equiv M_T \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}. \tag{7.22}$$

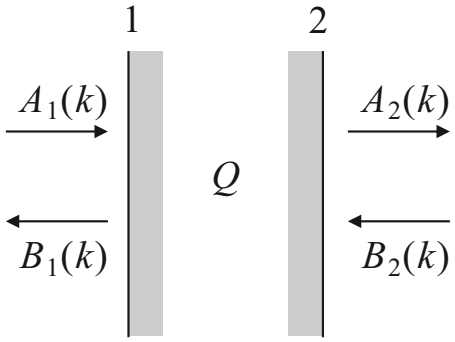


Fig. 7.4 Scheme of T-matrix

Then matrix $\{m_{ij}\}$ is obtained as

$$\begin{cases} m_{11} = \left[\cosh(\kappa L) + i\dfrac{k^2-\kappa^2}{2k\kappa}\sinh(\kappa L)\right], \\ m_{12} = -i\dfrac{k^2+\kappa^2}{2k\kappa}\sinh(\kappa L), \\ m_{21} = m_{12}^*, \quad m_{22} = m_{11}^*. \end{cases} \tag{7.23}$$

Specific form of M_T surely depends on shape of potential though the relation between input and output can always be written in the matrix form as in (7.22) guaranteed by the linearity of Schrödinger equation. A matrix like M_T is called **transfer matrix** (T-matrix).

In Eq.(7.23), M_T has the symmetry of $m_{21} = m_{12}^*$, $m_{22} = m_{11}^*$, which comes from the time-reversal symmetry and the even symmetry in the potential shape.

Let $B_2 = 0$, and the ratio of transmission wave A_2 and reflection wave B_1 to the incident wave A_1 can be given from (7.22), (7.23) as

$$t \equiv \frac{A_2}{A_1} = \frac{|m_{11}|^2 - |m_{12}|^2}{m_{11}^*} = \frac{1}{m_{11}^*} = \frac{2ik\kappa}{(k^2-\kappa^2)\sinh(\kappa L) + 2ik\kappa\cosh(\kappa L)}, \tag{7.24}$$

$$r \equiv \frac{B_1}{A_1} = -\frac{m_{21}}{m_{22}} = \frac{(k^2+\kappa^2)\sinh(\kappa L)}{(k^2-\kappa^2)\sinh(\kappa L) - 2ik\kappa\cosh(\kappa L)}. \tag{7.25}$$

t, r are called **imaginary transmission coefficient** and **imaginary reflection coefficient** respectively. They are related to the transmission and reflection coefficients as

$$\text{Transmission: } T = |t|^2, \quad \text{Reflection: } R = |r|^2, \quad |t|^2 + |r|^2 = 1, \tag{7.26}$$

and the T-matrix M_T can be expressed with them as

$$M_T = \begin{pmatrix} 1/t^* & -r^*/t^* \\ -r/t & 1/t \end{pmatrix}. \tag{7.27}$$

7.2.2 Transmission through double-barrier structure

Let us consider the transmission through the double barrier potential illustrated in Fig. 7.5. Quantum well and quantum barrier are upside down to each other and the double barrier may have the position in between them. Let the boundaries be 1∼4 as in the figure and the wavefunctions also as A_{1-4} and B_{1-4}. For the left barrier the setup is the same as that in the previous section and (7.23) is applcable. Next in the well part between the barriers, a particle gains a kinetic phase factor $\exp(ikW)$ during the traverse. Hence as T-matrix for this part we can adopt

$$M_W = \begin{pmatrix} \exp(ikW) & 0 \\ 0 & \exp(-ikW) \end{pmatrix}. \tag{7.28}$$

The right barrier is just the same as the left. The expression of T-matrix does not depend on local coordinates and M_T can be used as it it.

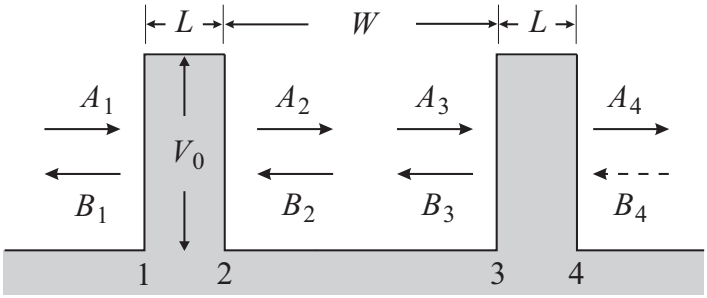


Fig. 7.5 Schematic illustration of double barrier potential.

Then the total T-matrix M_{DW} of the double barrier structure is, as obvious from the definition, obtained as the product of all T-matrices as

$$M_{DW} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} e^{ikW} & 0 \\ 0 & e^{-ikW} \end{pmatrix} \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \equiv \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}. \tag{7.29}$$

The transmission coefficient is, from (7.29),

$$T_{11} = m_{11}^2 \exp(ikW) + |m_{12}|^2 \exp(-ikW) \quad (\because m_{12} = m_{21}^*).$$

The interference effect due to the double barrier structure appears in the second term. Let the argument of m_{11} be φ, and writing $m_{11} = |m_{11}| \exp(i\varphi)$ we get

$$\begin{aligned} T_{11}T_{11}^* &= ((|m_{11}|^2 e^{2i\varphi} e^{ikW} + |m_{12}|^2 e^{-ikW})(|m_{11}|^2 e^{-2i\varphi} e^{-ikW} + |m_{12}|^2 e^{ikW}) \\ &= (|m_{11}^2 - |m_{12}|^2)^2 + 2|m_{11}|^2 |m_{12}|^2 \left(1 + \cos(2(\varphi + kW))\right) \\ &= 1 + 4|m_{11}|^2 |m_{12}|^2 \cos^2(\varphi + kW). \end{aligned}$$

The the transmission coefficient is obtained as

$$T = \frac{1}{|T_{11}|^2} = \frac{1}{1 + 4|m_{11}|^2 |m_{12}|^2 \cos^2(\varphi + kW)}. \tag{7.30}$$

The final form of transmission coefficient is then in combination obtained with (7.23).

Figure 7.6(a) shows thus calculated transmission coefficient T for various barrier widths L as a function of energy of incoming wave. The relation between the barrier width and the well width is fixed as $W = 2L$. Here L and E are transfomed into dimensionless parameters $l \equiv (\sqrt{2mV_0}/\hbar)L$ and E ほ $\epsilon \equiv E/V_0$ respectively. The points where the transmission coefficient hits 1 are due to **resonant scattering** and the condition is written as

$$\varphi + kW = \left(n - \frac{1}{2}\right)\pi \qquad (n = 1, 2, \cdots), \tag{7.31}$$

from (7.30), where φ is witten from (7.23) as

$$\varphi = \arctan\left[\frac{k^2 - \kappa^2}{2k\kappa} \tanh(\kappa L)\right], \tag{7.32}$$

where we restrict the region to $-\pi/2 < \varphi < \pi/2$. With this, n should take a natural number.

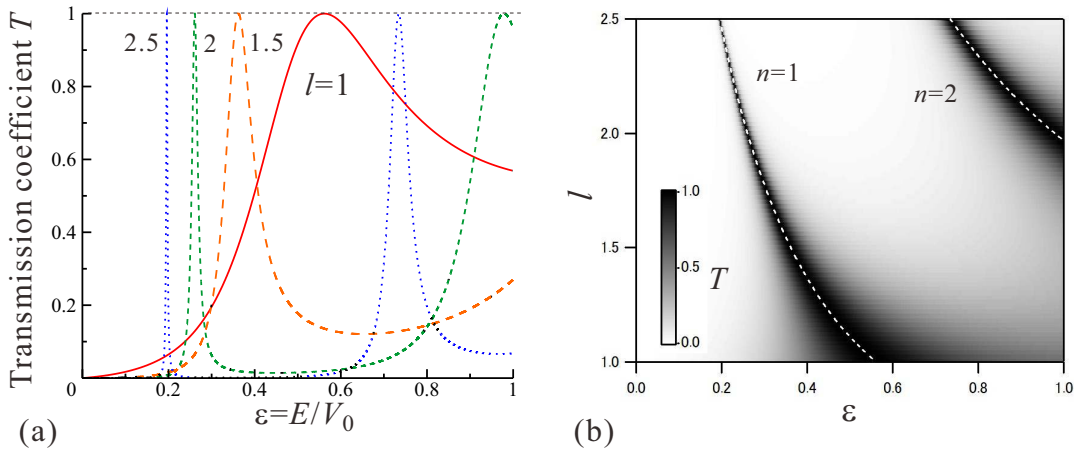


Fig. 7.6 (a) Transmission coefficient T calculated on (7.30) as a function of the energy of incoming wave for various barrier widths. Well width - barrier width relation is fixed to $W = 2L$. (b) The same results are plotted in a gray scale as a function of the incoming energy and the barrier width. White broken lines indicate the resonance condition (7.31), (7.32).

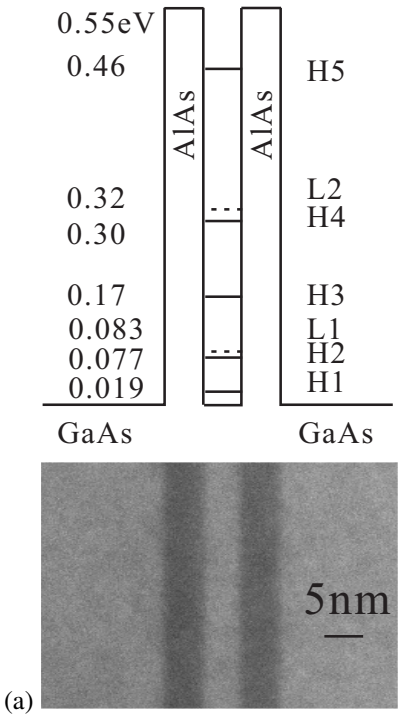
In Fig. 7.6(b), the same data are plotted in a gray scale versus a plane of energy and barrier width. White broken lines indicate the resonant scattering condition in the above equation. With increasing l, the peaks become sharper, which tendency is due to the elongation of time for staying inside the well, that makes the life width determined from the uncertainty relation smaller. If we take the limit $L \to \infty$ keeping W finite, the system becomes a quantum well with a finite barrier height and the resonant scattering condition approaches to that for bound eigenstates.

7.2.3 Transport of double barrier diode

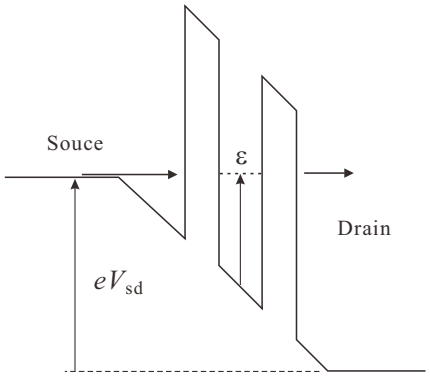
Double barrier diode is a device, which realized the double barrier structure with hetero-inferfaces. Here we introduce an experiment on such a device with GaAs-AlAs hetero-interfaces, p-type doped electrodes. Hence the device works as a double barrier for holes. The band discontinuity is $\Delta E_{\rm v}$ =0.47 eV. There are two species of holes at the top of valence band in GaAs with effective masses $0.51m_0$ and $0.082m_0$, which are called "heavy" and "light" holes (hh and lh) respectively. We ignore the mass difference in AlAs for simplicity (actually the difference is not small but does not affect the result significantly). The potential prepared has, as shown in the upper panel of Fig. 7.7(a), widths of 5nm both for the barriers and the well. The barriers and the well parts do not have any doping. Figure 7.7(a) shows a photograph of the sample cross section taken by a scanning transmission electron microscope, STEM.

The transmission coefficient T thus calculated with the above parameters and the structure shown in Fig. 7.7 is displayed as a function of energy in Fig. 7.8. Because the effective masses of holes are comparatively heavy and the barrier height is high, the transmission peaks are very sharp. We thus can see the behavior of tail only in the semi-log plot. We see below the barrier threshold, 5 heavy hole resonance peaks and 2 light hole peaks. Figure 7.7(a) shows the positions of resonance levels in the well numerically calculated from eq.(7.31).

In order to see the behavior of tunneling, usually source-drain voltage $V_{\rm sd}$ is applied as illustrated in Fig. 7.7(b). Inside the source and the drain, highly concentrated holes screen the electric field and the applied voltage should mainly consumed across the double barrier regions. In actual situation, however, the contact resistances also cause significant voltage drops.



(a)



(b)

Fig. 7.7 (a) Upper panel: Potential diagram of the double barrier diode prepared for the present experiment. The energy base is taken to the top of the valence band and the energy of holes is positive in this plot. $H1 - H5$, $L1, L2$ are the positions of resonant levels for heavy holes and light holes respectively. Lower panel: Scanning transmission electron micrograph of the sample. Darker regions are AlAs. (b) Schematic potential diagram for a source-drain biased double barrier diode.

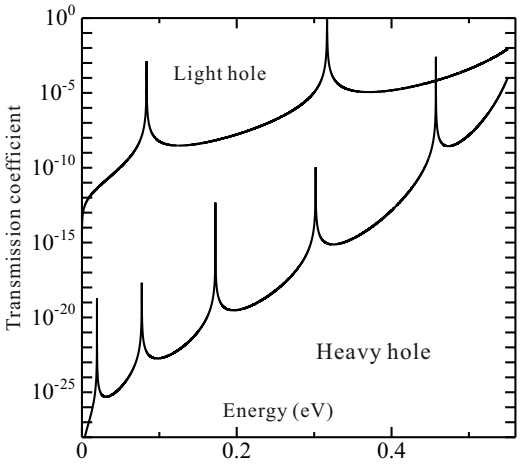
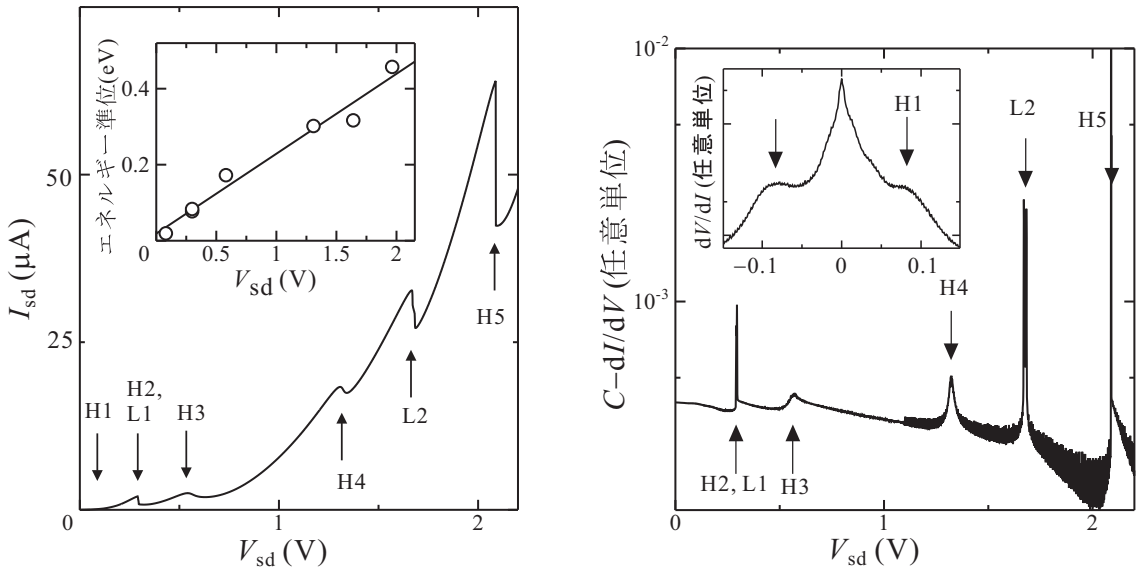


Fig. 7.8 Energy dependence of transmission coefficient for the double barrier structure with parameters given in the text and with eq.(7.30). The peaks hit 1 actually but too narrow to be sampled.

We ignore the distortion of the originally rectangular-shaped potential due to the applied electric field. Then, as in the illustration, the energy of an injected hole is in accordance with the resonant level when the applied voltage reaches twice of it. The transmission coefficient takes a peak at that time, that is the amount of holes passing through the barriers within a unit time, namely the current should take a peak (see Appendix E for more realistic current lineshape).

A measured current-voltage curve in a double barrier diode (the one in Fig. 7.7) is shown in Fig. 7.9(a). Several current peaks appear versus the voltage. To clarify the peak positions the absolute value of voltage-derivative the current with a constant bias C is plotted in a semi-log scale in Fig. 7.9(b) [*1].



(a) (b)

Fig. 7.9 (a) Current-voltage characteristics of the double barrier diode introduced in Fig. 7.7. Resonant levels corresponding to the peaks are indicated by the arrows. The inset indicates peak positions of energy levels on the voltage axis. (b) Emphasis is on the peak positions with differentiating the current with the voltage and the absolute value being plotted in semi-log scale. The inset is enlargement around the origin.

[*1] This transformation is just for the clarity in sight.

7.2.4 Superlattice

The next step, in the course of quantum mechanics, we have double quantum well, which is very important as a qubit. We skip it, to my regret, for the shortage of time. I would like to remind you we have lectures on "nano-quantum information" in the applied physics department (but in Japanese). Here I would like to give a short introduction of **heterojunction superlattice**, which was proposed by Leo Esaki and Raphael Tsu and has provided rich physics. The basic idea of heterojunction superlattice is realization of Kronig-Penny type potential, illustrated in Fig. 7.10. This, in a sense, recovers spatial translational symmetry of the lattice lost by the introduction of the interface but in a different manner.

Let us express a Kronig-Penny type potential as $V_{\rm KP}(x)$ and write down the Schrödinger equation as

$$\left[-\frac{\hbar^2 d^2}{2mdx^2} + V_{\rm KP}(x)\right]\psi(x) = E\psi(x), \qquad V_{\rm KP}(x) = V_{\rm KP}(x+d). \tag{7.33}$$

According to Bloch theorem, we write the eigenstate wavefunction as a product of a plane wave and a lattice periodic function with $d = L + W$ as the lattice constant.

$$\psi_K(x) = u_K(x)e^{iKx}, \quad u_K(x+d) = u_K(x), \quad K \equiv \frac{\pi s}{Nd}. \tag{7.34}$$

s takes an integer from $-N+1$ to $N-1$. The transfer matrix M_d corresponding to the unit cell of the system is

$$M_d(k) = \begin{pmatrix} e^{ikW} & 0 \\ 0 & e^{-ikW} \end{pmatrix}\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} = \begin{pmatrix} m_{11}e^{ikW} & m_{12}e^{ikW} \\ m_{21}e^{-ikW} & m_{22}e^{-ikW} \end{pmatrix}. \tag{7.35}$$

As before, we write the input/output in the left hand side of i-th cell as (a_i, b_i), then from (7.34),

$$\begin{pmatrix} a_{i+1} \\ b_{i+1} \end{pmatrix} = M_d \begin{pmatrix} a_i \\ b_i \end{pmatrix} = e^{iKd}\begin{pmatrix} a_i \\ b_i \end{pmatrix} \tag{7.36}$$

should hold, that is, this is a problem of engenvalue e^{iKd} of matrix M_d. From the unitarity of M_d, or from "reversed" equation of (7.36), the two eigenvalues $e^{\pm iKd}$ are obtained. We re-use $\{m_{ij}\}$ in (7.23) to get to the equations

$$e^{iKd} + e^{-iKd} = 2\cos Kd = {\rm Tr}M_d = 2{\rm Re}(e^{-ikW}m_{11}^*), \tag{7.37}$$

$$\cos\left[K(L+W)\right] = \cosh(\kappa L)\cos(kW) - \frac{k^2-\kappa^2}{2k\kappa}\sinh(\kappa L)\sin(kW). \tag{7.38}$$

By use of φ in (7.32), expression

$$\cos(Kd) = |m_{11}|\cos(kW+\varphi) = \frac{1}{|t|}\cos(kW+\varphi) \tag{7.39}$$

is available.

Transforming the Kronig-Penny potential to a series of δ-function potentials can be attained with taking limits $L \to 0$, $W \to d$, $V_0 \to \infty (V_0 L = C$(constant) to obtain the condition

$$\cos(Kd) = \cos(kd) + \frac{mC}{\hbar^2 k}\sin(kd). \tag{7.40}$$

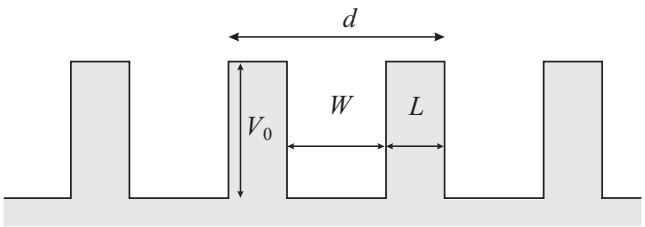


Fig. 7.10 One dimensional rectangular potential (Kronig-Penny type potential)

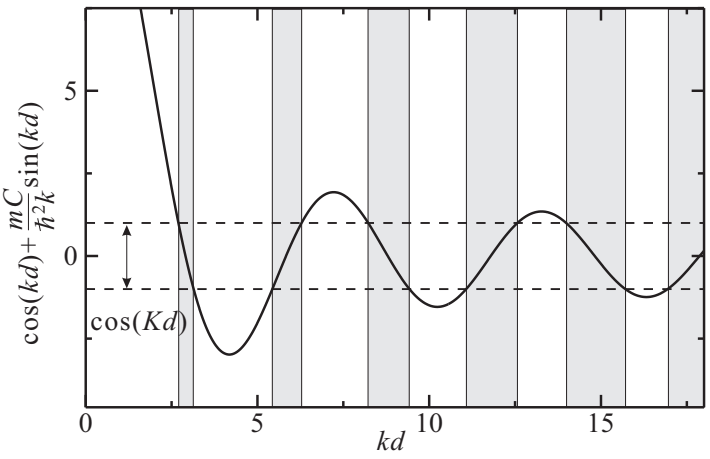


Fig. 7.11 RHS of (7.40) as a function of kd. Here $mdC/\hbar^2$ is taken to be 13. The gray belts indicate "allowed bands".

Figure 7.11 shows the RHS as a function of kd. The solution K for (7.40) exists for the RHS to be in $[-1, +1]$ corresponding to the gray bands namely the energy bands.

Let us simplify the energy dispersion relation of a single band as

$$E(K) = \frac{E_{nw}}{2}(1 - \cos Kd). \tag{7.41}$$

The group velocity and the effective mass are

$$v_g(K) = \frac{E_{nw}d}{2\hbar}\sin Kd, \qquad m^*(K) = \frac{\hbar^2}{E_{nw}d^2}\sec Kd. \tag{7.42}$$

The equation of motion of an electron in a periodic potential under a uniform electric field E_m is written as

$$m^*\frac{dv}{dt} = \hbar\frac{dK}{dt} = F = eE_m. \tag{7.43}$$

We see an effective mass in a periodic potential can be negative.

An acceleration according to (7.43) results in $K = eE_m t/\hbar$. Now we put a wave packet with zero-velociy at the origin $x = 0$, and observe the time evolution. From (7.42),

$$v_g(t) = \frac{E_{nw}d}{2\hbar}\sin\left(\frac{eE_m d}{\hbar}t\right), \qquad x(t) = \frac{E_{nw}}{2eE_m}\left[1 - \cos\left(\frac{eE_m d}{\hbar}t\right)\right]. \tag{7.44}$$

The result indicates that in spite of the constant acceleration, the wave packet oscillates in space. The phenomenon is called **Bloch oscillation**, an observation of which in an actual lattice is almost impossible due to various scattering. In a superlattice, however, the super-period devides the large original band into "mini-bands" and the acceleration to the top of a mini-band before scattering. The Block oscillation was thus observed in superlattices in optical measurements.

7.3 Modulation doping and two-dimensional electrons

The most popuar artificial structure made with heterojunctions is the two-dimensional electrons with modulation doped heterojunctions (two-dimensional electron gas, 2DEG). As is illustrated in Fig. 7.12, in a single heterojunction, doping is given just in the wider band region. Now let us see what happens here for n-type doping.

Let us take the z-axis vertical to the surface and the hetero-interface plane as in the figure. In a "rigid band" model, the conduction band discontinuity ΔE_c emerges and the carriers re-distribute. Let us take the plain case of the combination of $Al_xGa_{1-x}As$ and GaAs. Then we can adopt the approximation that the envelope function in the effective mass approximation as the electron wavefunction itself, and electron-electron interaction can be treated within the Hartree

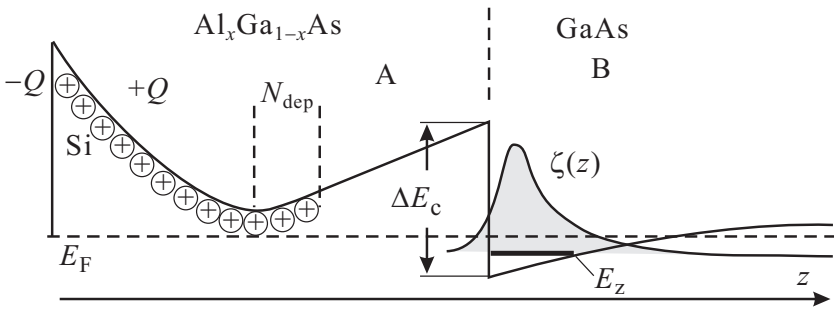


Fig. 7.12 Schematic cross sectional view of two-dimensional electrons at a modulation doped $Al_xGa_{1-x}As$/GaAs heterointerface.

approximation[*2]. Then, the Poission-Schrödinger equation including the electrostatic potential formed by ionized donor, the band discontinuity and the 2DEG itself should be solved self-consistently for obtaining quantilzed energy levels and wavefunction (envelope function) along the direction perpendicular to the 2DEG plane.

z-axis is taken to be perpendicular to the heterointerface plane. As in Fig. 7.12, the surface Schottky barrier creates a surface depletion layer. Let the charge at the surface be $-Q$ and the electric field from the charge should be compensated with that from charges at ionized donors (in the figure Si) $+Q$ in the amount and screened from inside. Let us write the number of all the residual ionized donors per unit area (integrated along z-axis) as $N_{\rm ddep}$. The electrostatic potential from the charges is, far inside the lattice from the doping region, $V_D(z) = (4\pi e^2/\epsilon\epsilon_0)N_{\rm dep}z$. Between the doped region and the hetero-interface, a non-doped region called "spacer " is often places. The spacer spatially separates the 2DEG and the ionized impurities, decreases scattering probabilities of two-dimensional electrons, resulting in very high mobility of electrons. A too thick spacer, however, lifts up the band depletes the well and throws out the 2DEG.

Let us adopt a variable separation type expression for 2DEG wavefuntion, $\Psi(\boldsymbol{r}) = \psi(x,y)\zeta(z)$. $\zeta(z)$ is the envelope function along z-axis. The areal concentration n_{2d} is the function of discretized energy level E_z, which is in other words the kinetic energy along z-axis for $\zeta(z)$. The areal charge density at position z' is then $-en_{2d}|\zeta(z')|^2$, the sheet charge of which creates the electric field $-(4\pi e^2/\epsilon\epsilon_0)n_{2d}|\zeta(z')|^2|z-z'|$ as calculated from the Gauss theorem. In the Hartree-only mean field approximation, the potential should include these terms. The potential created by the 2DEG itself is

$$V_{2d}(z) = -\frac{4\pi e^2}{\epsilon\epsilon_0}n_{2d}(E_z)\int_{-\xi}^{\infty}|\zeta(z')|^2|z-z'|dz'.$$

Here the integral cut-off ξ should be taken longer enough than the penetration depth of $\zeta(z)$ in to AlGaAs barrier. We write a step potential with discontinuity ΔE_c just at the interface as $V_h(z)$. Now the total potential can be written as

$$V(z) = V_h(z) + \frac{4\pi e^2}{\epsilon\epsilon_0}\left[N_{\rm dep}z - n_{2d}(E_z)\int_{-\xi}^{\infty}|z-z'||\zeta(z')|^2dz'\right]. \tag{7.45}$$

Schrödinger equation for $\zeta(z)$

$$\left[-\frac{\hbar^2}{2m^*(z)}\frac{\partial^2}{\partial z^2} + V(z)\right]\zeta(z) = E_z\zeta(z) \tag{7.46}$$

should be solved self-consistently to obtain (consistent) $\zeta(z)$. The effective masses m^* are different in the two species of semiconductors and the boundary condition should be

$$\zeta(0)^{(\rm A)} = \zeta(0)^{(\rm B)}, \quad \left.\frac{1}{m^*_{\rm A}}\frac{d\zeta^{(\rm A)}}{dz}\right| = \left.\frac{1}{m^*_{\rm B}}\frac{d\zeta^{(\rm B)}}{dz}\right|. \tag{7.47}$$

In the Poisson-Schrödinger procedure, one should solve the equations from (7.45) to (7.47) consistently. The above only treats the Hartree term. In general, the Fock term, or the correlation effect is also important in mean field theory. However, it is known that the correlation effect does not affect $\zeta(z)$ or E_z so much and here we ignore it for simplicity.

[*2] Even within the mean field theory, the interaction term contains the Fock term (exchange), but the contribution was calculated to be small.

It is comparatively easy to solve Poisson-Schrödinger equation numerically for a simple band with small spin-orbit interaction, like the conduction band in GaAs. For more complicated cases, *e.g.*, multiple valleys, strong spin-orbit interaction, etc., the scale of numerical calculation increases. If one needs to expand the calculation to other quantities with obtained $\zeta(z)$ for such a case, approximate formulas with simple mathematical forms are convenient. For example, in Fang-Howard approximation, the formula

$$\zeta(z) = \sqrt{\frac{b^2}{2}} z \exp\left(-\frac{bz}{2}\right) \tag{7.48}$$

is used as the trial function with b as a parameter for variational calculation. The result of the variational calculation is given as

$$b^3 = \frac{48\pi m e^2}{\epsilon\epsilon_0 \hbar^2}\left(\frac{11}{32} n_{2d} + N_d\right). \tag{7.49}$$

In this approximation, penetration of wavefunction into the barrier (spacer) is ignored. Another approximation form which takes such penetration into account is given in, *e.g.* ref.[3].

7.4 Fabrication of quantum wires

Nowadays we have so many methods to fabricate quantum wires and reviewing in this narrow space is impossible. Here we have a short look at a few examples of them.

7.4.1 Split gates, other physical approach

The split gate method starts from a 2DEG. Metallic films on the surface form Schottky barriers and deplete the electrons underneath them. Then we can build potentials with various shapes through those of the metals.

In the split gate method, enlargement of depleted regions with reverse (negative) bias voltage V as in Fig. 7.13(a) is often used. Let us consider a simple model illustrated in Fig. 7.13(b), where two half-infinite metals are placed with a distance w. The line density of charge σ, created by applied gate voltage is assumed to be uniform. The electric field formed by these charges has the z-component $\mathcal{E}_z(x, d)$ as,

$$\mathcal{E}_z(x, d) = \frac{-\sigma}{2\pi\epsilon\epsilon_0}\left[\pi + \arctan\left(\frac{x - w/2}{d}\right) - \arctan\left(\frac{x + w/2}{d}\right)\right]. \tag{7.50}$$

(7.50) depends on d, but as a coarse approximation, we ignore the dependence within the depth η of the two-dimensional electron gas (2DEG) potential and the potential modulation due to the split gate is summarized as $V_{\rm sg}(x) = e\eta\mathcal{E}_z(x, \eta)$.

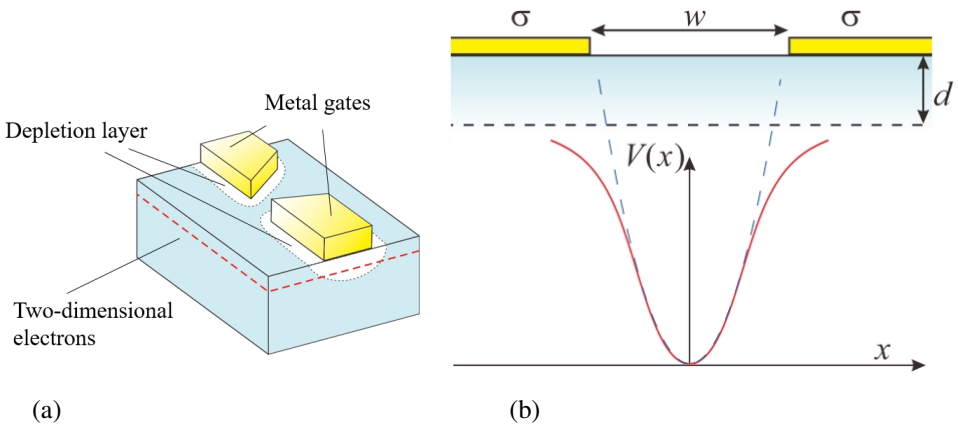
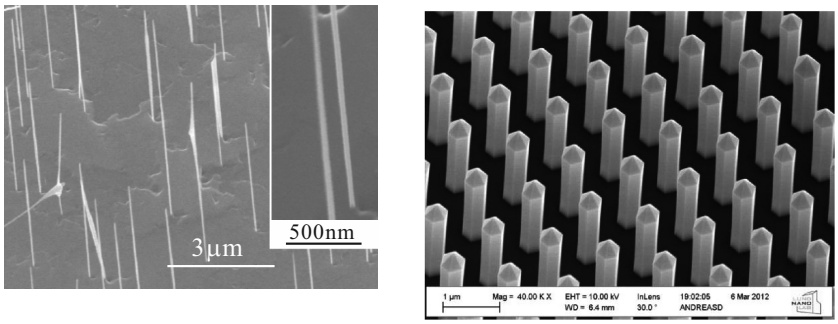


Fig. 7.13 (a) Schematic illustration of micro-fabrication by split-gate method. (b) Electrostatic potential formed by a split gate calculated on a simple model in the text.



(a) (b)

Fig. 7.14 (a) Scanning electron micrograh of InAs nanowires along [111] grown by vapor-liquid-solid phase method with metal particles as catalyst on an InAs(111)B substrate. (b) GaN-InGaN core-shell type nanowires.

An example of $V_{\rm sg}(x)$ is shown in the lower panel in Fig. 7.13(b). As indicated by the broken line, a parabollic potential well approximates $V_{\rm sg}(x)$ around the bottom [*3].

The gate electrodes in Fig. 7.13(b) hence give confinement along x in addition to z-direction heterinterface confinement potential. The kinetic energies for these two directions are quantized and quantum wires are realized. Density of states in one-dimensional systems has a divergence at the bottom of band, which corresponds to the descrete energy levels in the potential in Fig. 7.13(b).

Another way to form quantum wires with physical methods is "cutting" of 2DEG into thin and long shapes with wet (*i.e.* in some solvent) or dry (*i.e.* in some plasmas) etching method. There are also several ways such as regrowth of heterostructures onto a cleaved edge of another heterostructure to form T-shaped thin-line potential, ion-inplantation inactivation, etc.

7.4.2 Self-assembled nanowires

After putting some "seeds" onto semiconductor substrates by electron beam irradiation, etc., crystal growth onto it causes nanowire growth at the seeds under some conditions. In such a growth, with changing "flying" materials heterostructures or doping can be installed in the wire. Figure 7.14(a) shows InAs nanowire along [111] grown on an InAs(111)B substrate with Au nanoparticles as the catalyst by vapor-liquid-solid phase method. Fiugre 7.14(b) shows GaN-InGaN core-shell type nanowires.

Well-known **carbon nanotubes**, which are rolled up graphenes with nanometer-size diameters, may not be classified into one-dimensional systems, are also a kind of "self-assembled systems".

7.5 Fabrication of quantum dots

Fabrication methods of quantum dots are also classified into physical methods, self-assemble methods, and their combinations. I would like to introduce just a part of them.

7.5.1 Physical method

Quantum dots (QDs) are expected to have a very wide range of optical applications, and quantum dot lasers are already on the market. Here, however, we restrict ourselves to the QD for the study of transport. To measure the electric conduction, we need to touch electrodes to the dots. As the electrode material we consider a normal metal with an ordinary Fermi surface. And as the "connection", we consider tunneling junctions, through which electron can transmit

[*3] Anyway a rounded bottom of symmetric potential can often be approximated by a parabolla because the leading term in the power expansion is usually of the second order.

with quantum tunneling. At least a single electorode is required. And to compose structure for measurement of transport between two particle reservors just like quantum wires, FET, two electrodes, hence two tunneing junctions are required. As shown in Fig. 7.15, these electrodes are called just like FET, source and drain.

In quantum dots, the density of states is like a series of δ-functions. The electric conduction is determined by the tunneling probability of junctions and the relative positions of the δ-function like density of states and the Fermi levels in the drain and the source. If we place another metallic electrode close to the dot without tunneling probability, the electrochemical potential of the dot can be controlled with the electric field from the electrode. The electrode is called gate. Figure 7.15(a) shows the total schematic of the quantum dot for conduction measurement.

The split-gate method introduced for the quantum wire can also be applied to form quantum dots. Figure 7.15(b) illustrates a possible pattern of Schottky electrodes. For the tunnel junctions, quantum point contacts near the pinch-off condition are used. For the gates, also Schottky electrodes are used in the reverse bias condition. As can be imagined from the figure, because the reverse bias voltage enlarges the depletion layer to squeeze the dot region, the size of the quantum confinement potential is diminished that widens the level intervals other than the single electron effect, which will be discussed later. Hence the gate is sometimes called "plunger" gate. In this kind of configuration, the source, dot, drain line up side by side along the two-dimensional electrons, hence called "lateral" quantum dots. When the number of electrons in a quantum dot is reduced by the gate voltage, the dot size also becomes smaller, it is spatially separated from the source and drain, the tunnel probability becomes smaller, and conduction becomes unmeasurable. This once considered as a difficult problem but has been overcome by the remote charge detection. The structure in Fig. 7.15(c) is made from the double barrier structure The structure shown in Fig. 7.15(c) is made by cutting out a double barrier structure into a cylinder shape. Then Schottky gate electrodes are deposited on the side of the cylinder. This is called a "vertical" quantum dot. The tunnel couplings are determined by the double barrier structure, not affected by the electron number. The property makes the structure apropriate for the experiments for small number of electrons. There is a problem in connection with external quantum circuits, which requires various devicing.

7.5.2 Self assembling method

The epitaxial growth has various "mode" in the growth process. The layer-by-layer growth mode is called Frank-van der Merve (FvdM) mode (Fig. 7.16(a)). When the interface energy accumulation between the thin film and the substrate is large due to the combination of materials, the deposited material is repelled from the substrate in the beginning of the growth and a three dimensional growth begins. As a such growth mode, the Volmer-Weber (VM) mode is illustrated in Fig. 7.16(b). In Fig. 7.16(c) we show the Stranski-Krastanow (SK) mode, in which the growth is two-dimensional at the very beginning but changes into three-dimensional due to the lattice distortion inside the film. With such three

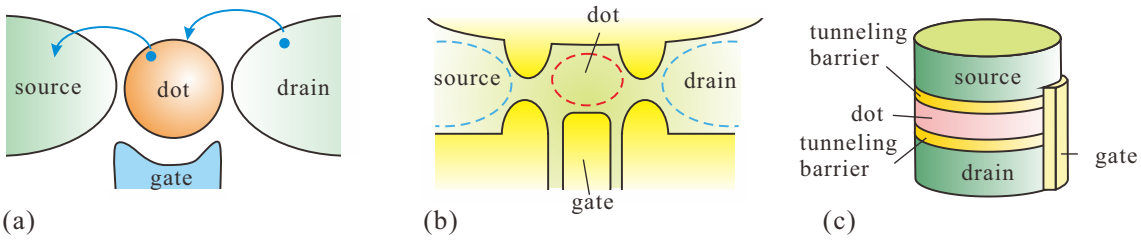


Fig. 7.15 (a) Schematic of quantum dot structures for transport measurement. Two electrodes for examining conduction, a source, and a drain are connected via a tunnel barrier across the quantum dot, and a gate electrode that controls the potential of the dot is arranged at a distance. (b) Illustration of a "lateral" quantum dot. Nano-fabricated metallic gates on two-dimensional electrons are used. (c) Illustration of a "vertical" quantum dot. The dot layer is between two barrier layers and the doped upper and lower layers are the source and drain. The layers are cut to a piller and the metallic gate is deposited surrounding it (in the figure a part of the gate is drawn).

dimensional growth, the structures with dimensions less than two can be obtained.

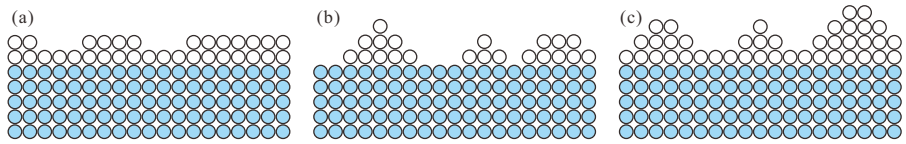


Fig. 7.16 Various epitaxial growth modes. Blue-gray circles represent atoms in the substrate, while open circles do the atoms in the fim. (a) Frank-van der Merve. (b) Volmer-Weber. (c) Stranski-Krastanow.

A representative example of self-assembling of low-dimensional systems is the growth of quantum dots with SK mode growth. Such examples are shown in Fig. 7.17. When some amount of a semiconductor InAs, which has the lattice constant 7% larger than that of GaAs, is deposited on a GaAs substrate and kept at some high substrate temperature for some time, the indium atoms on the substrate migrate and accumulate to form quantum dot structures.

In the case of InAs, dots are self-assembled because when a relatively small number of In atoms are present on the substrate, it is more energetically stable to perform three-dimensional growth to escape from the interface, which gets strong lattice distortion from the substrate. The indium atoms first form a two-dimensional wetting layer with thickness of a few lattice constant then dots are randomly formed in the shape that depends on the crystal direction of the surface. The quantum dots produced by the SK mode growth have random sizes and positions. On the other hand, the SK dots are with high crystal qualities and with high densities, and thus widely used for optical devices like quantum dot lasers. In addition, since InAs has low junction resistances with metals, conduction measurement is also performed by attaching metal electrodes. A method often adopted as a combination of self-assembling and physical methods is to attach gates and barrier electrodes to self-forming nanowires by lithography to make dots. In particular, the self-assembling of InSb or InAs, to which heterojunction technique is difficult (though not impossible) to apply, is used to form quantum dots and other structures with many gate electrodes. There is also a method to form quantum dots with implementing the barrier layers into nanowire during the growths.

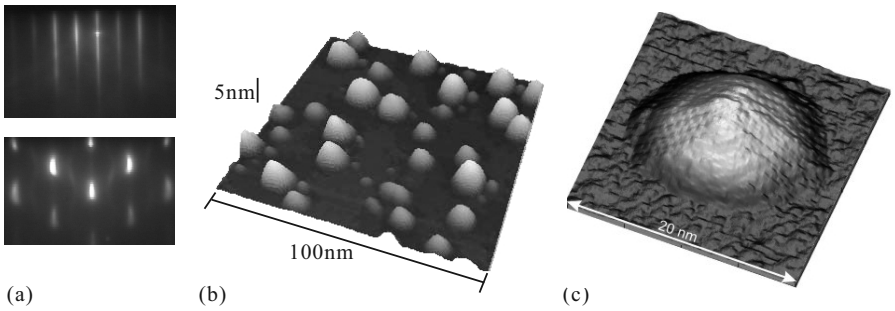


Fig. 7.17 (a) InAs quantum dot growth on GaAs (001) substrate. Upper panel: RHEED pattern of two-dimensional growth at the beginning of InAs growth. Lower penel: RHEED pattern of three-dimensional growth of quantum dot structure. (b) Atomic force micrograph of self-assembled quantum dots. (c) Scanning tunneling micrograph of a quantum dot. The lattice image can be seen, manifesting that the whole dot is a single crystal.

7.5.3 Colloidal quantum dots

In recent years, the colloidal manufacturing method has come to be widely used to form optical quantum dots for optical use. As shown in Fig. 7.18, this is a method of obtaining quantum dots by injecting a dot material called a ”precursor” into a solvent, dissolving it, making it supersaturated by a temperature change, and precipitating a part of it. From the relationship between surface area and volume, when the degree of supersaturation falls below a certain value, dots that continue to grow and dots that redissolve are separated, so dots with relatively uniform sizes can be obtained. This is

called Ostwald reipening. After the growth reached saturation, surface covering of the grown dots with another material is possible by adding a different precursor into the solvent. Such covered quantum dots obtained in this way are called **core-shell** type quantum dots. The luminescence wavelength of the quantum dots can be tuned by their size and hence it is possible to form high efficiency luminous materials. They are already applied to, *e.g.* quantum dot displays.

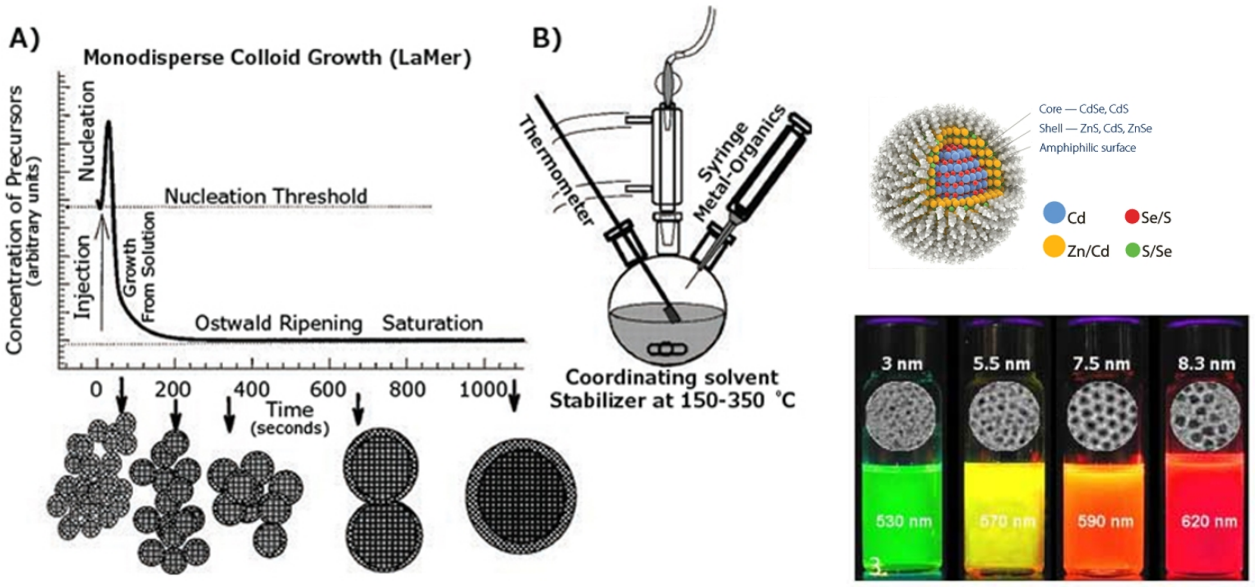


Fig. 7.18 Left (A): Illustration of quantum dot formation with precipitation from supersaturated solvent. The time dependence of the density of precursor is plotted in the graph. Left (B): Illustration of experimental setup of precipitation method. Right upper: Schematics of core-shell type quantum dots. [4] Right lower: Illumination from CdS-based core-shell quantum dots and TEM images. (Ocean Nanotech. web site. https://www.oceannanotech.com/)

References

[1] T. Ishibashi, S. Tarucha and H. Okamoto, Proc. GaAs and Related Compounds, p. 587 (Oiso, 1981).

[2] L. Esaki and R. Tsu, IBM J. Res. Dev. **14**, 61 (1970).

[3] T. Ando, J. Phys. Soc. Jpn. **51**, 3893 (1982).

[4] C. B. Murray, C. R. Kagan, and M. G. Bawendi, Annu. Rev. Mater. Sci. **30**, 545 (2000).

16th June (2021) Shingo Katsumoto, Institute for Solid State Physics, University of Tokyo

I would like to introcude some of optical devices with hetrojunctions though the confinement scales are not in quantum regime.

7.6 Confinement of injected minority carriers and optical devices

The minority carriers injected with pn junctions or with optical excitations, are transported by diffusion currents or drift currents in solids. Spatial geometries, injection currents etc. are used for the control of diffusion currents. A typical example is the bipolar transistor. Drift currents can be controlled through internal potentials introduced by hetero, Schottky, MOS junctions, and through bias voltages, and gate voltages. A simple example is the window layers of solar cells. As illustrated in Fig. 7.19(a), a window layer is placed on the top layer of a pn-junction solar cell. It should have a larger band gap than that of the material for the pn-junction.

One of the factors of lowering the conversion effciency of solar cells, is the non-radiative recombination of injected minority carriers via the highly dense surface states, which also cause the pinning of the Fermi level in Schottky junctions. The current through the device is driven by minority carriers swept out by the built-in potential of the pn-junction. Minority carriers created inside the semiconductor have a random initial momentum and diffuse also to the surface. Many of them are lost at the surface with non-radiative recombination and thier energies either as heat. When the surface has some decolation to prevent reflection, the increase of the surface area reults in the enhancement of surface recombinatn rate.

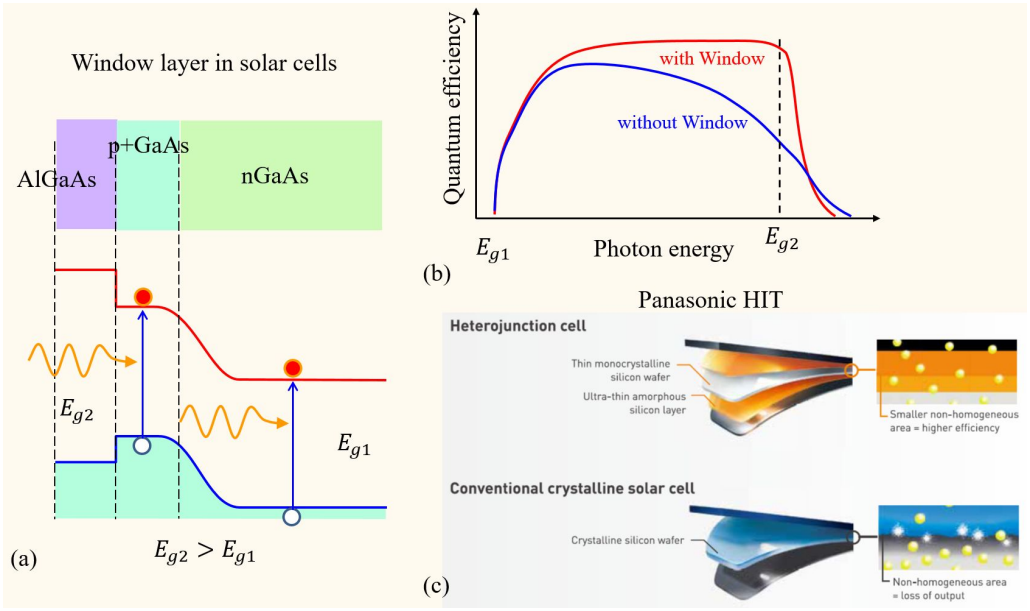


Fig. 7.19 (a) Upper: Illustration of a solar cell with a window layer. An example of AlGaAs/GaAs pn junction. Lower: Schematic band-diagram of the solar cell shown in the upper panel. (b) The schematic diagrams of quantum collection efficiencies of cells with (red line) and without (blue line) the window layer. (c) Illustrations from brochure of HIT solar cells, which demonstrates the enhancement in the conversion efficiency with the heterojunctions.

If the cell has the layer with a larger bandgap E_{g2} on the top as illustrated in Fig. 7.19(a), the diffusion of minority carriers to the surface is blocked with the heterojunction barrier. The materials are chosen so as to have a good connection at the junction and not to have in-gap recombination centers as AlGaAs-GaAs in the figure. Then the minority carriers reflected at the junction diffuse back to the pn junction and contribute to the photocurrent. Figure 7.19(b) illustrates the quantum efficiency spectra $\eta(h\nu)$ of pn junctions (energy gap E_{g1}) with and without the window layer with the energy gap E_{g2}. $h\nu$ is the energy of photons. ν is the photon frequency throuout this section. The quantum efficiency is defined as the ratio of the number of electrons in the photocurrent to that of the photons in incoming flux. Without the window layer (blue line), $\eta(h\nu)$ decreases with increasing the photon energy due to the increase of minority carrier creation close to the surface and hence the increase of surface recombination. With the window layer (red line), the surface recombination is reduced and the value of η is kept close to 1 up to around $h\nu \sim E_{g2}$. Above E_{g2}, due to the absorption in the window layer, the surface recombination increases and the efficiency decreases. If we can choose E_{g2} around the energy above which the mode density of sun light is small, we can expect large enhancement in the conversion efficiency. The window layer can be viewed as an example of reducing the diffusion current with a kind of drift current caused by barriers at heterojunctions.

An ingenious example of the application of above technique to market-selling devices is the solar cells named HIT (heterojunction with intrinsic thin-layer), which were developed in SANYO and now are produced and sold in Panasonic bland. The base is a crystal Si solar cell but they utilized the fact that amorphous Si has a larger effective band gap. In the structure a Si active layer is sandwitched by clad amorphus layers, which cause confinement of minority carriers inside the crystal Si. HIT still has a top-class conversion effciency but unfortunately, it has been announced that it will be discontinued due to various reasons.

7.6.1 Light emitting diodes

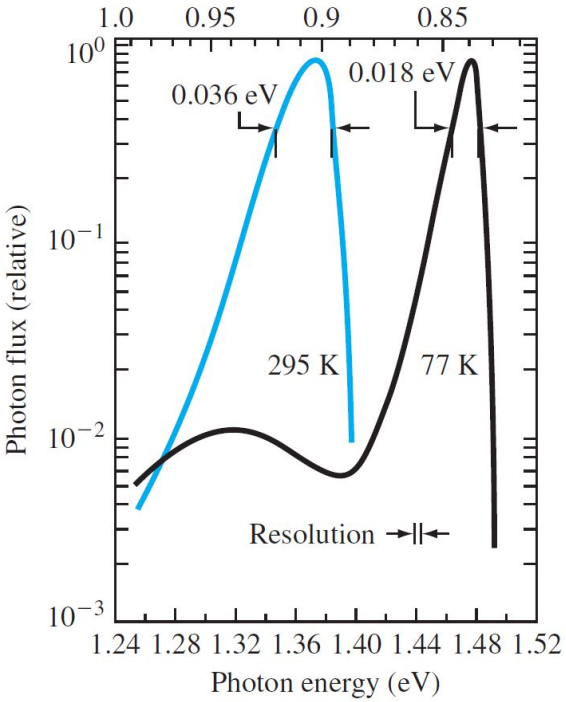


Fig. 7.20 Electroluminescence spectra of a GaAs pn-junction. From [1].

To take an important example of confinement of minority carriers, we consider luminescent devices with pn-junctions as the injectors of minority carriers. Such electroluminescent devices are called **light emitting diode** (**LED**).

There are various processes of photon emission by the recombination of injected minority carriers, but here we restrict ourselves to the direct recombination of electrons in the conduction band and holes in the valence band. The luminescent intensity $I(\nu)$ is written as

$$I(\nu) \propto \nu^2 (h\nu - E_g)^{1/2} \exp\left[\frac{-(h\nu - E_g)}{k_B T}\right]. \qquad (7.51)$$

Figure 7.20 shows an example of luminescent spectra from a GaAs pn-junction. With decreasing the temperature, the band gap E_g widens mainly due to the variation of lattice constant. As a result, the luminescent peak narrows and shifts to high-energy (blue-shift). As for the second peak in the spectrum at 77 K, the authors of Ref. [1] commented only the existence, but it looks like the luminescence from the impurities.

Important parameters of LED characteristics are the wavelength and the efficiency. In the case of homo pn-junction luminous layer, the wavelength is almost determined by the band gap as in Eq. 7.51. The quantum efficieency η_q is

defined as the ratio of the radiative recombination processes $R_{\rm r}$ to the total recombination processes R of the injected carriers as

$$\eta_{\rm q} \equiv \frac{R_{\rm r}}{R} = \frac{\tau_{\rm nr}}{\tau_{\rm nr}+\tau_{\rm r}} = \frac{\tau_{\rm tot}}{\tau_{\rm r}}, \quad \frac{1}{\tau_{\rm tot}} \equiv \frac{1}{\tau_{\rm nr}} + \frac{1}{\tau_{\rm r}}, \tag{7.52}$$

where $\tau_{\rm nr}$, $\tau_{\rm r}$ are the lifetimes limited by non-radiative recombination and radiative recombination respectively. $\tau_{\rm tot}$ is the total lifetime of minority carriers obtained from Matthiessen's rule. The interband radiative recombination probability is proportional to the electron-hole density product, that is

$$R_{\rm r} \propto np. \tag{7.53}$$

Under the injection of minority carriers, the law of mass action naturally does not hold: $np \neq n_i^2$.

The current density of minority carriers is, as seen in Eq. (6.11), gvien by the sum of the electron flux density, the hole flux density,

$$j_e + j_h = e\left[\frac{D_e n_{p0}}{L_e} + \frac{D_h p_{n0}}{L_h}\right]\left[\exp\left(\frac{eV}{k_{\rm B}T}\right) - 1\right], \tag{7.54}$$

and the recombination rate inside the depletion layer (width w_d) expressed in the form of current as

$$j_{\rm R} = \frac{e n_i w_d}{2\tau_0}\left[\exp\left(\frac{eV}{2k_{\rm B}T}\right) - 1\right]. \tag{7.55}$$

The recombination in the depletion layer mainly occurs at mid-gap deep levels resulting in the factor 1/2 in the term eV just as in Eq. (6D.13). We take the case of an n^+p junction, in which the n-side is heavily doped and the luminescence is mainly by electron-hole recombination in the p-layer. Then the injection efficiency of the junction is

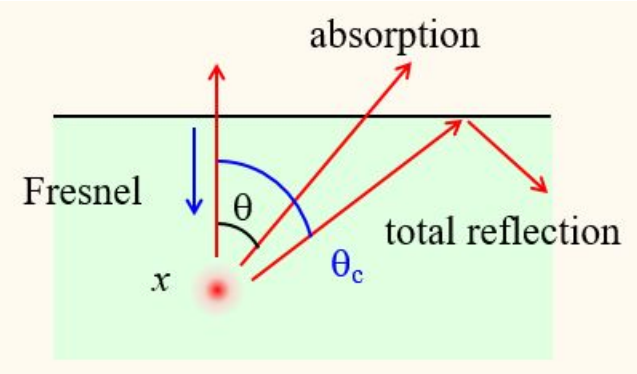
$$\gamma = \frac{j_e}{j_e + j_h + j_{\rm R}}. \tag{7.56}$$

The **internal quantum efficiency**is thus defined from Eq. (7.52) and Eq. (7.56) as

$$\eta_{\rm iq} = \gamma\eta_{\rm q}. \tag{7.57}$$

There are various limiting factors of the internal quantum efficiency, some of which are related to the crystalinity as deep levels. The device structures also affect the efficiency through the surface recombination, etc.

The **external quantum efficiency**has the ultimate importance for the LEDs. As we can see from Eq. (7.51), the energy of photons emitted by direct interband transition has a peak slightly above the energy gap. Then the reabsorption of light by the crystal itself occurs and the absorbed photons result in the absorption loss.



As shown in the left figure, we take x as the distance from the surface to the emission point, θ as the angle of the ray from vertical to the surface. Let α be the absorption coefficient and we have the absorption loss

$$\zeta_{\rm abs} = 1 - \exp(-\alpha x/\cos\theta). \tag{7.58}$$

When a light pass through the interface between the materials with refractive indices $\bar{n}_1$ and $\bar{n}_2$, we have the reflectance

$$\Gamma = \left(\frac{\bar{n}_2 - \bar{n}_1}{\bar{n}_1 + \bar{n}_2}\right)^2. \tag{7.59}$$

The loss at the surface by the reflection is called Fresnel loss. Because the reflactive index inside semiconductors is generally larger than that in the vacuum or in the air, when θ exceeds the critical angle $\theta_{\rm c}$, the surface causes total reflection, which reults in the optical loss. The ratio of the number of photons finally emitted from the surface $n_{\rm f}$ to that of photons once produced inside the crystal is called **optical efficiency**. The ratio of $n_{\rm f}$ to the number of injected

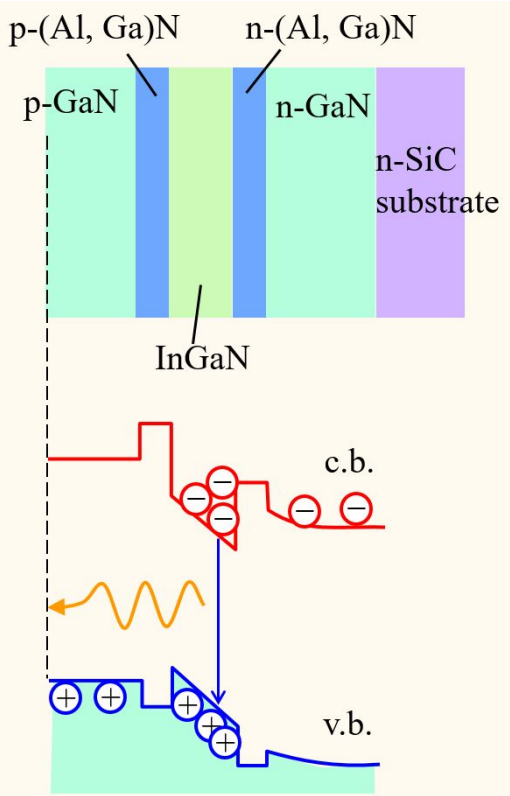


Fig. 7.21 Conceptual diagram of double-heterojunction LED. InGaN is taken as the material for the active layer.

carriers is called external quantum efficiency. Let $\eta_{\rm opp}$, $\eta_{\rm exq}$ be the optical efficiency and the external quantum efficiency respectively, then from the definitions

$$\eta_{\rm exq} = \eta_{\rm opp}\eta_{\rm iq}. \tag{7.60}$$

Generally, simple pn-junctions have very low external quantum efficiency less than few %.

As in the solar cells, surface textures to cause multiple reflection are effective to reduce the Fresnel loss and the total reflection loss. And also like the case for solar cells, **double-heterojunction(DH)** is frequently used to enhance the internal quantum efficiency and to reduce the absorption loss. The concept is shown in Fig. 7.21. The radiative layer is inserted between the cladding layers of materials with larger band gap than the active material. In the figure, the chemical dopings are just done in the cladding layers. The elimination of chemical doping in the active layer reduces the recombination in the depletion layer. Injected minority carriers are confined into the thin active layer, resulting in high np product and in high internal quantum efficiency. Furthermore, the energy of emitted photon is less than the band gap of cladding layers and the absorption does not occur there. Mirror-like layers are often placed at the back planes to reflect forward the photons emitted backward. In Ref. [2], the authors reported the external quantum efficiency of 77% in YAG active type LED with InGaN-LED activation. $\eta_{\rm exq} \sim 30$ % were the highest then, and the case of YAG is extraordinary. But now the technology is widely used for LED illumination.

7.6.2 Laser diode

The light emission in LEDs is by spontaneous emission drawn in Fig. 4.1(b). Now we consider the stimulated emission drawin in (c). $|A_0|^2$ in the transition probability is proportional to the energy density of electromagnetic field. We write the energy density as $n_\lambda \hbar\omega_\lambda/V$, where V is the system volume, n_λ is the number of photons in mode λ. Such a coherent electromagnetic field excites electric dipole moment $\boldsymbol{\mu}$ in the material, creating the transition element between $|a\rangle$ and $|b\rangle$ (see Fig. 4.1(c) for the two-level system). If we write $\boldsymbol{r} = \boldsymbol{r}_0\cos(\omega_0 t)$, then $\boldsymbol{p} = m\omega_\lambda \boldsymbol{r}$. We rewrite $\vec{e}\cdot\boldsymbol{p}$ as $(\omega_\lambda m/e)\vec{e}\cdot\boldsymbol{\mu}$

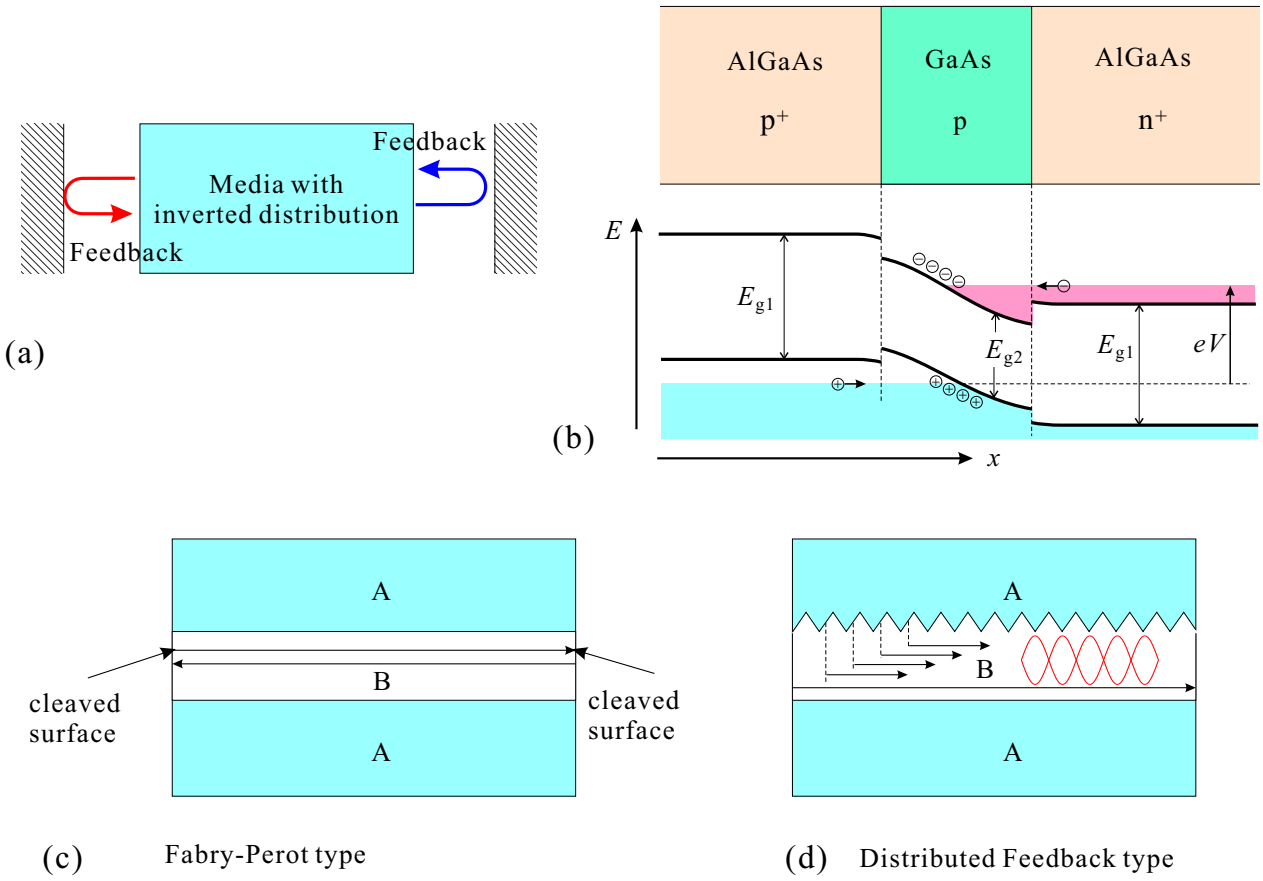


Fig. 7.22 (a) Confinement of photons with parallel mirrors. Also can be viewed as the feedback by mirrors. (b) Illustration of double heterojunction laser diode structure. AlGaAs and GaAs are taken as representative materials. With forward bias, the distribution inversion is realized. (c) Fabry-Pérot type lase structure, which is constituted of parallel mirrors of cleaved surfaces. Material B is sandwitched by material A with larger band gap. (d) Distributed feedback (DFB) type laser structure, in which corrugation type grating is built in at the A-B interface.

and put $\omega = \omega_0 = \omega_\lambda$. Then the probability of stimulated emission with the transition $|b\rangle \to |a\rangle$ is

$$P_{ba}(t) = \frac{\omega_\lambda}{\epsilon\epsilon_0 \hbar V} |\langle a|\vec{e}\cdot\boldsymbol{\mu}|b\rangle|^2 n_\lambda \frac{t^2}{2}, \tag{7.61}$$

which is proportional to n_λ. The symmetry of Eq. (7.61) tells that the probability of light absorption with transition $|a\rangle \to |b\rangle$ is $P_{ab} = P_{ba}$. Equation (7.61) tells that the more photons in a mode the higher the probability of stimulated emission to this mode. The phenomenon can be interpreted as a **Bosonic stimulation**, which is the origin of Bose-Einstein condensation. And the photonic state is described as a coherent state[3]. The coherence can be understood in a classical picture that $\boldsymbol{\mu}$ is excited coherently by the electromagnetic field.

As a model of the medium of photon propagation, we consider a set of such two-level systems. Let N_a, N_b be the concentrations of the two-level systems at the state $|a\rangle$, $|b\rangle$ respectively. When the light of ω_λ propagates the media, the enery absorbed by the media in unit volume is written as

$$\mathcal{E} = (N_a - N_b) P_{ba}(\tau) \hbar\omega_\lambda, \tag{7.62}$$

where τ is the averaged interaction time of light with each two-level system. If the state $N_b > N_a$ is realized $\mathcal{E} < 0$, namely the light absorbs energy from the media and the light is amplified. The light is in coherent state with a common phase of photon. Such amplification of photons (increment in the photon number in the same mode) and the device (apparatus) to realize the phenomenon is called **light amplification by stimulated emission of radiation**, **LASER**.

A laser diode (LD) is a light emitting element that uses a pn junction like an LED, but uses a double heterojunction (or a stronger confinement structure) to create an inversion distribution and to cause laser action. In order to strongly amplify light, it is necessary to advance the light in the population inversion medium, but the light is also strongly amplified by confining it in the resonator using a mirror surface and reciprocating in the same medium(Fig. 7.22(a)). Figure 7.22(b) shows an LD structure in the beginning of the research. An example of the simplest Fabry-Pérot type cavity of the laser oscillation is illustrated in Fig. 7.22(c). In Fig. 7.22(d), the structure called distributed feedback (DFB) type laser diode is illustrated. A corrugation is introduced at the hetero-interface as a grating and to make the structure a cavity.

Chapter 8 Basics of quantum transport

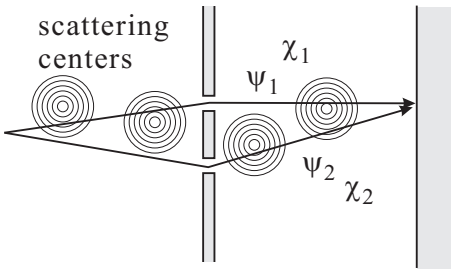
Let us go into quantum transport, which is one the major subjecs in semiconductor physics, because one-dimensional systems are the best for the construction of theoretical models.

8.1 Classical transport and quantum transport

We treated electrons as particles in the secton of "classical transport" besides counting the number of cases while in the sections of heterojunction and quantum confinement, they were treated as quantum mechanical waves. The difference in the treatment of electrons in the same material comes from the scales of energy and spatial range. Until now we have treated pn-junctions classically and double barrier diodes quantum mechanically, but these two are actually marginal cases. In a pn junction, the deletion layer becomes thinner with increasing the doping concentrations. When both the p-layer and the n-layer are highly doped, the depletion layer is very thin and the Fermi level penetrates into the conduction band in n-layer and the valence band in p-layer. Now in the both layers the density of states exist around the Fermi level and they are very close with a thin separation. Then a quantum tunneling, which is nothing but a quantum phenomenon occurs through the depletion layer. The structure is called "Esaki diode", a representative quantum device. On the other hand in some double barrier diodes, depending on the materials and the structures, no resonance peak can be observed *e.g.* at room temperatures. Let us briefly discuss here in what case we need to treat a phenomenon quantum mechanically. We already had a very short discussion in the beginning section of classical transport. Let us go a bit deeper here.

Now in what case does quantm coherence appear in transport? The "length" which expresses the criteria is **quantum cohrence length** [*1] . In very short, when an electron travels in a solid, the averaged length over which the electron propagates with quantum coherence, is called "quantum coherence length" and often written as l_ϕ.

Whether we can observe quantum coherence in experiments or not depends not only on the essential quantum coherence of each particle but also on the coherence between the particles *i.e.*, statistical fluctuation of the interference. The former is, in very short, due to **quantum entanglement** with a large number of surrounding freedoms (environment).



Let us consider a double slit experiment shown in the left. The interference pattern on the screen is

$$|\psi|^2 = |\psi_1 + \psi_2|^2 = |\psi_1|^2 + |\psi_2|^2 + 2|\psi_1||\psi_2|\cos\theta,$$

where the third term in the RHS is the quantum interference. Consider the situation that from the starting point to the screen there exist some scatterings, at which the electron has interactions with a quantum mechanical state χ (quantum mechanical freedom other than the electron). The interaction should be different for the two paths 1 and 2, then

$$\psi_1 \to \psi_1 \otimes \chi_1, \quad \psi_2 \to \psi_2 \otimes \chi_2.$$

As a result, the interference term changes into

$$2|\psi_1||\psi_2|\cos\theta\langle\chi_1|\chi_2\rangle.$$

[*1] The word "coherence length" is used in various ways in different meanings. In condensed matter physics, for example, it appears in treating superconductivity for multiple meanings.

Hence if $\chi_1 \perp \chi_2$ the inner product is zero and the interference term vanishes. In such a state, two freedoms ψ and χ are in **maximally entangled state** (Appendix F). In other words, the quantum coherence length in this case is the length over whch the electron (freedom) makes up a maximally entangled state with another degree of freedom (**envirnment**). A little question here is that χ_1 and χ_2 may be orthogonal at one-moment but the time evolution after that may restore the interference killing the orthogonality. There are, of course, many such setups [*2]. Here we adopt that in not-specially designed quantum systems, with time evolves the entanglement spreads over many other freedoms and disentanglement never occurs.

There is another kind of "dephasing" in experiment. Even if each particle is able to interfere with itself, when the wavelengths of particles are widely distributed, in other words monochromaticity is not high enough, the interference patterns are also distributed and averated out. Let us estimate the characteristic length, over which the difference in the patterns is small enough for them to survive after averaging. Electrons being fermion, the energy of movable electron at absolute zero is $E_{\rm F}$, *i.e.*, they are completely monochromatic. The energy width appears at a finite temperature T as $\Delta E = k_{\rm B}T$. The difference in the electron phase accumulated during time τ is $2\pi\Delta f\tau = 2\pi\Delta E\tau/h = 2\pi k_{\rm B}T\tau/h$. A criterion in time can be the time for the phase difference becomes 2π, that is

$$\tau_c = \frac{h}{k_{\rm B}T}.$$

In diffusive transport, the diffusion length is written as $l = \sqrt{D\tau}$, and this determines a kind of coherence length $l_{\rm th}$ as

$$l_{\rm th} = \sqrt{\frac{hD}{k_{\rm B}T}}, \tag{8.1}$$

which is called **thermal diffusion coherence length**. In ballistic transport, the electrons get few scatterings during the traversal through the sample. They get through the sample with the velocity $v_{\rm F}$ and

$$l_{\rm th} = \frac{hv_{\rm F}}{k_{\rm B}T}. \tag{8.2}$$

After traversal over the above **themal length**, the coherence is lost from the result of averaging even though intrinsic quantum coherence survives. Attention should be paid for $l_{\rm th}$ particularly in experiments.

After knowing l_ϕ, what are the conditions for the quantum mechanics to appear in transport? Firstly we should list up the case that the sample size is shorter than l_ϕ. For even shorter sample size, shorter than the representative de Broglie length (*i.e.* the Fermi wavelength), as we already saw, quantum confinement effect (descretized energy levels) emerges, into which we do not go into here. Secondly, we often have some characteristic lengths in transport besides the sample size. A representative is the **magnetic length**, which appears when an external magnetic field is applied. The magnetic length, or minimum cycrotron radius is written as $l_B = \sqrt{h/eB}$ for magnetic flux density B. When $l_B \leq l_\phi$, there appear various quantum effects in magneto-transport.

We finish up this very short section for quantum coherence and decoherence here. Below we go into quantum coherent transport.

8.2 Landauer formula

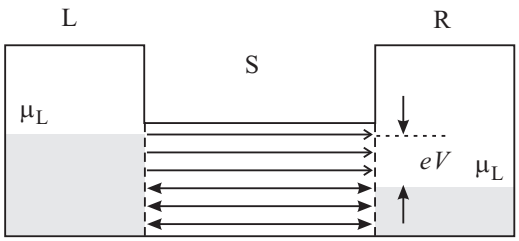
In this lecture, I would like to introduce one view point, from which we view the transport in quantum systems as the conduction in "quantum circuits"[*3]. In this section we see the most basic part of it.

[*2] This can be experimentally verified. With this fact, some people claim that the theory of decoherence based on environmental freedom is wrong. But this is, of course, misunderstanding. In the theory of decoherence from the environmental freedom, "intrinsic decoherence" does not exist besides the thermodynamic limit. In real systems no thermodynamic limit has been achieved though the results of statistical mechanics apply. In the same way, with progress of entanglement with many degree of freedoms, the disentanglement becomes more difficult and impossible at last.

[*3] Here "quantum circuit" is different from what we use in schematization of quantum information manipulation.

The **Kubo formula** is an ultimate form of linear response, which was studied from the beginning to middle of 20th century in Bell labs and other places. It is now an indispensable tool for theoretical studies in condensed matter physics. On the other hand in practical analyses of experiments, the **Landauer formula**, which can be derived as a one expression of the Kubo formula[4], is often used. I hope students refer to other lectures *e.g.*, statistical physics, for the introduction of the Kubo formula and here we go into quantum transport with the simplest introduction of the Landauer formula.

8.2.1 Conductance quantization



The lowest dimension in which "transport" exists is one. We thus first consider the conductance of a one-dimensional fermion system. Here we adopt an ingenious modeling by Rolf Ladauer, illustrated in the left figure. A one dimensional conductor without scattering is connected to two particle reservoirs, in which the chemical potentials are well defined as they have huge (infinit) number of particles and are in thermal equilibrium. Let the chemical potentials of left and right reserviors as $\mu_{\rm L}$, $\mu_{\rm R}$ respectively. The current brought by a state with wavenumber k can be written as

$$j(k) = \frac{e}{L} v_{\rm g} = \frac{e}{\hbar L} \frac{dE(k)}{dk}, \tag{8.3}$$

where L is the length for normalization, thus e/L is the charge density. The total current J then is

$$J = \int_{k_{\rm L}}^{k_{\rm R}} j(k) \frac{L}{2\pi} dk = \frac{e}{h} \int_{\mu_{\rm R}}^{\mu_{\rm L}} dE = \frac{e}{h}(\mu_{\rm L} - \mu_{\rm R}) = \frac{e^2}{h} V. \tag{8.4}$$

The conductance is finally obtained as

$$G = \frac{J}{G} = \frac{e^2}{h} \equiv G_q \equiv R_q^{-1}. \tag{8.5}$$

This is the conductance for one-dimensional conductor without scattering and called **conductance quantum**. If we consider the spin degree of freedom, and when the spin can be treated as just a double degeneracy of quantum states, we simply multipy $G_{\rm q}$ by two and may call $2e^2/h$ a conductance quantum. R_q is called quantized resistance.

The above discussion is, in a sense, a paraphrase of te uncertainly principle. Let us see that in a more transparent form. The problem is equivalent to that we pack wavepackets with a width Δk in k-space into a one-dimensional fermion system. The highest charge density in the system is $e/\Delta x$ for wavepackets with a width Δx in the real space. The velocity of the packet is $\Delta E/\hbar\Delta k$, giving the current as

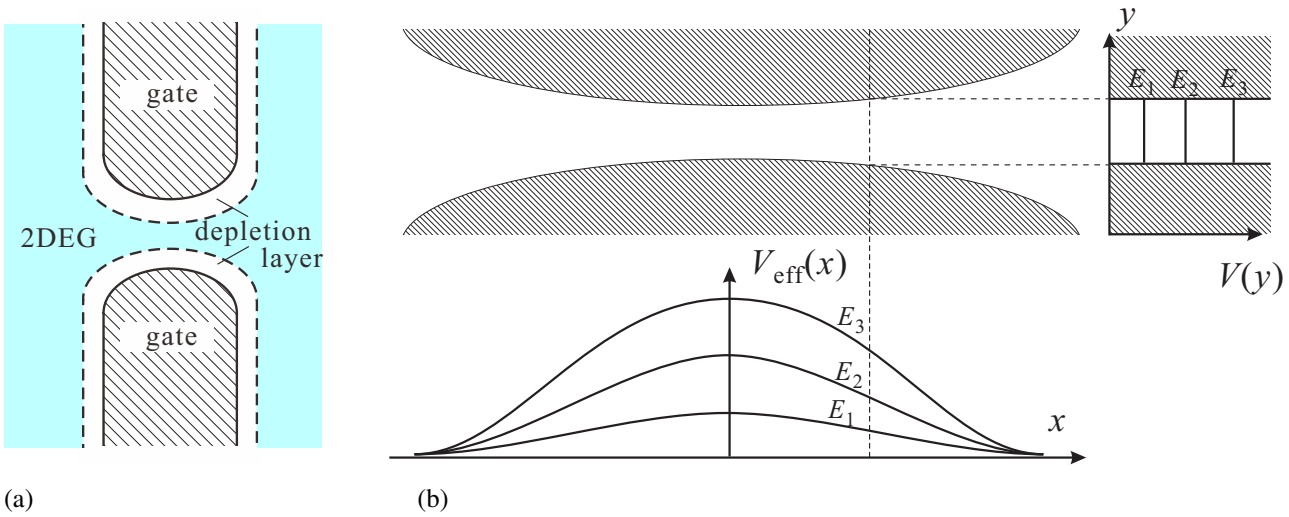
$$J = \frac{e}{\Delta x} \frac{\Delta E}{\hbar \Delta k} = \frac{e^2}{h} V, \tag{8.6}$$

which is the same result as before. Here we write $\Delta x \Delta k = 2\pi$, $\Delta E = eV$.

8.2.2 Quantum point contact and concept of conductance channel

One dimensional fermion system discussed above is in other expression **quantum wire** (QW) or **quantum point contact** (QPC). A way to realize them in semiconductor structures is to confine a two-dimensional electron gas (2DEG) into a narrow region.

In the case of QPC, "a narrow region" means, as in Fig. 8.1(a), a narrow short region gradually squeezed from a wide 2DEG. As is easily imagined, such a structure can be realized with the split gate technique introduced in Sec.**??**. This can be modelized as in Fig. 8.1(b). x-axis is taken to longitudinal to the QPC "waveguide". Here we assume **adiabatic**



(a) (b)

Fig. 8.1 (a) Schematic of quantum point contact. (b) Simplified model of a QPC. Upper panel: Electrons are excluded from hatched regions and confined in white region, one dimensionally at the narrow gap. The right figure shows the confinement potential along the cross section at the broken line. Descrete eigen energies $E_{1,2,3}$ correspond to the three effective potential drawn in the lower panel. Lower panel: Illustrates effective potentials $V_{\rm eff}(x)$ in eq.(8.8).

propagation of electrons through a QPC, that is, the total energy of an electron $E = E_{kx} + E_{ky}$ does not change during the traversal though E_{kx}, E_{ky} transform each other.

Though harmonic potentail like in Fig. **??**(b) is generally a good approximation for such kind of confinement, here we take, for simplicity, the hard-wall approximation illustrated in Fig. 8.1(b). With W being the width of confinement, the wavefunction in y-direction is written as $\varphi_n(y) = \cos(n\pi y/2W)$ (n: an odd number), $\sin(n\pi y/2W)$ (n: an even number). We assume that change of W for x is slow enough so that we can separate x and y dependencies as $\psi(x,y) = \varphi_n(y)\phi(x)$. Then, the equation is

$$\begin{aligned} H\psi(x,y) &= \frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\varphi_n(y)\phi(x) \\ &= \varphi_n(y)\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \left(\frac{n\pi}{2W}\right)^2\right)\phi(x) = E\varphi_n(y)\phi(x). \end{aligned} \tag{8.7}$$

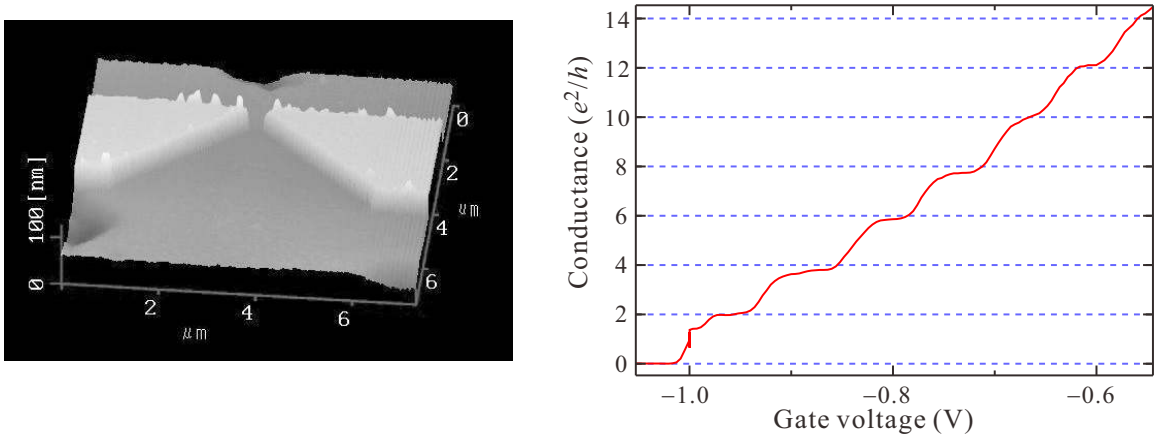
(8.7) depends on x, and the assumption of adiabaticity requires it holds for each x-position. This can thus be viewed as a potential problem with effective potential along x-direction

$$V_{\rm eff}(n,x) = \frac{\hbar^2}{2m}\left(\frac{n\pi}{2W(x)}\right)^2. \tag{8.8}$$

The situation is illustrated in the lower panel of Fig. 8.1(b). The effective potential $V_{\rm eff}(n,x)$ has index n, which is for the descrete quantized energy along y. The partitioning of total energy is, then,

$$E_{\rm tot} = E_{kx}(n,x) + V_{\rm eff}(n,x), \tag{8.9}$$

and we can treat a propagating state as a one-dimensional one indexed with n. Such a one-dimensional state is called **conductance channel**, the density of states to which is proportional to $1/\sqrt{E - E_{\rm c}}$. When the system is in equilibrium, $E_{\rm F}$ is, of course, constant over the system though the effecive potential $E_{kx}(n,x)$ is channel dependent and thus the Fermi wavenumber $k_{x\rm F}$, the density of states should be determined for each channnel.



(a) (b)

Fig. 8.2 (a) Atomic force microscope of a QPC gate structure. White raised regions are the gate electrodes placed on an AlGaAs/GaAs two-dimensional electrons. (b) Conductance of a QPC at 30 mK as a function of the gate voltage.

8.2.3 Transport experiments in QPCs

Let us see some experimental results on transport through real QPCs. With increasing the negative voltage to the gate electrodes, the effective potential in (8.8) becomes higher due to the narrowing of $W(x)$ and the number of conduction channels whch can go over the potential top decreases.

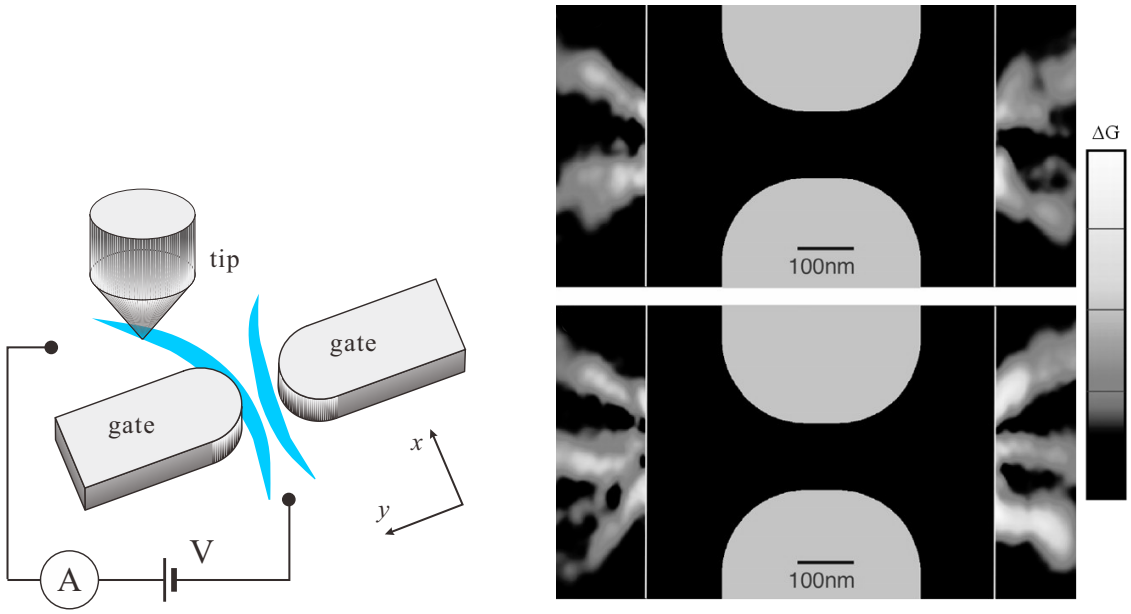
Figure 8.2(a) shows an AFM image of a split gate structure fabricated with nano-fabrication techniques. In Fig. 8.2(b), we plot the electric conductance G of the QPC as a function of the gate voltage $V_{\rm g}$. G shows staircase-like variation versus $V_{\rm g}$ with a constant height of stairs about $2e^2/h$. Namely G is quantized to an integer times $2e^2/h$. The system holds the time-inversion symmetry and the spin degeneracy. Hence the experiment confirms the result of Eq. (8.5).

In the experiment shown in Fig. 8.3, the conductance of a QPC is adjusted on the plateau of $n \times 2e^2/h$ (n: integer) and the tip of an atomic force microscope (AFM) is placed just on the surface close to the QPC. Then the image potential of the tip in the two-dimensional electrons causes weak scattering of the electron wave resulting in a small shift of the conductance from the quantized value. With scanning the tip the shift is plotted versus the tip position, then as in Fig. 8.3(b), on the plateau of $n = 2$, we observe a wave with two anti-node is flowing out from the QPC. The number of anti-node is three for $n = 3$ and one for $n = 1$. The above results shows the number of anti-node of standing wave along y, that is the number of channels transmitting through a quantum wire is equal to the quantization number n of the conductance [*4].

8.2.4 Conduction channel and transmission probability

In the above we have introduced the concept of conduction channel referring to the QPC experiments. The shortness of QPC is to escape from the scattering and longer structure is available if the mean free path exceeds the size. Actually, in longer quantum wires made of high-mobility two-dimensional electron systems, the quantization of conductance has been observed. Next we consider the widening of the quantum wire. With the increase of width, the level spacing of quantization along the width narrows, the number of states below the Fermi level $E_{\rm F}$ increases if the position of $E_{\rm F}$ from

[*4] Some of you may think that if the number of channels is, say 3, then the waves with antinodes 1, 2, and 3 should be overlapped. The argument is correct. However, the density of states of one-dimensinonal systems is expressed as $1/\sqrt{\epsilon - \epsilon_0}$ with ϵ_0 as the band edge. Then in actual measurement, the amplitude of wavefunction with highest channel is detected. Also, the electrons traversing on the highest channel have the lowest the kinetic energy along x and easily scattered by the probe potential, detected in the experiment.



(a) (b)

Fig. 8.3 (a) Illustration of experimental setup to measure the wavefunction amplitude with a scanning probe microscope (SPM). With measuring the conductance of the QPC, the tip is scanned over the conduction channel. (b) The image of the shift in the conductance from the quantized value measured with this setup. The center part is drawn from the topography obtained by AFM measurement. The upper is measured on the $n_{\rm ch} = 2$ conductance step. The lower is for $n_{\rm ch} = 3$. (The data are taken from Topinka *et al.*, Science **289**, 2323 (2000))

the bottom of the band is fixed. We take the limit of infinity in width and the system is now a two-dimensional. We write the electron density as $n_{\rm 2D}$ then we find the number of channels per unit length is $\sqrt{n_{\rm 2D}}$.

So far we have considered systems without scattering. What we are treating here is coherent quantum transport and random inelastic scatterings by phonons etc. are out of scope. However, the potential scatterings by impurities or lattice imperfections do not break quantum coherence and they can be taken into account. The scatterings are transitions between the propagating states, or from the view point introduced here, transitions between the conduction channels. Hence we express the scattering centers with points as in Fig. 8.4(b) and through the points electrons enter different channels. Note that they are quantum mechanical scattering and the electrons do not completely "change" their tracing lines at the scatters. Instead the electron waves are devided at the scattering points and continue propagation. The conduction channels play the role of waveguide for microwaves. With such waveguides and joints in various shapes, we can separate or join microwaves. At such joints there also should exist reflection which reverse the direction of propagation. The same should happen at the scattering centers. When the number of scattering center increases and the system can be viewed as a "disordered metal," the system in the channel expression is like a cobwebs as illustrated in Fig. 8.4(c). At first sight, it looks difficult to treat. But instead of treating inside, we just pay our attentions to the channels at the inlet and at the outlet. We write the transmission probability of electron propagation from i-th channel at the inlet to j-th channel as T_{ij}. From the fact that the single channel without scattering has the conductance of e^2/h with the transmission probability $T = 1$, the conductance G of a conductor that has the matrix of transmission probability $\{T_{ij}\}$ is (with consideration of the spin degree of freedom 2)

$$G = 2\frac{e^2}{h}\sum_{i,j} T_{ij}. \tag{8.10}$$

Equation (8.10) is called **Landauer formula for 2-terminal conductance**.

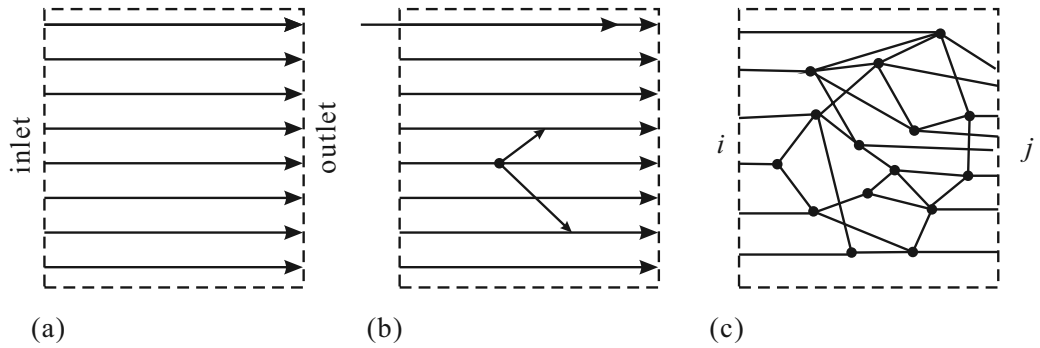


Fig. 8.4 (a) A two-dimensional conductor is expressed as a set of one-dimensional channels. Two-terminal configuration. (b) Introduction of a scattering center, which causes transitions between the conduction channels. (c) Disordered conductor with multiple scattering centers.

8.3 S-matrix

We have introduced scattering centers (joints) through the comparison of conduction channels with waveguides for microwave. Actually researchers often call conduction channels **electron waveguides**. Also, we often use interference circuits, in which quantum wires are joined/devided at some points just like joints of waveguides. For the treatment of such joints, **scattering matrix**, **S-matrix**is often used as is in the case of microwave circuits. As in Fig. 8.5(b), we write the wavefunctions coming into a scatterer from left and right as $a_1(k)$, $a_2(k)$ respectively, and the same for outgoing ones as $b_1(k)$, $b_2(k)$. The S-matrix representing the scatterer is defined as

$$\begin{pmatrix} b_1(k) \\ b_2(k) \end{pmatrix} = S \begin{pmatrix} a_1(k) \\ a_2(k) \end{pmatrix} = \begin{pmatrix} r_{\mathrm{L}} & t_{\mathrm{R}} \\ t_{\mathrm{L}} & r_{\mathrm{R}} \end{pmatrix} \begin{pmatrix} a_1(k) \\ a_2(k) \end{pmatrix}, \tag{8.11}$$

where $t_{\mathrm{L,R}}$, $r_{\mathrm{L,R}}$ are complex transmission and reflection ratios from left and right respectively. They bare **phase shifts** occurring at the scattering in their complex phases. Here we adopt the lower case expression for the "wavefunction flows" in order to distinguish them from $A_i(k)$ etc. so far used because the directions of the flows are different by definition. There are the relations to transmission and reflection probabilities $T_{\mathrm{L,R}}$, $R_{\mathrm{L,R}}$ as

$$T_{\mathrm{L,R}} = |t_{\mathrm{L,r}}|^2 = 1 - R_{\mathrm{L,R}} = 1 - |r_{\mathrm{L,R}}|^2. \tag{8.12}$$

Unlike T-matrices, S-matrices cannot have the output as the next input because the channels are mixed in the operand vectors. On the other hand, as seen in Eq. (8.11), each element has clear physical meaning and the parameters of the scatteing can be readily extracted.

In the above, in a sense, we have considered a connection of two channels with the same wavenumber. If we consider the extension to more general cases, we need to take care that each channel has different wavenumbers, dispersions. Even in the simplest case of a QPC, when it is on the plateau of $G = n \times 2e^2/h$, it has n conduction channels and the Fermi

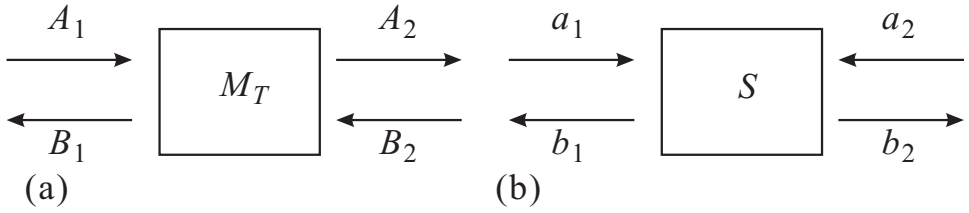


Fig. 8.5 (a) Conceptual diagram of T-matrix M_T. (b) Conceptual diagram of S-matrix S.

wavelengths are different for different channels. In such a case, we cannot simply use wavefunctions for $a_1(k)$. Instead, we write

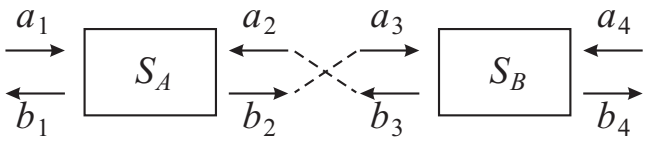
$$a_i(k) = \sqrt{v_{\rm Fi}}\psi_{ai}(k_{\rm F}), \tag{8.13}$$

where $\psi_{ai}(k_{\rm F})$ is the wavefunction (the same for b_i). Under this definition, the norms of input/output vectors represent the strengths of "probability density fluxes." We call t as a comlex transmission probability and $|t|^2 = T$ is (real) transmission probability. Then in this way, we can call $a_i(k)$ etc. in (8.13) as **complex probability flow**.

8.3.1 Connection (joint) of S-matrices

For the series connection of T-matrices, as we did in the double barriers, we can simply take the product of them, which procedure simplifies the calculation and saves the trouble. On the other and for the series connection of S-matrices, as in the figure shown below, the eight lines for input/ouput should be in cross connection and the results should be expressed in terms of a new S-matrix. For the calculation we first write

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = S_A \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} r_{\rm L}^{(\rm A)} & t_{\rm R}^{(\rm A)} \\ t_{\rm L}^{(\rm A)} & r_{\rm R}^{(\rm A)} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad \begin{pmatrix} b_3 \\ b_4 \end{pmatrix} = S_B \begin{pmatrix} a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} r_{\rm L}^{(\rm B)} & t_{\rm R}^{(\rm B)} \\ t_{\rm L}^{(\rm B)} & r_{\rm R}^{(\rm B)} \end{pmatrix} \begin{pmatrix} a_3 \\ a_4 \end{pmatrix}. \tag{8.14}$$



By using the boundary conditions

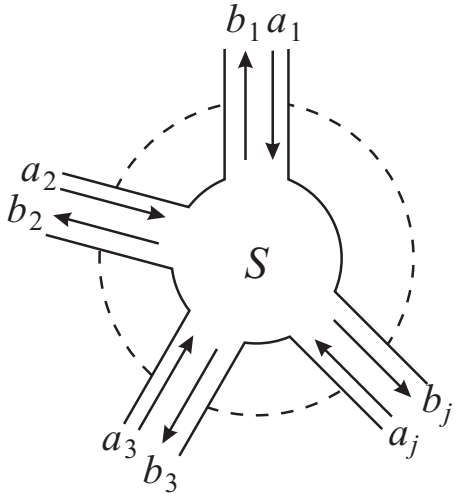
$$b_2 = a_3, \quad a_2 = b_3, \tag{8.15}$$

we drop these variables from the final simultaneous equations, to get the single S-matrix. The result is the following S-matrix $S_{\rm AB}$.

$$S_{\rm AB} = \begin{pmatrix} r_{\rm L}^{(\rm A)} + t_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)} \left(I - r_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)}\right)^{-1} t_{\rm L}^{(\rm A)} & t_{\rm R}^{(\rm A)} \left(I - r_{\rm L}^{(\rm B)} r_{\rm R}^{(\rm A)}\right)^{-1} t_{\rm R}^{(\rm B)} \\ t_{\rm L}^{(\rm B)} \left(I - r_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)}\right)^{-1} t_{\rm L}^{(\rm A)} & r_{\rm R}^{(\rm B)} + t_{\rm L}^{(\rm B)} \left(I - r_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)}\right)^{-1} r_{\rm R}^{(\rm A)} t_{\rm R}^{(\rm B)} \end{pmatrix}. \tag{8.16}$$

At the first sight, it looks just complicated and you may wonder why we need to take such a way for calculation. However, the expression shows the behavior of wave propagating over two series scatterers. To see that we take out (1,1) element of Eq. (8.16) and expand the second term as

$$\left(I - r_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)}\right)^{-1} = I + r_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)} + (r_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)})^2 + (r_{\rm R}^{(\rm A)} r_{\rm L}^{(\rm B)})^3 + \cdots. \tag{8.17}$$



This clearly shows that the second term is the summation of the processes, each of which is a reflection including multiple reflections between the two scatterer A and B. Elements of S-matrices have clear meanings as in Eq. (8.11) and are easy to be interpreted. And because the inputs and the outputs are separated, we can easily apply them for multiple channels or electrodes.

In the above, we did not consider the evolution of phase when the wave propagetes between the scatterers. This can be taken into account by inserting T- or S-matrix to express the phase rotation. With this we can treat the case that the transmission line works as a resonator.

We have redefine the input/output as in Eq/ (8.13), and expansion to multi-channel systems can be done with increasing the dimension of input/output vectors. When we write down (8.17), the care is taken for the

order of product, the denomination is expressed as the multiplication of the inverse, and 1 is expressed as I. These are for the expansion to multi-channel systems with converting a_1 etc. to vectors, r $(\mathrm{A})_{\mathrm{r}}$ etc. to matrices.

We do not consider any interaction (scattering) between the channels on the transmision line. Therefore, in the case of multiple channel, assignment of the channels to the transmission lines is not a crutial problem in the S-matrix treatment. Hence, as illustrated in the figure, the **wire connection** can be done regardless of the lines. In that sense, S-matrices are "nodes" of the lines. T-matrices shceme is not so easy to be applied to channel multiplication. We utilize both methods to treat the electron waveguide circuits.

8.4 Onsager reciprocity

An important property of S-matrices is the **unitarity**. From the definition of complex probability flow in (8.13) and the conservation of probability requires $|\boldsymbol{a}|^2 = |S\boldsymbol{a}|^2$. Then it is almost trivial that S-matrices should be unitary. From the unitarity, a very important property in the symmetry called **Onsager reciprocity** is derived. The Onsager reciprocity, which holds generally in the transport phenomena, is expressed in the form of S-matrix as

$$S(\boldsymbol{B}) = {}^tS(-\boldsymbol{B}) \quad (S_{mn}(\boldsymbol{B}) = S_{nm}(-\boldsymbol{B})), \tag{8.18}$$

where $\boldsymbol{B}$ is the external magnetic field.

The derivation is as follows. The problem here is essentially the potential scattering described by the Schrödinger equation

$$\left[\frac{(i\hbar\nabla + e\boldsymbol{A})^2}{2m} + V\right]\psi = E\psi. \tag{8.19}$$

We take the complex conjugate of (8.19) and revert the direction of the magnetic field with $\boldsymbol{A} \to -\boldsymbol{A}$ to get

$$\left[\frac{(i\hbar\nabla + e\boldsymbol{A})^2}{2m} + V\right]\psi^* = E\psi^* \qquad \therefore \{\psi^*(-B)\} = \{\psi(B)\}. \tag{8.20}$$

This means $\psi(B)$ and $\psi^*(-B)$ forms the same set of solutions (here $\{\cdots\}$ means the set of $\cdots$). Remember that $\psi(B)$ is a scattering solution of Schrödinger equation (8.19). Let us express a scattering state as $\mathrm{Sc}\{\boldsymbol{a} \to \boldsymbol{b}\}$($\boldsymbol{a}$ is the incoming wave to the S-matrix, $\boldsymbol{b}$ is the scattered wave).

$$\mathrm{Sc}\{\boldsymbol{a}(B) \to \boldsymbol{b}(B)\} \in \{\psi(B)\}, \tag{8.21}$$

$$i.e., \quad \boldsymbol{b}(B) = S(B)\boldsymbol{a}(B). \tag{8.22}$$

We take the complex conjugate of (8.22) as

$$\boldsymbol{b}^*(B) = S^*(B)\boldsymbol{a}^*(B). \tag{8.23}$$

Now to take the complex conjugate of a propagating wave $\exp(\pm i\boldsymbol{kr})$ corresponds to the inversion of direction of propagation[*5]. That is, by taking the complex conjugate, the incoming wave and the scattered wave are exchanged.

$$\mathrm{Sc}(\boldsymbol{b}^*(B) \to \boldsymbol{a}^*(B)) \in \{\psi^*(B)\} \tag{8.24}$$

$$\therefore \quad B \to -B \text{ results in } \quad \mathrm{Sc}\{\boldsymbol{b}^*(-B) \to \boldsymbol{a}^*(-B)\} \in \{\psi^*(-B)\} = \{\psi(B)\} \tag{8.25}$$

$$i.e. \qquad \boldsymbol{a}^*(-B) = S(B)\boldsymbol{b}^*(-B). \tag{8.26}$$

From (8.26)

$$\boldsymbol{b}^*(B) = S^{-1}(-B)\boldsymbol{a}^*(B), \tag{8.27}$$

[*5] Schrödinger equation (8.19) does not depend on time, and then taking the complex conjugate of means the sign reversal of $i\boldsymbol{kr}$ keeping the sign of $i\omega t$.

and the comparison with (8.23) gives

$$S^*(B) = S^{-1}(-B) = S^\dagger(-B) \quad (\because \text{unitarity} \quad SS^\dagger = S^\dagger S = I)$$
$$\therefore S(B) = {}^tS(-B). \tag{8.28}$$

Q. E. D.

The following simple symmetric property is derived for the case of so far discussed **two-terminal transport**, in which the system only has single inlet and single outlet, and the resistance (ρ_{xx}) is defined as the ration of the voltage drop between the electrodes to the current.

$$\rho_{xx}(\boldsymbol{B}) = \rho_{xx}(-\boldsymbol{B}). \tag{8.29}$$

In the above proof, the linearity of the transport coefficients is assumed. Hence in non-linear devices, the reciprocity is broken under finite bias. Even for the non-linear devices, if the I-V characteristics is symmetric to the origin, the reciprocity recovers with including reversing the bias.

8.5 Landauer-Büttiker formula

So far we have treated coherent transport in two-terminal conductors. As in the S-matric scheme, experiments of coherent transport can be seen as a kind of scattering experiments. The terminals correspond to the detectors catching the scattered wave, and the number of terminals can be larger than two in general transport measurements. The scattering theory, which treats many terminals with equal footings, is the **Landauer-Büttiker** formalism.

Let us index the terminals with p, q (Fig. 8.6). Terminal p is connected to the particle reservor which has the chemical potential $\mu_p = -eV_p$. The net current J_p which flows from terminal p into the sample is obtained as follows. We consider the sum of the electron fluxes times $-e$ flowing into p from other terminals to p. And we subtract the sum from the electron flux times $-e$ flowing from p to the sample.

$$J_p = -\frac{2e}{h}\sum_q [T_{q\leftarrow p}\mu_p - T_{p\leftarrow q}\mu_q]. \tag{8.30}$$

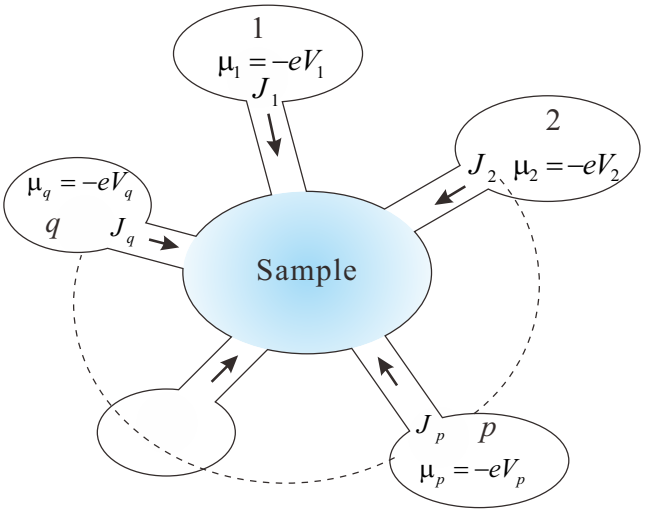


Fig. 8.6 Model to derive LB formalism.

First, the current conservation tells that

$$\sum_q J_q = 0. \tag{8.32}$$

With expressing $T_{p\leftarrow q}$ etc. in the form of matrix $\mathscr{T}$ as

$$\mathscr{T}_{pq} \equiv T_{p\leftarrow q} \quad (p \neq q), \qquad \mathscr{T}_{pp} \equiv -\sum_{q\neq p} T_{q\leftarrow p},$$

and with writing $\boldsymbol{J} = {}^t(J_1, J_2, \cdots)$, $\boldsymbol{\mu} = {}^t(\mu_1, \mu_2, \cdots)$ (column vectors), we can express

$$\boldsymbol{J} = \frac{2e}{h}\mathscr{T}\boldsymbol{\mu}.$$

Also

$$V_q = \frac{\mu_q}{-e}, \qquad G_{pq} \equiv \frac{2e^2}{h}T_{p\leftarrow q} \quad \text{とおくと}$$

$$J_p = \sum_q [G_{qp}V_p - G_{pq}V_q]. \tag{8.31}$$

The above is the essence of Landauer-Büttiker formalism but it still needs some constraints as below.

Next, when all the terminals are at the same potential, the currents should be zero, i.e.

$$\sum_q [G_{qp} - G_{pq}] = 0. \tag{8.33}$$

Further, for the external magnetic field B, the Onsager reciprocity

$$G_{qp}(B) = G_{pq}(-B) \tag{8.34}$$

holds. This can be proven from the Onsager reciprocity of S-matrix. The above is **Landauer-Büttiker formalism** (LB formalism) of electron transport.

Let us apply the LB formalism to a sample with four terminals. We take the origin of energy so as for the fourth chemical potential to be zero, *i.e.* $\mu_4 = -eV_4 = 0$. Then we can write down

$$\begin{pmatrix} J_1 \\ J_2 \\ J_3 \end{pmatrix} = \begin{pmatrix} G_{12}+G_{13}+G_{14} & -G_{12} & -G_{13} \\ -G_{21} & G_{21}+G_{23}+G_{24} & -G_{23} \\ -G_{31} & -G_{32} & G_{31}+G_{32}+G_{34} \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix}. \tag{8.35}$$

Now we consider the boundary condition

$$J_1 = -J_3, \qquad J_2 = -J_4, \tag{8.36}$$

which is called Casimir problem. The problem is reduced to ordinary situation of four probe measurement with $J_2 = 0$ in that the current is flowing through 1 and 3 while the voltage between 2 and 4 is measured without current. With writing $V_{ij} \equiv V_i - V_j$, the solution of this problem is given as

$$\begin{pmatrix} J_1 \\ J_2 \end{pmatrix} = \begin{pmatrix} \alpha_{11} & -\alpha_{12} \\ -\alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} V_{13} \\ V_{24} \end{pmatrix}, \tag{8.37}$$

where

$$\alpha_{11} = 2G_q[-\mathscr{T}_{11} - S^{-1}(\mathscr{T}_{14}+\mathscr{T}_{12})(\mathscr{T}_{41}+\mathscr{T}_{21})], \tag{8.38a}$$
$$\alpha_{12} = 2G_q S^{-1}(\mathscr{T}_{12}\mathscr{T}_{34} - \mathscr{T}_{14}\mathscr{T}_{32}), \tag{8.38b}$$
$$\alpha_{21} = 2G_q S^{-1}(\mathscr{T}_{21}\mathscr{T}_{43} - \mathscr{T}_{23}\mathscr{T}_{41}), \tag{8.38c}$$
$$\alpha_{22} = 2G_q[-\mathscr{T}_{22} - S^{-1}(\mathscr{T}_{21}-\mathscr{T}_{23})(\mathscr{T}_{32}+\mathscr{T}_{12})], \tag{8.38d}$$
$$S = \mathscr{T}_{12}+\mathscr{T}_{14}+\mathscr{T}_{32}+\mathscr{T}_{34} = \mathscr{T}_{21}+\mathscr{T}_{41}+\mathscr{T}_{23}+\mathscr{T}_{43}. \tag{8.39}$$

In Eq. (8.37), the current is expressed with the voltages, but in real experiments often the current is given by the external circuit and the voltages (chemical potentials) $V_1 \sim V_3$ are rearranged to fulfill the condition (8.36).

The reciprocity (8.34) gives the constraint

$$\alpha_{11}(B) = \alpha_{11}(-B), \qquad \alpha_{22}(B) = \alpha_{22}(-B), \qquad \alpha_{12}(B) = \alpha_{21}(-B) \tag{8.40}$$

to the solution (8.37). We apply this to ordinary four-termial problem and assign 1 and 3 to the current probes, 2 and 4 to the voltage probes and write the resistance obtained from LB formalism as $\mathcal{R}_{13,24}$. Then we see

$$\mathcal{R}_{13,24} = \frac{V_2 - V_4}{J_1} = \frac{\alpha_{21}}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}, \tag{8.41}$$

which does not show the symmetry to the magnetic field inversion like (8.29) though each matric element fulfills the Onsager reciprocity. On the other hand, the resistance measured with current-voltage exchanged terminals is

$$\mathcal{R}_{24,13} = \frac{\alpha_{12}}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}, \tag{8.42}$$

which is, from Eq. (8.40) symmetric to the reversal of magnetic field.

Generally from

$$\mathcal{R}_{mn,kl} = R_q \frac{\mathscr{T}_{km}\mathscr{T}_{ln} - \mathscr{T}_{kn}\mathscr{T}_{lm}}{D}, \qquad D \equiv R_q^2(\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21})S, \tag{8.43}$$

the reciprocity

$$\mathcal{R}_{mn,kl}(B) = -\mathcal{R}_{kl,mn}(-B) \tag{8.44}$$

holds. The minus sign is just due to the order of terminals.

The above results bring intresting information in measuring magnetoresistances of four terminal samples in quantum coherence. That is, generally in four terminal measurement, the magnetoresistance is not symmetric to $B = 0$ ($\rho_{4\mathrm{t}}(B) \neq \rho_{4\mathrm{t}}(-B)$. However, if we exchange the set of voltage probes and that of current probes and reverse the field direction $B \to -B$, then the resistance is unchanged.

References

[1] W. N. Can, IEEE Trans. Electron Dev., ED-**12**, 531 (1965).

[2] Y. Narukawa *et al.*, Jpn. J. Appl. Phys. **46**, L963 (2007).

[3] 本格的に学ぶには例えば，R. Loudon, "The Quanutm Theory of Light" (3rd ed., Oxford, 2000); P. Meystre and M. Sargent III, "Elements of Quantum Optics" (Springer, 1990); 松岡正浩 「量子光学」 (裳華房, 2000) など.

[4] 早川尚男 「非平衡統計力学」 (サイエンス社，2007).

[5] S. Datta, "ElectronTransport in Mesoscopic Systems" (Cambridge Univ. Press, 1995).

[6] 勝本信吾 「メゾスコピック系」 (朝倉書店，2002)

8.5.4 Aharonov-Bohm ring

As an application of S-matrix, we consider the transmission coefficient of an Aharonov-Bohm (AB) ring. The channel configuration is shown in Fig. 8.19(a). We write the S-matrix for the two junctoins with three channels as [*1]

$$S_t = \begin{pmatrix} 0 & -1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & 1/2 & -1/2 \\ -1/\sqrt{2} & -1/2 & 1/2 \end{pmatrix}. \tag{8.51}$$

The AB phase is taken into account by inserting

$$S_{\mathrm{AB}} = \begin{pmatrix} 0 & e^{i\theta_{\mathrm{AB}}} \\ e^{-i\theta_{\mathrm{AB}}} & 0 \end{pmatrix}, \qquad \theta \equiv 2\pi\frac{\phi}{\phi_0} = \frac{e}{\hbar}\phi \ \ (\phi \text{ is the flux piercing the ring}) \tag{8.52}$$

into one of the parallel paths. We insert the S-matrix

$$S_{\mathrm{w}} = \begin{pmatrix} 0 & e^{i\theta_0} \\ e^{i\theta_0} & 0 \end{pmatrix} \tag{8.53}$$

into the counter arm to express the phase difference between the two paths. The phase shift θ_0 from the path difference does not depend on the direction of the propagation while the sign of θ_{AB} is inverted with inversion of propagation. The Onsager reciprocity (8.18) is kept with these mathematical settings.

From the total S-matrix, the complex transmission coefficient of the ring is obtained as[3]

$$t = \frac{4\sin\theta_0}{1 + e^{i\theta_{\mathrm{AB}}}\left(e^{i\theta_{\mathrm{AB}}} + e^{i\theta_0} - 3e^{-i\theta_0}\right)}. \tag{8.54}$$

The transmission coefficient $T = |t|^2$ shows AB oscillation of the period ϕ_0 in ϕ (the magnetic flux piercing the ring). T also oscillates versus θ_0 with the period of 2π. $|t|^2$ is symmatric for $\phi = 0$, which is due to the reciprocity induced on (8.54) by the Onsager reciprocity introduced into S-matrix (8.52).

The phase of the oscillation with period ϕ_0 varies on θ_0 as a rectangular wave with amplitude π. The amplitude of the oscillation disappears around the phase jumps, which does not mean the disappearance of the magnetoresistance oscillation and the $\phi_0/2$ period and higher frequency components increase in the amplitudes. As above, the ϕ_0-oscillation only takes the phase offset of 0 or π, which property is called "phase rigidity[4]." The phase rigidity means that we cannot detect the phase shift over the quantum dot inserted into one of the arms of an AB ring[*2] .

8.6 Qauntum transport and particle statistics

As a transport phenomenon in a semiconductor, the transport of electrons (charge and spin) is often considered, but some quasiparticles baring transport behave differently from electrons. We will have a look how we apply the quantum transport theory (or not apply).

[*1] This form is frequently adopted though it is completely symmetric and a bit special in that sense.

[*2] If we consider multiple conductance channels and also restrict the region of magnetic field, the phase looks smoothly changes with flux[5] though this does not mean the breaking of the Onsager reciprocity.

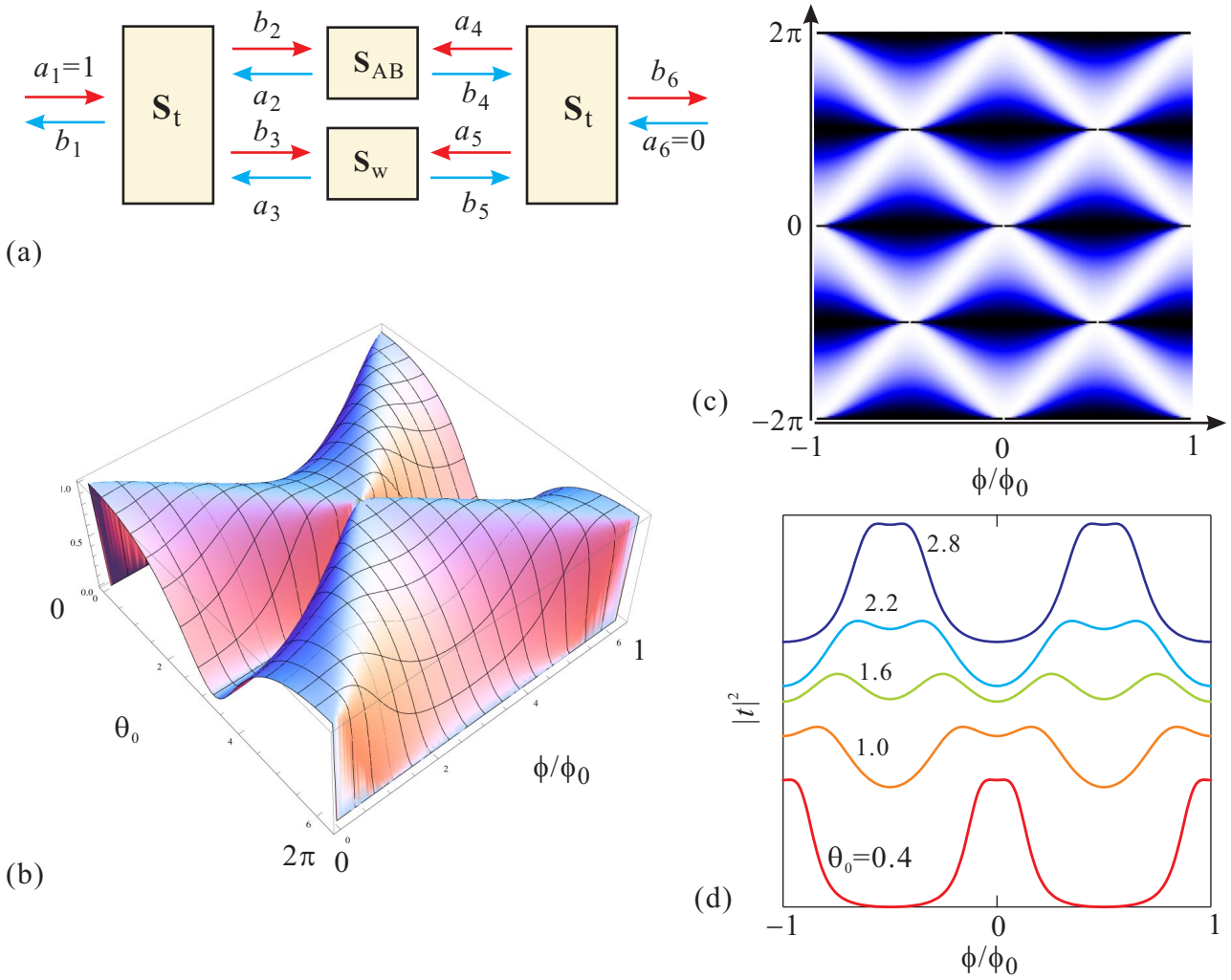


Fig. 8.19 (a) S-matrix modeling of an AB ring. (b) The transmission coefficient of the AB ring $|t|^2$ is plotted (surface plot) as a function of the phase shift from the path difference (θ_0) and the magnetic flux piercing the ring ϕ/ϕ_0. (c) Color plot of the same calculation over a bit wider region. (d) The same transmission coefficient as a function of ϕ/ϕ_0 with θ_0 as a parameter. The AB oscillation of period ϕ_0 once disappears around $\theta_0 = 1.6$ and then the reverted oscillation, that is, with π-shift in the phase appears.

8.6.1 Bunching, anti-bunching

In the previous section, we have introduced the Landauer formula to treat the electric conduction in semiconductor quantum structures. In the discussion, we assumed the transport of electrons and used the unit charge and the Fermi distribution in the derivation. And the conductance quantization is derived from the anti-bunching of fermions on quantum wires. On the other hand, in order to calculate the transmission coefficient T_{ij}, we have introduced T-matrices and S-matrices scheme. These are only to calculate the transmission and reflection of waves without the relation to the particle statistics. Hence the method should be applicable regardless of particle statistics.

Let us have a look on bunching and anti-bunching properties. The wavefunction for identical two particls is in real coordinate representation as

$$\psi(\boldsymbol{r}_1, \boldsymbol{r}_2) = \frac{1}{\sqrt{2}}[\phi_1(\boldsymbol{r}_1)\phi_2(\boldsymbol{r}_2) \pm \phi_1(\boldsymbol{r}_2)\phi_2(\boldsymbol{r}_1)]. \tag{8.55}$$

In the double sign, $+$ corresponds to bosons, and $-$ to fermions.

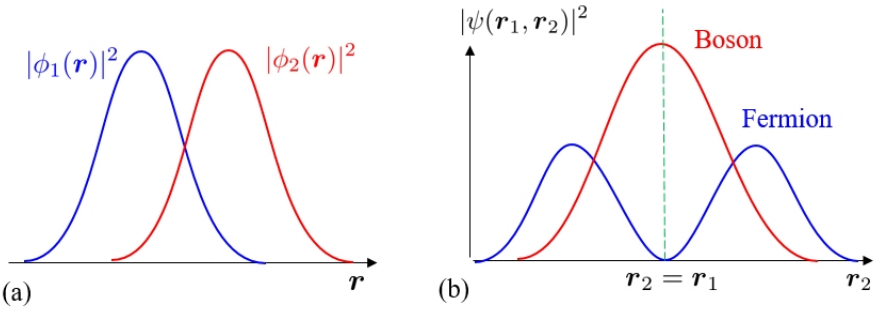


Fig. 8.20 (a) Schematic probability densities in coordinate representation $|\phi_1(\boldsymbol{r})|^2$, $|\phi_2(\boldsymbol{r})|^2$ of one-particle wavefunctions. The two wavefunctions have a partial overlap. (b) In the case of (a), the spatial probability density of particle 2 is plotted as a function of coordinate $\boldsymbol{r}_2$ with taking the position of the particle 1 $\boldsymbol{r}_1$ as the origin in the two-particle wavefunction ϕ in (8.55).

As shown in Fig. 8.20(a), we consider gaussian shaped wavefunctions $\phi_{1,2}(\boldsymbol{r})$ with partial overlap. Figure 8.20(b) shows the probability density of the two-particle wavefunction versus the relative position. As we can easily see by putting $\boldsymbol{r}_2 = \boldsymbol{r}_1$, the probability densities are

$$|\psi(\boldsymbol{r}_1,\boldsymbol{r}_1)|^2 = \begin{cases} 2|\phi_1(\boldsymbol{r}_1)|^2|\phi_2(\boldsymbol{r}_1)|^2 & \text{(boson)}, \\ 0 & \text{(fermion)}. \end{cases} \tag{8.56}$$

That is, in the case of bosons, the probability density is twice the case of single particle while in the case of fermions the density is zero. This means that the **bunching** occurs for bosons while the **anti-bunching** occurs for fermions.

The discussion of (8.3)∼(8.5), which leads to the Landauer formula can be understood in this context of anti-bunching. That is, the difference in the chemical potentials of two electron reservoirs, eV is the energy window to be used to form the wave packets of elctrons *i.e.*, $\Delta E \sim eV$. The time for such a wave packet to go through a point in the real space is, from the uncertainty relation, $\Delta t = h/\Delta E = h/eV$. From the anti-bunching property or Fermi statistics, a single wave packet can accomodate a single elecrtron (if the spin degree of freedom is taken into account, two electrons). Then the current flowing though the one-dimensional system is $J = e/\Delta t = (e^2/h)V$ (with spin freedom, $(2e^2/h)V$), which results in the same conclustion in the previous section.

The above discussion is the same calculation of that done in the k-space though it gives an important insight of the flow of electrons on quantum wires. When the conductance is quantized in such a quantum wire, the electrons flow with the time interval of h/eV. As is describe in the shot noise section of Appendix 8A, the current flow is approximated as a periodic series of delta-functions and the shot noise disappears. In the Landauer's discussion, the conductance quantization is the consequence of the fermion's anti-bunching property and the above consideration means that can be experimentally confirmed through shot noise measurement.

We expand the above to quantum wires with transmission coefficients $\mathcal{T}$ less than 1 to get $G = \mathcal{T}G_{\rm q}$. In this case, there appear free spaces between the wave packet due to the electron reflection and the packing of electrons becomes stocastic to some degree. This states can be viewed as follows. A perfectly ordered series of wave packets are occupied by electrons with probability $\mathcal{T}$, by holes with probability $1-\mathcal{T}$. Identical electrons (also holes) cannot be distinguished and the number of cases for vacancies, *i.e.* the degree of randomness is proportional to $\mathcal{T}(1-\mathcal{T})$. In the limit $\mathcal{T} \to 0$, this goes to Eq. (8A.3) and with using the relation $J = 2\mathcal{T}G_{\rm q}V$ for voltage V, the noise power spectrum is given as[**?**]

$$S \equiv \frac{\langle(\delta J)^2\rangle}{\Delta f} = 2e\frac{2e^2}{h}V\mathcal{T}(1-\mathcal{T}). \tag{8.57}$$

The above S is suppressed from $S_{\rm Poisson}$ in Eq. (8A.3) by factor $1-\mathcal{T}$. Generally, we refer to **Fano factor** as the ratio of variance to the average. In the present case that corresponds to the ratio of the shot noise to the Poisson noise and is $1-\mathcal{T}$.

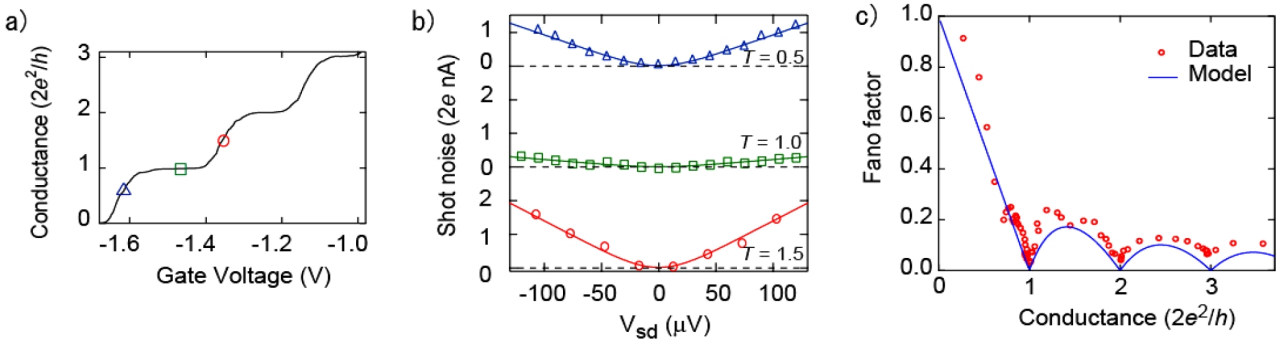


Fig. 8.21 Shot noise measured on a quantized conductance plateau of a QPC, and at around transition regions. (a) QPC conductance as a function of the gate voltage. Positions for noise measurements are indicated by color-framed open symbols. (b) Shoit noise measured at the three points indicated in (a) as a function of the source-drain voltage. (c) Fano factor (red circles) as a function of the conductance of the QPC. The blue line is obtained from a simple model, in which Fano factor should be $1 - \mathcal{T}$. From [6].

Figure 8.21 shows an example of shot noise measurement of a QPC. In panel (a), the conductance is shown as a function of the gate voltage and the current noise (spectral density of the square of current fluctuation) were measured at the three points indicated by color-framed symbols (on a conductance plateau and at neighboring transition regions). In panel (b), the current noise is plotted versus the source-drain voltage. In the transient regions, the noise increases with the voltage indicating the appearance of shot noise while around the center of plateau, the increase of noise is very small, indicating the reduction of noise. In panel (c), the noise data are converted to Fano factor and shown as a function of the conductance. The blue line shows the consequence of a simple model, in which the Fano factor should be $1 - \mathcal{T}$. The data distribute a bit above the model line confirming the noise reduction with the conductance quantization.

8.6.2 Transport of exciton-polaritons

As a bosonic quasiparticle, we consider exciton-polariton (E-P), which is introduced in Sec. 4.4.2. An E-P is a composite of photons and excitons created as a result of strong coupling of light and matter. Being pairs of fermions, excitons obey Bose statistics but the effective mass is the sum of those for electrons and holes as $m_e + m_h$. On the other hand, as can be seen from the dispersion relation in Fig. 4.7, an E-P has a very small effective mass around $k \approx 0$. This makes the control of phase of E-Ps easier and the reseachers are trying to apply E-Ps for optical integrated circuits. The light effective mass makes the critical temperature of BEC (8B.10) very high. Actually, the observations of BEC have been reported.

8.6.2.1 Cavity exciton-polariton

In the section of laser diode, we have seen a structure of two-dimensional cavity. In Fig. 8.22(a) we show a transmission line made by cutting the two-dimensional cavity into a thin mesa structure. Here, the structure is such that GaAs is used as a quantum well, which is sandwiched between GaAs / AlGaAs superlattices (SL), and a GaAs clad layer is placed on the outside. The effective refractive index of the SL part is lower than that of GaAs, and photons are confined in this region. On the other hand, excitons are confined in the central GaAs quantum well because the SL regions work as barriers due to the band discontinuity. The excitons in this case are confined in the two-dimensional plane of the cavity and as we saw in Sec. 7.1.3 the binding energy of the excitons becomes larger and they are stabilized. With the above devicing, E-Ps can propagate the waveguide but for stable propagation, we need to prepare some low temperature environments. The limit of temperature is extimated from the gap between the upper and the lower branches in the dispersion relation in Fig. 8.22(c) and that can be go over the liquid nitrogen temperature.

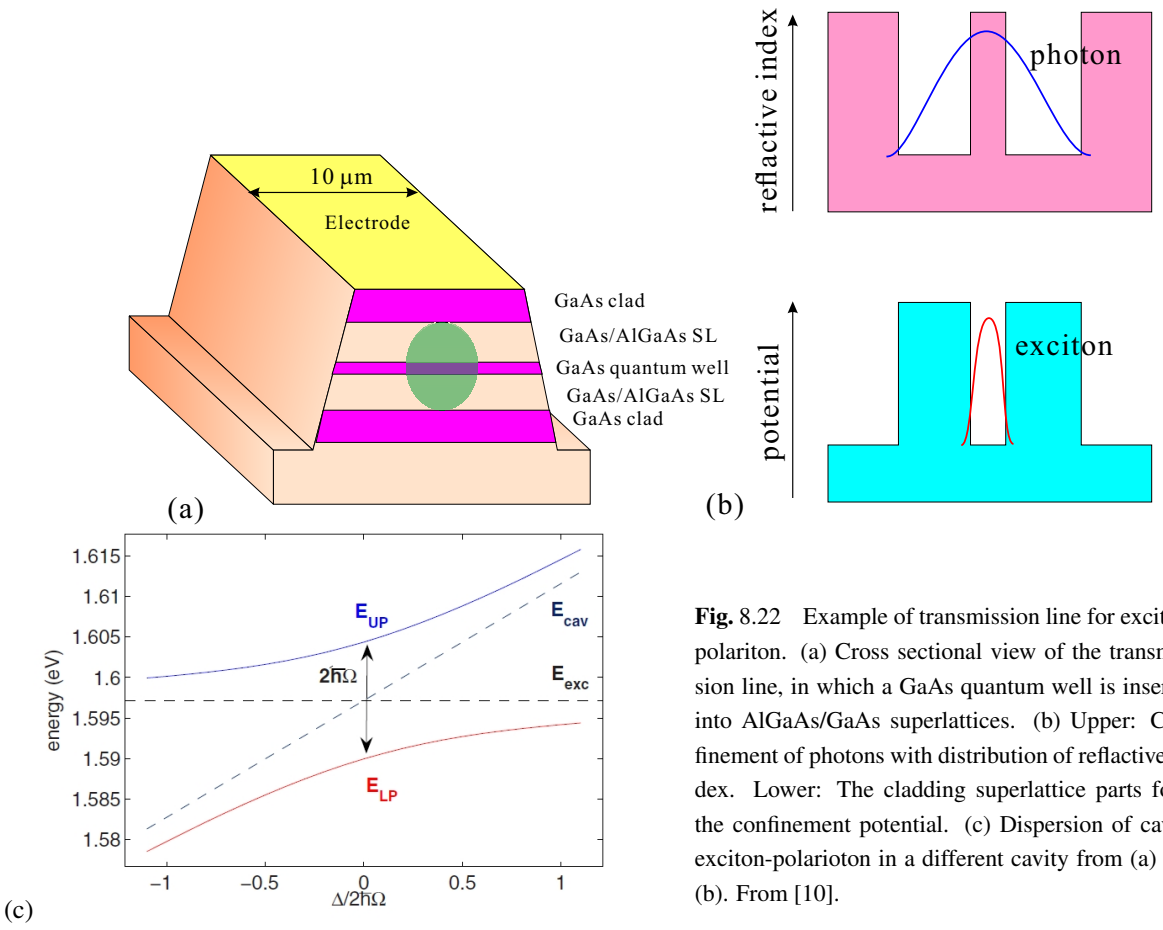


Fig. 8.22 Example of transmission line for exciton-polariton. (a) Cross sectional view of the transmission line, in which a GaAs quantum well is inserted into AlGaAs/GaAs superlattices. (b) Upper: Confinement of photons with distribution of reflactive index. Lower: The cladding superlattice parts form the confinement potential. (c) Dispersion of cavity exciton-polarioton in a different cavity from (a) and (b). From [10].

The dispersion relation in Fig. 4.7 is made up from photons in crystals and excitons. In the present E-P waveguide, the photons are strongly confined into the micro-cavity and the photon dispersion relation changed from that in bulk. Figure 8.22(c) shows the avoided crossing between the cavity photon and the exciton caused by the strong coupling of light and matter. The structure of cavity here for the calculation of the dispersion is shown in Fig. 8.24.

In the left panel of Fig. 8.23, a conceptual figure of a Mach-Zehnder (MZ)-type interference device composed of the cavity transmission line shown in Fig. 8.22(a). As mentioned in the section of exciton, because the response to the electric field is opposite for electrons and holes, if is is a comletely single-body composite particle, the effect of the electric field can be hardly observied. However, in the structure of Fig. 8.22(a), the binding energy of excitons can be varied by the electric field and with that, the wavenumber varies as

$$\Delta\varphi = L\left[\frac{\sqrt{2mE_k}}{\hbar} - \frac{\sqrt{2m(E_k - \delta E)}}{\hbar}\right]. \tag{8.58}$$

δE represents this variation in the kinetic energy and L is the length of the gate region. This gives modulation in the output of the two-path circuit shown in the left panel of Fig. 8.23. Finally the output is transposed into light at the edge of the transmission line and the output can be detected as the light strength. The transmission circuit in Fig. 8.23 is called in the paper[7], as an MZ interferometer though, because it has a single output line, some reflection exists at the joint, it should be called as a two-path AB-type (in the present case, the AB phase does not exist and "ring-type" may be a better expression). As shown in Fig. 8.23(a), (b), with the voltage the light ouput power can be controlled by over 10 dB and voltage-light switching function is realized.

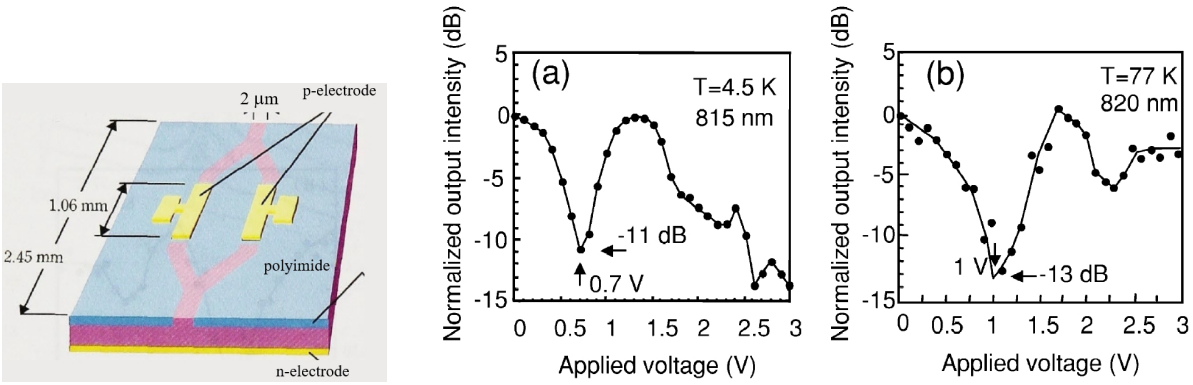


Fig. 8.23 Left panel: Concept of Mach-Zehnder (MZ)-type interference device. (a) Variation in the ouput of MZ interference device versus gate voltage (4.5 K). (b) The same for 77 K.

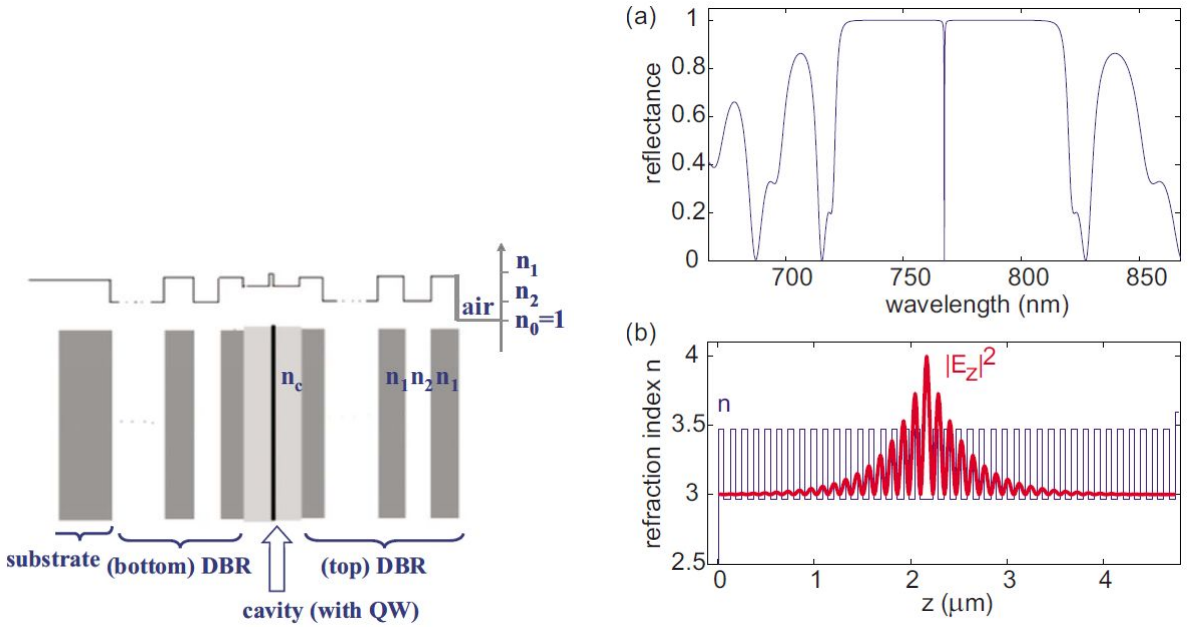


Fig. 8.24 Left panel: Schematic view of the cavity transmission line for the observation of exciton-polariton condensation. Upper figure shows the refractive index. Right panel: (a) Calculated reflection coefficient of hte left cavity. (b) Structure of reflactive index in the cavity. The red line is the distribution of electric field along z-axis for the localized mode showing the sharp resonant dip around the center of the spectrum shown in (a). From [10].

8.6.2.2 Condensation of exciton-polariton

While electrons are the representative of fermions flowing through quantum circuits, as we have seen above, E-Ps in microcavities is a system, with which we can explore the boson flow through quantum circuits experimentally. The consequence of Fermi statistics on fermion flow in quantum circuits is the conductance quantization and the reduction of shot noise. On the other hand, bunching of the identical particles is the characteristics of the Bose statistics, as we have seen in Sec. 8.6.1. As a result, **Bose-Einstein condensation** (BEC) or similar phenomenon with condensation occurs. The stimulated emission is also a phenomenon similar to the boson bunching, and photons in a cavity of a laser can be viewed as a kind of condensation though the lasing occurs in non-equilibrium open systems while BEC is a phenomenon in equilibrium. There is thus a clear difference[8, 9].

Figure 8.24 shows the diagram of refractive index in the cavity prepared for the observation of BEC. This figure also

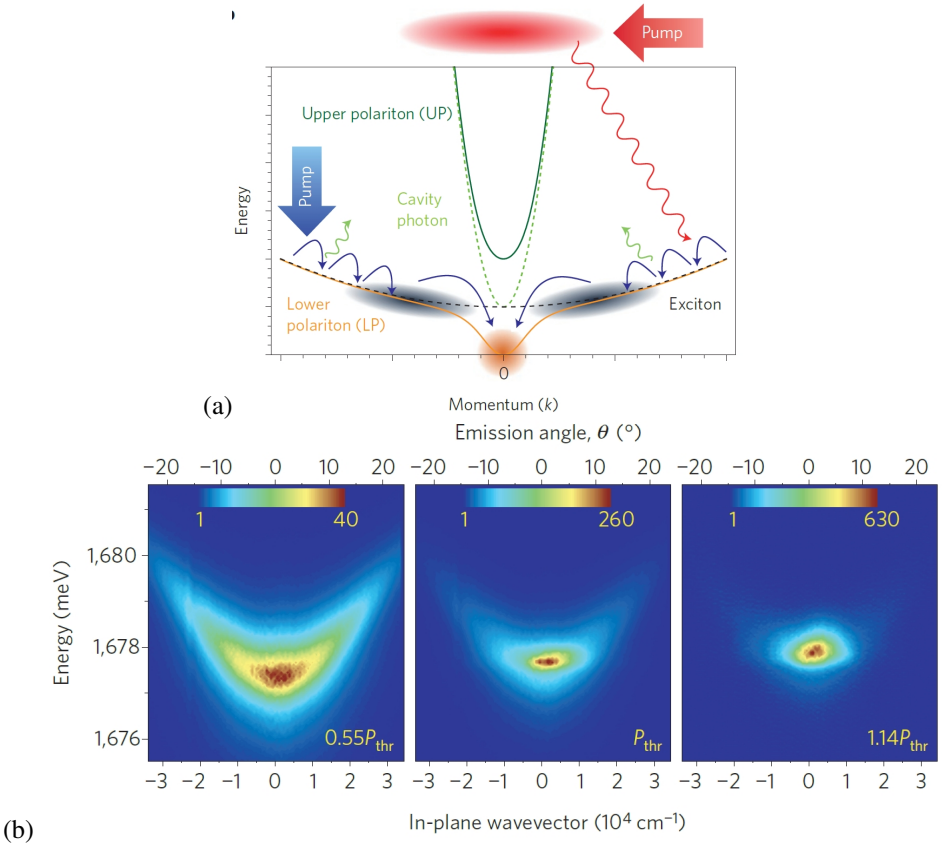


Fig. 8.25 (a) Conceptial diagram illustrating the "cooling" process for micro-cavity E-P to cause BEC. (b) Ditribution of E-Ps in the space of wavenumber and the energy measured from the optical emission. $P_{\rm th}$ is the critical excitation power to create E-Ps with the critical particle density for BEC. from [9].

shows the energy density of electric field along z-axis in the mode localized at the center of cavity, which is calculated with T-matrix method. T-matrice and S-matrices are thus used to calculate various quantities related to wave propagations.

A short summary on BEC of three dimensional ideal boson gas is given in Appendix 8B. As the expression of the critical temperature $T_{\rm c}$ in Eq. (8B.10), the lighter the effective mass is, and also the higher the particle density is, the higher $T_{\rm c}$ becomes. Conversely, when the mass and the temperature are given, the critical particle density for BEC to appear is defined.

Figure 8.25(a) illustrates the process of creation of a BEC with laser light irradiation on the cavity system. In the beginning of the process, many E-Ps with high energies are excited with the lase pulse. They emit energies as phonons to the crystal and are cooled down. If the laser power is higher than the critical value $P_{\rm th}$, as cooling, a BEC is created and a macroscopic number of E-Ps fall into the lowest energy state. In Fig. 8.25(b), the wavenumber-energy distributions of E-Ps measured from light leakage, are given around the BEC critical power.

Here we need to be careful about the meaning of "BEC." The present E-P system is composed of modes confined to a 2-dimensional plane and 3-dimensional BEC in App. 8B cannot be directly applied. In the space with dimension lower than or equal to 2, no infinitely long range order does not exist as mentioned in Mermin-Wagner theorem[11]. Instead, Berezinskii-Kosterlitz-Thouless (BKT) transition occurs and the order decays with some power of the distance[12]. Actually, the existence of BKT transition was evidenced in detailed analysis of experiments. And the observation of vortex-pair is announced. There are so many reports on the BEC of E-P systems and the research is active both in theory and experiment.

Appendix 7A: Laser diode and waveguide

Here a short supplement on the structure of waveguide for laser diode (LD) is given. Let us consider the Fabry-Pérot type LD with waveguide (cavity) length L. Let m_j, $\bar{n}$ and λ be an integer, the reflactive index, and the wavelength in the vacuum respectively, then the condition of resonance is

$$m_j \frac{\lambda}{\bar{n}} = 2L. \tag{7A.1}$$

Therefore the interval in the resonant wavelengths and that in the resonant frequencies are

$$\Delta\lambda = \frac{\lambda^2}{2L\bar{n}}, \quad \Delta\nu = \frac{c}{2L\bar{n}}, \tag{7A.2}$$

respectively. In usual systems, $\lambda \ll L$. When the amount of carrier injection is large and the luminescence is broad in wavelength, precise determination of l is not requied monstly and multi-mode oscillation around a center wavelenth is observed.

In the above, we write the light intensity simply as $I_0 \exp(-\alpha' z)$. Then α' can be expanded as $I(z) = I_0 \exp((g-\alpha)z)$, where g is the optical gain, α is the material specific absorption coefficient. Let us write the reflection ratio of the two mirrors as R_1 and R_2 respectivly, then the condition for the amplification to occur is

$$R_1 R_2 \exp[(g-\alpha)2L] > 1.$$

Thus the threshold optical gain $g_{\rm th}$ for the total amplification is

$$g_{\rm th} = \alpha + \frac{1}{L} \ln\left(\frac{1}{R_1 R_2}\right). \tag{7A.3}$$

The reflactive index $\bar{n}_1$ is common for both sides of the homo pn-junction, while the reflactive index in the active layer $\bar{n}_2$ is larger than $\bar{n}_1$. z-axis is taken as in the figure and we consider the electromagnetic wave propagate along the z-axis. The propagation mode is transverse electric (TE), i.e., the electric field along z axis is absent ($\mathcal{E}_z = 0$). Also the mode is assumed to be uniform in y direction. Thus we only need to consider the elecric field in y direction, which is determined from

$$\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2} - \mu_0 \epsilon_0 \epsilon \frac{\partial^2}{\partial t^2}\right] \mathcal{E}_y = 0. \tag{7A.4}$$

The change of magnetic permeability in semiconductors from the vacuum is little then we use μ_0 here.

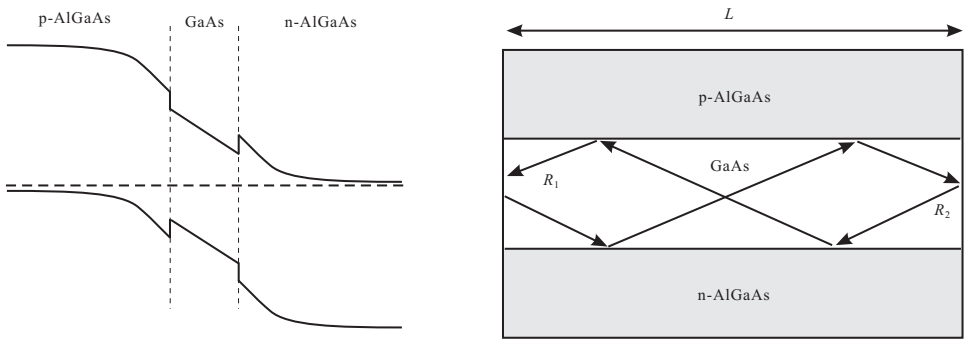


Fig. 7A.1 Left panel: Schematic band diagram of a pn-junction for LD. The active layer is non-doped GaAs and the doping layers with larger band gap than that of the active layer, are AlGaAs. Right panel: Substrate edges are formed by cleaving and work as half mirrors, which make the active GaAs layer into a cavity.

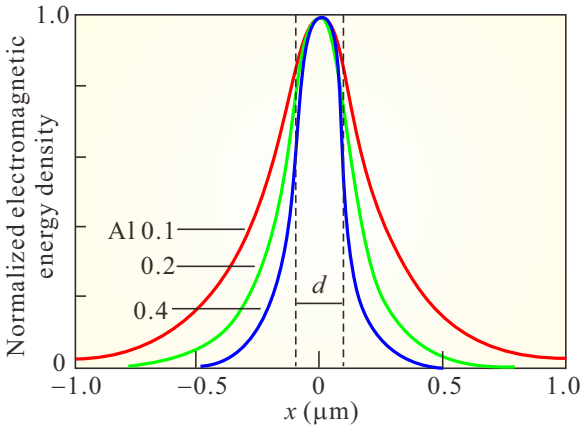


Fig. 7A.2 Distribution of normalized electromagnetic energy density of fundamental mode on x (origin at the center of active layer) in a cavity with an active (i-)layer of GaAs, The parameter is the content of Al.

We treat the system as a waveguide and look for a solution of standing wave on x and propagating wave on z. Then we find a solution inside the active layer ($|x| \leq d/2$) as

$$\mathcal{E}_y(x,z,t) = A\cos(\kappa x)\exp[i(\omega t - \beta z)], \tag{7A.5}$$

where

$$\kappa^2 = \mu_0\epsilon\epsilon_0\omega^2 - \beta^2 = \bar{n}_2^2 k_0^2 - \beta^2, \quad k_0 = \frac{\omega}{c^*\bar{n}_2}. \tag{7A.6}$$

And outside the active layer ($|x| > d/2$), the solution should decay with $|x| \to \infty$. The condition of connection at $x = \pm d/2$ での $\mathcal{E}_y$ restricts the solution to

$$\mathcal{E}_y(x,z,t) = A\cos\left(\frac{\kappa d}{2}\right)\exp\left[-\gamma\left(|x| - \frac{d}{2}\right)\right]\exp[i(\omega t - \beta z)]. \tag{7A.7}$$

Equation (7A.4) requires

$$\gamma^2 = \beta^2 - \bar{n}_1^2 k_0^2. \tag{7A.8}$$

From Maxwell equation, the continuity in z-component of magnetic field at $x = \pm d/2$ requires

$$\tan\left(\frac{\kappa d}{2}\right) = \frac{\gamma}{\kappa} = \frac{\sqrt{\beta^2 - \bar{n}_1^2 k_0^2}}{\sqrt{\bar{n}_2^2 k_0^2 - \beta^2}}. \tag{7A.9}$$

The values of κ, γ and β are determined from the above equations. Because tangent is a π-periodic function, there are multiple solutions, each of which forms a descrete mode.

Appendix 8A: Shot noise

We express information as a time-varying physical quantity and transmit it using various transport phenomena. Every physical quantity has fluctuations and among them time varing ones are called **noise** [*3]. The noise can be classified into external noise and **intrinsic noise**. While the former comes from "outside" of the system, the latter is included in the physics of the quantity itself. Particularly in the case of electric current by electron flow, the representative intrinsic noises are thermal noise (Johnson-Nyquist noise) that caused by random thermal motion of electrons and **shot noise** that originates from the particle nature of electrons and the randomness in the flow.

Let us consider first the current by a single electron observed at time t_p, expressed as $J_p(t) = e\delta(t - t_p)$. From the Fourier expansion

$$J_p(t) = e\int_{-\infty}^{\infty}\exp[2\pi i f(t-t_p)]df = 2e\int_0^{\infty}\cos[2\pi f(t-t_p)]df, \tag{8A.1}$$

[*3] Fluctuation means the distribution of observed values in multiple identical measurements and the parameter of sampling is not restricted to time. An example of non-time dependent fluctuation is aperiodic conductance oscillations in disordered mesoscopic conductors(P. A. Lee and A. D. Stone, Phys. Rev. Lett. **55**, 1622 (1985)).

we see that the current itself has an amplitude of $2e$ independent of frequency. In the infinitesimal frequency width df at frequency f, we take the average $\langle \cdots \rangle$ over one period. Let us write the integrand in (8A.1) as j_p and the current flusctuation is $\delta J_p = \sqrt{\langle j_p^2 \rangle} df = \sqrt{2} e df$.

Next we consider an electric current by two electrons observed at t_p and t_q, $J_{pq} = e[\delta(t - t_p) + \delta(t - t_q)]$. In the Fourier transform of J_{pq}, there is a phase difference $\phi = f(t_q - t_p)$ between the two Fourier components from the two delta-functions. The phase difference appears in the square of Fourier transformed function as an interference term:

$$j_{pq}^2 = j_p^2 + j_q^2 + 2 j_p j_q \cos\phi. \tag{8A.2}$$

The interference terms, however, cancel out when we add up many such two-electron currents and take the average (represented as $\overline{\cdots}$ due to the randomness in $t_q - t_p$, i.e. $\overline{j_{pq}^2} = 2(\sqrt{2}e)^2$. A current by many electrons randomized on time is equivalent to this many sampling. Hence, let N be the time averaged number of flowing electrons then the averaged current is $J = eN$ and the current fluctuation over the bandwidth Δf is

$$\langle (\delta J)^2 \rangle / \Delta f (\equiv S_{\text{Poisson}}) = N \times 2e^2 = 2eJ. \tag{8A.3}$$

The square of current fluctuation is proportional to the average of current corresponds to the fact the variance of Poisson distribution is the average (the number of electrons per unit time N). This case of complete randomness is called **Poisson noise**.

On the other hand, when the electrons flow with a constant interval, there is no fluctuation(timing of sampling would result in shifts of e in counted charge, but this is not a random variation). This can be understood from Fourier analysis of the current. Let us write the regular series of delta function with interval as τ as $\delta_\tau(t)$. Because $\delta_\tau(t)$ is a τ-periodic function, the Fourier series expansion on the region $[-\pi/\tau, \pi/\tau]$ is possible as follows.

$$\delta_\tau(t) = \frac{1}{\tau} \sum_{n=-\infty}^{\infty} \exp\left(-in\frac{2\pi}{\tau}t\right). \tag{8A.4}$$

Then the Fourier transform is written as

$$\begin{aligned} \mathscr{F}\{\delta_\tau(t)\} = \int_{-\infty}^{\infty} \left[\frac{1}{\tau}\sum_{-\infty}^{\infty} e^{-in(2\pi/\tau)t}\right] e^{i\omega t} dt &= \frac{1}{\tau}\sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[i\left(\omega - n\frac{2\pi}{\tau}\right)t\right] dt \\ &= \frac{2\pi}{\tau}\sum_{n=-\infty}^{\infty} \delta\left(\omega - n\frac{2\pi}{\tau}\right) = \frac{2\pi}{\tau}\delta_{2\pi/\tau}(\omega), \end{aligned} \tag{8A.5}$$

that is, it is also a regular series in ω space and there is no continnum spectrum, which is the sign of random variation. This means the disappearance of shot noise.

Appendix 8B: Bose-Einstein condensation

The Bose-Einstein Condensation (BEC) [*4] is called a phase transitoin that is not due to the interaction between freedoms (quantum statistical phase transition). Though phase transitions caused by interaction beween some freedoms can be intuitively understood, there are different types of phase transitions, in which the transitions are caused as the results of competition between various factors. A representative is BEC.

In the case of bosonic systems, in spite of the absence of "force" betwen the particles, there exists the tendency for them to occupy the same quantum state originating from their statistical property. Let us see that for the case of two

[*4] The acronym of BEC is applied to both Bose-Einstein Condensation and Bose-Einstein Condensate. In actual use, the confution is not serious.

particles. We write a solution of the wave equation for two particles as $\psi(\boldsymbol{x}_1, \boldsymbol{x}_2)$. For the composition of wavefunctions of the system $\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2)$ that reflects the statistical property of bosons, the symmetrization of ψ results in

$$\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{\sqrt{2}} \left[\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) + \psi(\boldsymbol{x}_2, \boldsymbol{x}_1)\right]. \tag{8B.1}$$

Hence the probability of finding the system at $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is

$$|\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2)|^2 = \frac{1}{2} \left[|\psi(\boldsymbol{x}_1, \boldsymbol{x}_2)|^2 + |\psi(\boldsymbol{x}_2, \boldsymbol{x}_1)|^2 + \psi(\boldsymbol{x}_1, \boldsymbol{x}_2)^* \psi(\boldsymbol{x}_2, \boldsymbol{x}_1) + \psi(\boldsymbol{x}_1, \boldsymbol{x}_2) \psi(\boldsymbol{x}_2, \boldsymbol{x}_1)^*\right]. \tag{8B.2}$$

This reveals that the last two interference terms intensify the probability of finding the system under the condition of $\boldsymbol{x}_1 = \boldsymbol{x}_2$. Let us write the de Broglie wavelength as λ and the averaged distance between the particles as l. Then at low temperatures $\lambda \sim l$, this tendency of bosons makes many of them to occupy the state of $k = 0$, which behavior leads to BEC. The above discusstion is expressed as

$$E_k = \frac{p^2}{2M} = k_{\mathrm{B}} T,$$

$$\begin{aligned} &\Delta p \sim \sqrt{M k_{\mathrm{B}} T} \\ &\therefore \lambda = \frac{h}{\Delta p} \sim \frac{h}{\sqrt{M k_{\mathrm{B}} T}}. \end{aligned} \tag{8B.3}$$

λ elongates as $1/\sqrt{T}$ with lowering the temperature. And with growing of the overlapp between the single particle wavefunctions makes them undistinguishable and the symmetrization of the wavefunction cause the condensation to the ground state in the phase space $(\boldsymbol{r}, \boldsymbol{p})$. The phase transition to the condensate at a certain temperature is BEC.

8B.1 Bose-Einstein condensation of ideal gas

Let us consider spin 0 ideal Bose gas. For the Bose distribution

$$f(\epsilon) = \frac{1}{e^{(\epsilon - \mu)\beta} - 1} \qquad (\beta \equiv (k_{\mathrm{B}} T)^{-1}) \tag{8B.4}$$

we define the point of $\mu = 0$ as follows. At $T = 0$, from (8B.4) all the particles fall into the ground state, there we define

$$\mu(T = 0) = 0. \tag{8B.5}$$

At finite temperatures, let N be the number of particles in the system:

$$N = \sum_i f(\epsilon_i).$$

In the usual case we can write

$$N \to \int f(\epsilon) \mathcal{D}(\epsilon) d\epsilon. \tag{?}$$

Here the number of particle at the ground state N_0 should be

$$N_0 = \frac{1}{e^{-\mu\beta} - 1} \sim \frac{1}{-\mu\beta} = -\frac{k_{\mathrm{B}} T}{\mu} \to \mu \sim -\frac{k_{\mathrm{B}} T}{N_0}. \tag{8B.6}$$

If we calculate the particle distribution on this line, for three dimensional ideal gas

$$\epsilon(k) = \frac{\hbar^2 k^2}{2m} \quad \text{then} \quad \mathcal{D}(\epsilon) = \frac{m^{3/2} V}{\sqrt{2} \pi^2 \hbar^3} \sqrt{\epsilon}. \tag{8B.7}$$

Therefore

$$N = \frac{V m^{3/2}}{\sqrt{2} \pi^2 \hbar^3} \int_0^\infty \frac{\sqrt{\epsilon}}{e^{(\epsilon - \mu)\beta} - 1} d\epsilon = \frac{(m k_{\mathrm{B}} T)^{3/2}}{\sqrt{2} \pi^2 \hbar^3} V \int_0^\infty \frac{\sqrt{x}}{e^{x - \alpha} - 1} dx, \tag{8B.8}$$

where $x \equiv \epsilon\beta$ and $\alpha \equiv \mu\beta$. We write the definite integral term as $I(\alpha)$, then I is

$$I(0) = \int_0^\infty \frac{\sqrt{x}}{e^x - 1} dx = \frac{\sqrt{\pi}}{2} \zeta\left(\frac{3}{2}\right) \sim 2.6, \tag{8B.9}$$

which decreases with increasing of the absolute value of $\alpha < 0$. Then, in this logic, with $T \to 0$ the maximum number of N determined from (8B.8) goes to zeo. It is apparent that we have dropped something from the counting. That is, of course, the macroscopic number of particles fall into the ground state.

From Eq. (8B.8),

$$I(\alpha) = \frac{\sqrt{2}\pi^2\hbar^3}{(mk_{\mathrm{B}}T)^{3/2}} \frac{N}{V}.$$

When this excessds (8B.9) at low temperatures the anomaly (increase in the particle number at the ground state.) occurs. This critical temperature T_c is

$$T < T_c \equiv \frac{2\pi\hbar^2}{mk_{\mathrm{B}}} \left[\frac{N}{\zeta(3/2)V}\right]^{2/3}. \tag{8B.10}$$

Here $l \equiv (V/N)^{1/3}$ is the average distance between the particles and Eq. (8B.10) is interpreted as

$$l = \frac{h}{\zeta(3/2)\sqrt{2\pi m k_{\mathrm{B}} T_c}} \sim \lambda(T = T_c). \tag{8B.11}$$

This confirms the statement that the BEC takes place when the average de Broglie wavelength is comparable with the average particel distance.

Below T_c, we add the number of ground state particles N_0 to Eq. (8B.8):

$$N = \frac{Vm^{3/2}}{\sqrt{2}\pi^2\hbar^3} \int_0^\infty \frac{\sqrt{\epsilon}}{e^{(\epsilon-\mu)\beta} - 1} d\epsilon + N_0. \tag{8B.12}$$

From Eq. (8B.6), N_0 becomes a macroscopic number fro $T < T_c$, then $\mu = 0$. Therefore

$$N_0 = N - \frac{Vm^{3/2}}{\sqrt{2}\pi^2\hbar^3} \int_0^\infty \frac{\sqrt{\epsilon}}{e^{\epsilon\beta} - 1} d\epsilon = N\left[1 - \frac{V}{N}\frac{(mk_{\mathrm{B}}T)^{3/2}}{\sqrt{2}\pi^2\hbar^3} I(0)\right] = N\left[1 - \left(\frac{T}{T_c}\right)^{3/2}\right]. \tag{8B.13}$$

This is just like a spontaneous magnetization rapidly grows to finite values below the critical temperature in the ferromagnetic transition.

The total energy of the system for $T < T_c$ is calculated as

$$E = \frac{Vm^{3/2}}{\sqrt{2}\pi^2\hbar^3} \int_0^\infty \frac{\epsilon^{3/2}}{e^{\beta\epsilon} - 1} d\epsilon \tag{8B.14}$$

ここで $T < T_c$ では $\displaystyle\int_0^\infty \frac{x^{3/2}}{e^x - 1} dx = \frac{3\sqrt{\pi}}{4} \zeta\left(\frac{5}{2}\right)$ より

$$E = \frac{3}{2}\zeta\left(\frac{5}{2}\right)\left(\frac{m}{2\pi\hbar^2}\right)^{3/2} V(k_{\mathrm{B}}T)^{5/2}. \tag{8B.15}$$

Then the heat capacity at constant volume is calculated as

$$C_v = \frac{15}{4}\zeta\left(\frac{5}{2}\right)\left(\frac{m}{2\pi\hbar^2}\right)^{3/2} V k_{\mathrm{B}}^{5/2} T^{3/2}. \tag{8B.16}$$

C_v shows a cusp at T_c indicating that this is the phase transition.

8B.2 Bosonic stimulation

Here we have a look at **bosonic stimulation** for N particles, which is, though, essentially the same as what has been mentioned on the case of two particles in Sed. 8.6.1. As we have seen, the bosonic stimulation works as if it is a driving

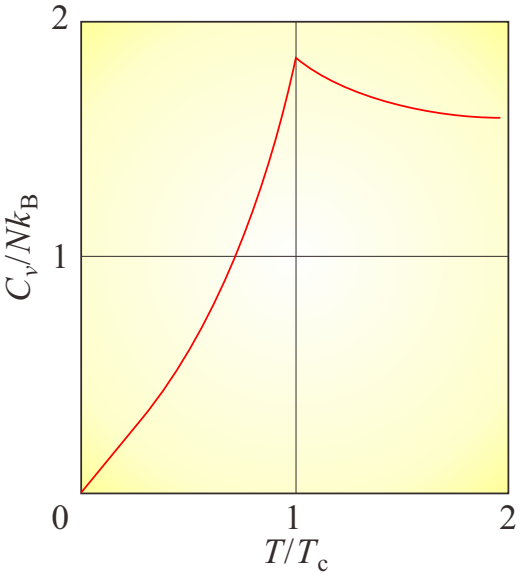


Fig. 8B.1 Specific heat at constant volume of three dimensional ideal Bose gas as a function of temperature. T_c is the critical temperature of the BEC.

force in BEC or laser oscillation. Let us consider a identical boson system the case a particle in state $\varphi_{\rm ini}$ gets perturbation and transitions to other single particel state $\varphi_{\rm fin}$. Now the problem is the difference in the transition probabilities to the state occupied with N particles and to the empty state. We write the initial state as

$$\psi_+^{(i)}(\boldsymbol{r}_1,\cdots,\boldsymbol{r}_{N+1}) = \frac{1}{\sqrt{(N+1)N!\prod_l n_l!}} \prod_{m=1}^{N} \hat{R}_{m,N+1} \det{}^{(+)}\{\varphi_i(\boldsymbol{r}_j)\}\varphi_{\rm ini}(\boldsymbol{r}_{N+1}). \tag{8B.17}$$

The symbol $\det^{(+)}$ represents permanent, which is obtained by making the signs of all the terms into $+$. The final state $\psi_+^{(f)}$ is obtaned by exchanging $\varphi_{\rm ini}$ with $\varphi_{\rm fin}$. Let the matrix elements of perturbation Hamiltonian be a, i.e. $\langle\varphi_{\rm fin}|\hat{H}_1|\varphi_{\rm ini}\rangle = a$.

Assuming that φ_i $(i \le N)$ is orthogonal to $\varphi_{\rm fin}$, among $\langle\psi_+^{(f)}|\hat{H}_1|\psi_+^{(i)}\rangle$, number of terms that give non-zero a is $(N+1)N!\prod_l n_l!$. This is equal to the sqare of the denominator in normalization constant. Then finally $\langle\psi_+^{(f)}|\hat{H}_1|\psi_+^{(i)}\rangle = a$.

On the other hand, assuming all of φ_i $(i \le N)$ are $\varphi_{\rm fin}$, we can write

$$\psi_+^{(i)} = \frac{1}{\sqrt{(N+1)}} \prod_{m=1}^{N} \hat{R}_{m,N+1}\varphi_{\rm fin}(\boldsymbol{r}_1)\cdots\varphi_{\rm fin}(\boldsymbol{r}_N)\varphi_{\rm ini}(\boldsymbol{r}_{N+1}). \tag{8B.18}$$

All of the $N!$ terms in $\det^{(+)}$ are $\varphi_{\rm fin}(\boldsymbol{r}_1)\cdots\varphi_{\rm fin}(\boldsymbol{r}_N)$ and devided by $N!$ in the denominator of normalization constant to 1. However the final state is

$$\psi_+^{(f)} = \varphi_{\rm fin}(\boldsymbol{r}_1)\cdots\varphi_{\rm fin}(\boldsymbol{r}_N)\varphi_{\rm fin}(\boldsymbol{r}_{N+1}). \tag{8B.19}$$

Then we get $\langle\varphi_{\rm fin}|\hat{H}_1|\varphi_{\rm ini}\rangle = a\sqrt{N+1}$, and from the Fermi's golden rule, the transition probability should be $N+1$ times larger.

References

[1] S. Datta, "ElectronTransport in Mesoscopic Systems" (Cambridge Univ. Press, 1995).

[2] 勝本信吾 「メゾスコピック系」 (朝倉書店，2002)

[3] Y. Gefen, Y. Imry, and M. Ya. Azbel, Phys. Rev. Lett. **52**, 129 (1984).

[4] A. Yacoby, R. Schuster, and M. Heiblum, Phys. Rev. B **53**, 9583 (1996).
[5] A. Aharony, O. Entin-Wohlman, T. Otsuka, H. Aikawa, S. Katsumoto and K. Kobayashi, Phys. Rev. B **73**, 195329 (2006).
[6] M. Hashisaka, Y. Yamauchi, S. Nakamura, S. Kasai, K. Kobayashi, and T. Ono, J. Phys.: Conf. Ser. **109**, 012013 (2008).
[7] K. Ogawa, T. Katsuyama and H. Nakamura, Phys. Rev. Lett. **64**, 796 (1990).
[8] 山本喜久，宇都宮聖子，日本物理学会誌 **67**, 96 (2012).
[9] T. Byrnes, N.-Y. Kim, and Y. Yamamoto, Nat. Phys. **10**, 803 (2014).
[10] H. Deng, H. Haug, and Y. Yamamoto, Rev. Mod. Phys. **82**, 1489 (2010).
[11] N. D. Mermin and H. Wagner, Phys. Rev. Lett. **17**, 1133 (1966).
[12] P. Minnhagen, Rev. Mos. Phys. **59**, 1001 (1987).

8.7 Single electron effect and quantum confinement

In the end of the chapter, we will have a look on single electron effect and quantum confinement to zero-dimensional system in quantum dots.

8.7.1 Single electron effect

In the transport through quantum dots the first importantce is on the single electron effect. The single electron effect is in very short, the the electrostatic energy of an electrostatically isolated system changes with adding(extracting) an electron, and when this increase is larger than the thermal fluctuation, the tunneling of the electron is prohibited. This effect is called **Coulomb blockade**. The electrostatic energy of a quantum dot is described with a capacitance C of the dot and the electrostatic energy of charging by a single electron is $E_c = e^2/2C$, which is finite and even can be large for small C. As a first approximation we separates the electronic states into two: states inside the dot and those outside the dot. Hence the number of electrons in the dot takes an integer (descrete value). There are two possible simplest transport processes of single electrons from a source to a drain: the dot catches an electron from the source then releases one to the drain, and conversely the dot releases one to the drain first, then catches one from the source.

Let us take the simplest **constant interaction** model, in which any pair of electrons in the dot has the same (constant) Coulomb interaction energy U. Then the total Coulomb energy in the dot with N electrons is

$$E_{cN} = {}_NC_2U = \frac{N(N-1)U}{2} = \frac{U(N-1/2)^2}{2} - \frac{U}{8}. \tag{8.59}$$

The variation in the Coulomb energy with the transition $N \to N+1$ is

$$\Delta E_+(N) = (N-1)U. \tag{8.60}$$

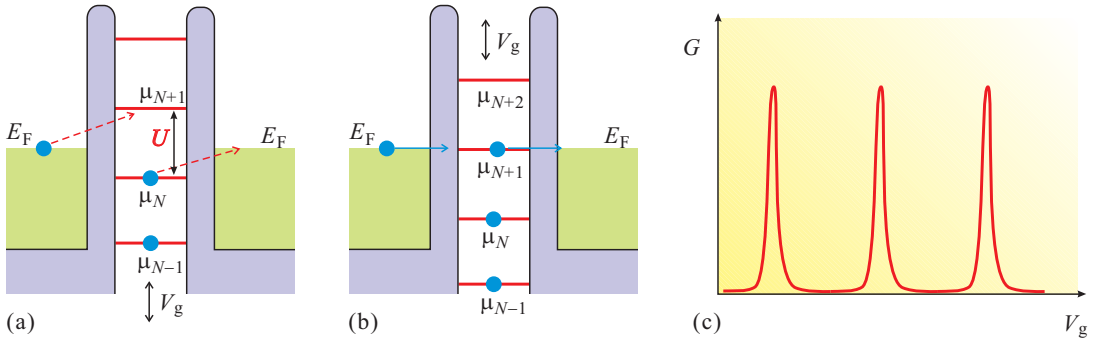


Fig. 8.26 (a) Schematic diagram of chemical potentials in the constant interaction model, in which the chemical potentials are descrete with the same distance U (the amplitude of two-electron interaction). At zero-bias, when none of the descrete chemical potentials meets the Fermi level E_F in the source and drain, a finite energy is required for an electron to tunnel, which prohibits the tunneling (Coulomb blockade). (b) When the origin of the descrete chemical potentials is shifted by the gate voltage V_g and one of them hits E_F, the tunneling thus the electric conduction becomes possible. (c) When V_g is swept, a repetition of processes (a) and (b) results in the series of sharp peaks with a regular interval in the quantum dot conductance G (Coulomb oscillation).

If we ignore other kinds of energy, $\Delta E_+(N)$ should be the electrochemical potential of N-th electron and from Eq./ (8.60) we see that the electrochemical potentials are ordered with the same distance being proportional to N.

8.7.2 Coulomb oscillation, Coulomb diamond

Let us write the electrochemical potential of the dot with N-electrons μ_N, and let N_0 be the electron number when the dot is in equilibrium with the electrodes, then $\mu_{N_0} < E_{\rm F} < \mu_{N_0+1}$. When μ_{N_0} is equal to $E_{\rm F}$ (Fermi energy in the electrodes), electrons can go into or out from the dot with tunneling from the electrodes, then at zero source-drain voltage ($V_{\rm sd} = 0$) the electric conductance ($G(0)$) takes a finite value. When μ_{N_0} does not hit $E_{\rm F}$, the tunnleing of an electron between the electrodes and the dot requires a finite energy and is prohibited (Coulomb blockade). As in Fig. 8.26(a), (b), that condition of finite $G(0)$ appears with a constant interval. Hence $G(0)$ forms regular peaks for a sweep of $V_{\rm g}$ as shown in Fig. 8.26(c), which is called **Coulomb oscillation**.

The constant interaction model can also be described as a simple circuit model illustrated in Fig. 8.27(a). Here the charge of an electron is $-e$.

$$Q_1 + Q_2 = -eN, \quad Q_1 = CV_{\rm d}, \quad Q_2 = C_{\rm g}(V_{\rm d} - V_{\rm g}), \tag{8.61}$$

and the charging energy is

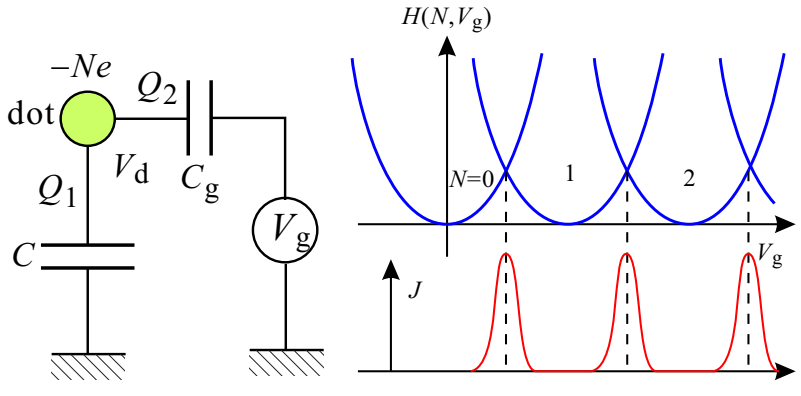
$$E = \frac{1}{2}CV_{\rm d}^2 + \frac{1}{2}C_{\rm g}(V_{\rm d} - V_{\rm g})^2, \tag{8.62}$$

in which the second term is the integral of the work done by the power source connected to the gate electrode from voltage 0 to $V_{\rm g}$. When we thermodynamically treat the problem whether the process proceeds or not under the condition that some system outside automatically provides energy, we need to consider **enthalpy**, which in the case of the pressure of atomosphere, written as $H = U - PV$. Here PV, the product of pressure and volume, corresponds to the automatic energy supply corresponding to the second term in Eq. (8.62). Then from (8.61) and (8.62),

$$H(N, V_{\rm g}) = \frac{(Ne - C_{\rm g}V_{\rm g})^2}{2(C + C_{\rm g})}. \tag{8.63}$$

If we plot this as a function of $V_{\rm g}$, as shown in Fig. 8.27(b), parabollas are lined up corresponding to N and the Coulomb peaks appear at the crossing points of the prabollas.

Next we consider the case that the gate voltage is fixed, the drain is grounded, and the source voltage is swept. The simplest model for such situation is shown in Fig. 8.28. At the positions of Coulomb peaks, the topmost chemical potential of the dot hits the Fermi level in the source and drain electrodes. The number of electrons differs by 1 at (a) and at (c). When $V_{\rm g}$ is at (b), the current is blocked at zero bias. Then if we increase the source voltage (decrease $E_{\rm F}$ in the source), and the chemical potential position used for the conduction at (a) goes in between the Fermi levels of source



(a) (b)

Fig. 8.27 (a) Simple model of the single electron charging in a quantum dot described with a self capacitance C and a gate capacitance $C_{\rm g}$. (b) Enthalpy $H(N, V_{\rm g})$ calculated in the model as a function of $V_{\rm g}$.

and drain, a finite conductance appears. Then a conduction appears outside the yellow parallelogram, inside which the conduction is Coulomb blocked. The parallelogram is called **Coulomb diamond** [*1].

The color plot in Fig. 8.29 shows an example of measured Coulomb diamonds. The sizes of the diamonds are not the same mostly because of the quantum confinement effect explained in the next section. There also should be the variation in the dot size by V_g and the variation in the effective capacitance. Various other effects should affect the sizes of the diamonds and conversely from the size we can know various physical propaties of the dot[2].

We see some line structures outside the diamonds, which come from the quantum confinement and the level descreteness (next section). There are also vague tile-like structures outside the diamond, which is called Coulomb staircase and due to the increase in the number of possible chemical potential levels. In the experiment, we also see that the vertical boundaries have a bit slanted. This is due to the capacitance between the source electrode and the dot. The capacitance mediates some of the electric force lines from the source to the dot.

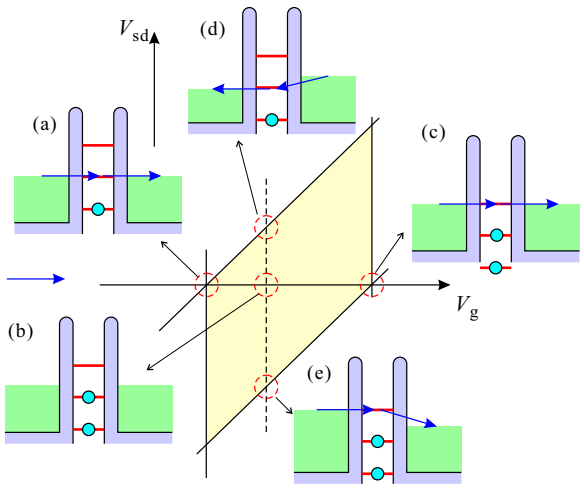


Fig. 8.28 Simple description of Coulomb diamonds. In the present case the bias voltage is applied to the source electrode (the drain is grounded). Yellow colored region is a Coulomb diamond (Coulomb blocked region). (a), (c) At zero-bias condiction, electric conduction occurs with tuning the dot chemical potential through the gate voltage. In (b), the system is out of the above resonance condition and the conduction is Coulomb blocked. At finite bias voltages on the source, the conduction appears, in (d) with the use of the chemical potential position that used in (a), vice versa in (e). Finally, the conduction is prohibited in the yellow colored region.

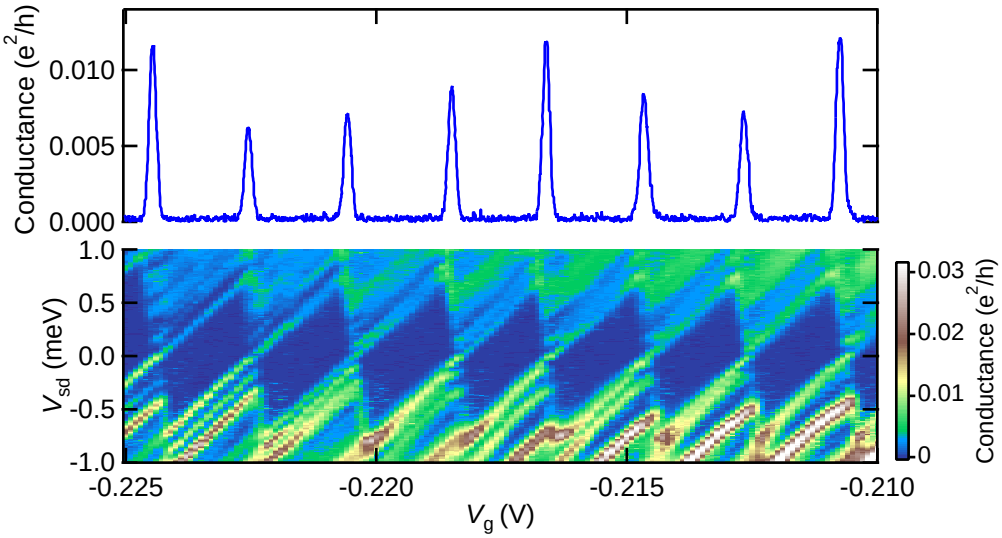


Fig. 8.29 Coulomb diamond structure appeared in the transport through a quantum dot made from 2DEG at a heterointerface. The abscissa is the gate voltage V_g. The upper panel shows the zero-bias conductance, which shows Coulomb oscillation. In the lower, the conductance is color plotted on the plane of V_g-V_{sd}. Clear diamond structures are observed. The parallel lines outside the diamonds are from the conduction through excited states in the dot.

[*1] The parallelogram becomes a diamond for symmetric configuration of voltage sources.

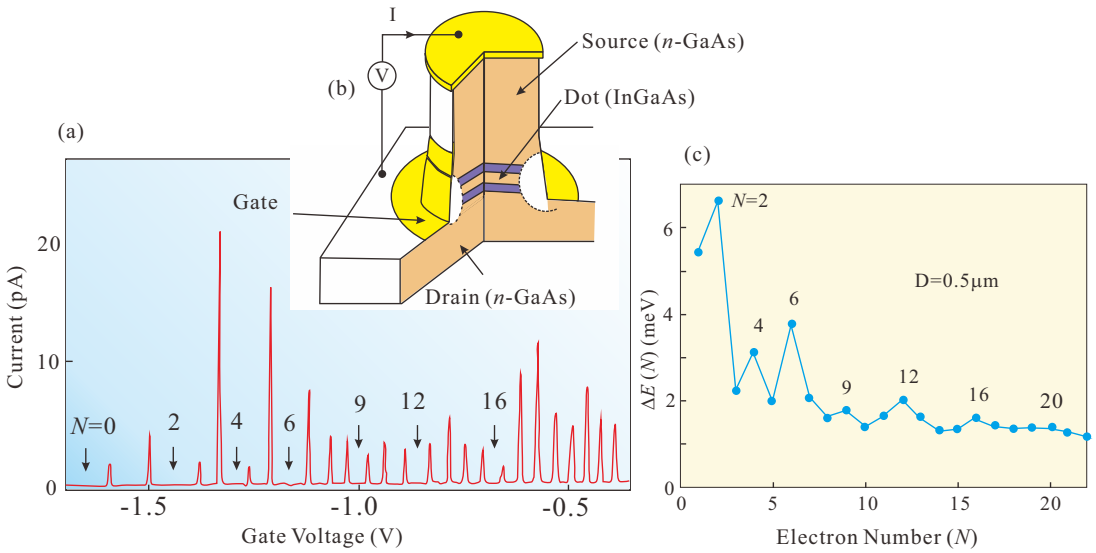


Fig. 8.30 Coulomg oscillation in a vertical type quantum dot. (a) Dot current as a function of the gate voltage. There is no Coulomb peak for further negative $V_{\rm g}$ than the region indicated as $N = 0$. The inset illustrates the sample structure. (b) Distances between the Coulomb peaks as a function of the electron number. The data are from [5].

8.7.3 Quantum confinement

Next we consider the case we cannot ignore the descreteness of the orbital energy due to quantum confinement. We number the orbital energy levels from the ground state (different numbers are assigned to all the degenerate states). Let the energy of i-th state be ϵ_i. We also ignore the terms that do not have relation with N. Then enthalpy H is

$$H(N) = \frac{(Ne - C_{\rm g}V_{\rm g})^2}{2C_s} + \epsilon_N. \tag{8.64}$$

The crossing points are obtained as

$$\begin{aligned} \Delta H(N, N+1) = H(N+1) - H(N) = \frac{e}{C_s}\left\{\left(N + \frac{1}{2}\right)e - C_{\rm g}V_{\rm g}\right\} + \Delta\epsilon_N \qquad \Delta\epsilon_N \equiv \epsilon_{N+1} - \epsilon_N \\ V_{{\rm g}X}(N, N+1) = \frac{1}{C_{\rm g}}\left\{\left(N + \frac{1}{2}\right)e + \frac{C_s}{e}\Delta\epsilon_N\right\}, \end{aligned} \tag{8.65}$$

which has a shift from the Coulomb peak position $\Delta\epsilon_N$ as in Eq. (**??**). From the shift we can get the energy spectrum in the quantum dots. This method is called addition energy spectroscopy. Because in the case of degeneracy, $\Delta\epsilon_N = 0$ and from the position of Kramers degeneracy, we can perform quantitative spectroscopy with this as a standard.

Let us have a look on a famous example, in which the researchers realized a two-dimensional harmonic potential. In this experiment, a two-dimensional quantum well was inserted into barrier layers and metallic doped "electrode" layers (source and drain) were placed at the top and the bottom of a cylindrical specimen (vertical type quantum dot). The confinement along vertical direction is strong and we only consider the ground state for this direction. Figure 8.30(a) shows the Coulomb oscillation and there is no peak in the left side (negative $V_{\rm g}$) of a small peak at about -1.6 V, which fact indicates that the dot is empty in this region. Then we can assign $N = 0$ to this region and then we can also assign the other number of electrons to blockade regions.

We review two-dimensional harmonic oscillator shortly. The two dimensional coorinate is take to xy. The in plane confinement potential $V(x, y)$ and the descrete eigenenergies can be written with the parameter ω_0 representing the strength of the potential as

$$V(x, y) = \frac{m\omega_0^2}{2}(x^2 + y^2), \qquad E_{n_h} = \hbar\omega(n_h + 1) \ \ (n_h = 0, 1, 2, \cdots), \tag{8.66}$$

which have an equidistance. Bound eigenstates of an isotropic potential can be indexed by the quantum number of angular momentum l and the radial quantum number n_r. In the above case $n_h = 2n_r + |l|$. The number of possible combinations (n_r, l) is $n_h + 1$, then with spin degeneracy, E_{n_h} has $2(n_h + 1)$ fold degeneracy.

As a simplest analysis of the data in Fig. 8.30(a), the peak intervals are plotted versus the number of electrons N in Fig. 8.30(b). Clear peaks are observed at $N = 2, 6, 12$. This reflects the fact that the bound states in two dimensional harmonic potential take shell structures at $\sum_{j=0}^{n_h} 2(n_h + 1) = (n_h + 1)(n_h + 2)$. In the simplest model, the shift of Coulomb peaks should correspond to $\hbar\omega_0$, hence the peak height in Fig. 8.30(a) should be common. In the experimental data, however, there is a strong tendency that the peak interval decrease with the number of electrons. The tendency is considered to be mainly due to the increase in the effective capacitances. We also see small peak structures at the middle of the clear peaks $N = (n_h + 2)^2$ (4,9). This comes from **Hund's rule**, which tells that the states should be occupied by electrons as to maximize the total spin due to the exchange energy. With quantum dots we can perform experiments knowing the number of electrons and informaiton on potential, quantum dots are sometimes called as "artificial atoms."

Let us see the effect of magnetic field vertical to the 2d-plane. Here we ignore the Zeeman effect. The effect of magnetic field on the orbitals appears in two terms in Hamiltonian. First is the inner product of angular momengum $\boldsymbol{l}$ and the field flux density vector. Second is the confinement into two-dimensional harmonic potential due to the cyclotron motion. The second effect modifies the effective confinement potential as

$$V_{\text{eff}}(x, y) = \frac{m\Omega^2}{2}(x^2 + y^2), \qquad \Omega \equiv \sqrt{\omega_0^2 + (\omega_c/2)^2}, \tag{8.67}$$

where $\omega_c = eB/m$ is the cyclotron frequency for magnetic flux density B. Then the energy corresponds to (n_r, l) is

$$E(n_r, l) = \hbar\Omega(2n_r + |l| + 1) + \hbar\omega_c l/2. \tag{8.68}$$

For general finite fields the orbital degeneracy is lifted by the angular momentum. The eigenstates with energies in (8.68) are called **Fock-Darwin states**. The energies in Eq. (8.68) vary with magnetic field as plotted in Fig. 8.31(a). We write

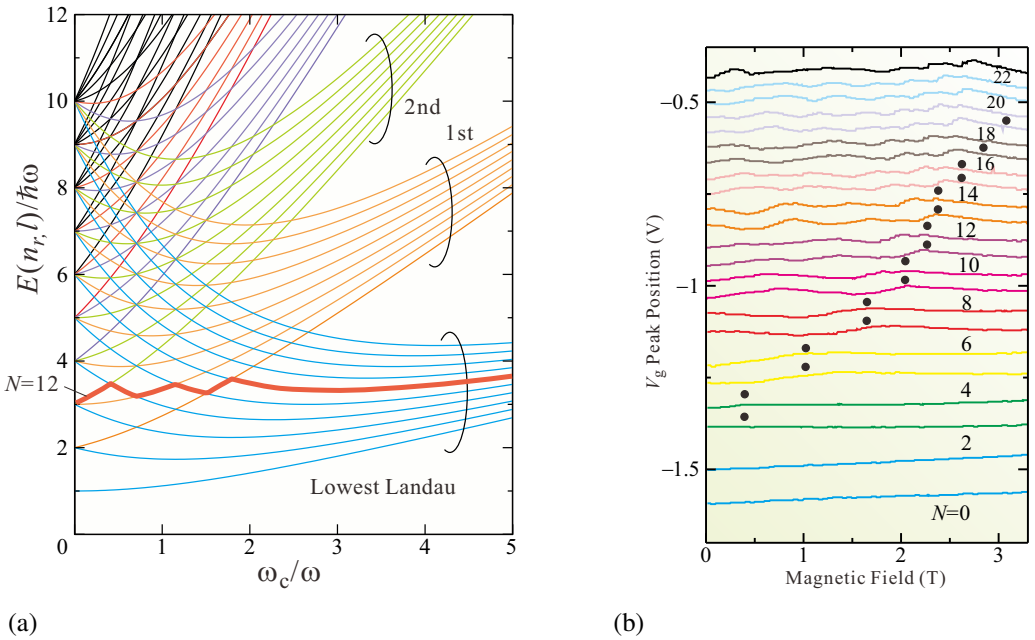


Fig. 8.31 (a) Energy levels of Fock-Darwin states (from the ground state to 10th excited state at zero field) as a function of magnetic field. They converge into Landau levels at high fields. Line colors are assigned from corresponding Landau levels. Thick red line is the trace of ground state for electron number $N = 12$. (b) Coulomb peak positions of the quantum dot in Fig. 8.30 versus vertial magnetic field. Black dots are calculated from Eq. (8.69), which represents the position of last crossing point for fixed N. The potential parameter ω is determined from the peaks $N = 3 \sim 6$.

$n_{\rm L} \equiv n_r + (|l| + l)/2$ and take the limit $B \to \infty$, to obtain $E(n_r, l) \to \hbar\omega_c(n_{\rm L} + 1/2)$. That is, they converge into **Landau quantized** levels.

As shown in Fig. 8.31(a), the levels depend on magnetic field with many crossings. The ground state of electrons with a fixed number is given by packing electrons from lower levels. The line of topmost level accommodating electrons should have kinks at such crossings. In Fig. 8.31(a), one of such a line is indicated as a thick red line for $N = 12$. The series of kinks ends up at the field where all the electrons are accomodated into the states corresponding to the lowest Landau level. The last crossing is between the line going to the lowest Landau level for N and the line for $(n_r, l) = (0, 1)$. Because of the spin degeneracy (ignoring Zeeman splitting), the former is given as $(n_r, l) = (0, -{\rm int}(N/2))$ (${\rm int}(x)$ is the largest integer smaller than or equal to x). This condition is given as

$$2\hbar\Omega + \hbar\omega_c/2 = \hbar\Omega({\rm int}(N/2) + 1) - \hbar\omega_c{\rm int}(N/2)/2$$
$$\therefore \quad \left(\frac{\omega_c}{\omega}\right)^2 = {\rm int}(N/2) - 2 + \frac{1}{{\rm int}(N/2)}. \tag{8.69}$$

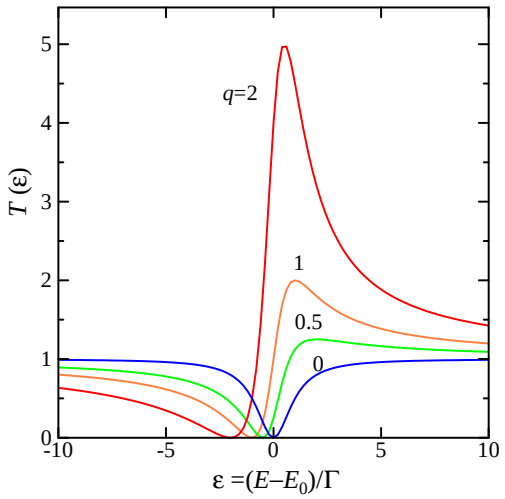
In the first approximation the Coulomb peak distance is constant and ignoring the last term for large N, the last crossing points depends on magnetic field prabolically. Actually such behavior is observed in Fig. 8.31(b). If we determine ω to fit Eq. (8.69) to kinks of $N = 3 \sim 6$, and put dots to the predicted end points of kinks, they agree nicely with the data up to $N = 14$.

8.8 Quantum dots and quantum circuits

Quantum dots (QDs) can be connected with quantum wires to form quantum circuits. A QD affects the circuit conductance though the transmission probability and the phase shift as characteristics of resonant scattering. Here the effect of single electron charging is only on the positions of chemical potentials for resonant scatterings. Hence in the simplest approximation, QDs are treated simple resonators and we do not consider the single electron effect explicitly.

8.8.1 Quantum dot and scattering experiment

Transport in mesoscopic systems can be viewed as scattering experiments in solids. We can see that clearly in quantum circuits with QDs.



For example, in the scattering of electrons with an atom, the **Fano resonance** occurs as the interference between the incident wave with continuous energies and the wave scattered by discrete atomic levels. And the same effect is observe in circuits with QDs. In the Fano effect, the scattered wave gets rapid phase shift by π over the resonance position and the interference results in distorted lineshape in the resonance. There the energy dependence of transmission coefficient is given as

$$T(\tilde{\epsilon}) \propto \frac{(\tilde{\epsilon} + q)^2}{\tilde{\epsilon}^2 + 1}, \quad \tilde{\epsilon} \equiv \frac{E - E_0}{\Gamma}. \tag{8.70}$$

Here q is called Fano parameter, which determines the lineshape as in the left figure. The larger absolute value in q results in the larger asymmetry and $q = 0$ gives a symmetric dip (anti-resonance).

Here we do not go into the derivation of (8.70)[2]. Rather we modelize the circuit with S-matrices and obtain the lineshape numerically. One-dimensional band of quantum wires are assigned to "continuum states" and quantum confined

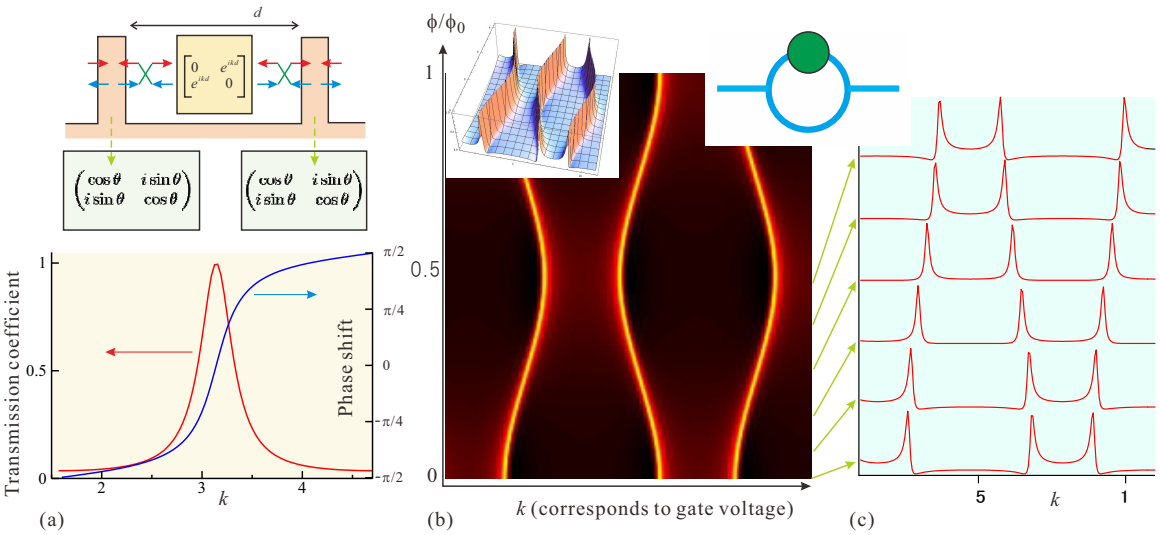


Fig. 8.32 (a) Upper: Double barrier model. Lower: Transmission coefficient (red) and phase shift (blue) of the model in the upper panel. k corresponds to gate voltage. The reflaction coefficient of a barrier is 0.7. (b) Color plot of conductance of the system shown in the right upper inset (AB ring+QD) versus plane of k (gate voltage) and magnetic flux ϕ piercing the ring. The conductance is higher for black $\rightarrow$ red $\rightarrow$ yellow. In the AB ring model in Fig.8.19(a), $\boldsymbol{S}_{\rm w}$ is replaced with the S-matrix obtained in (a), and a finite transmission coefficient is introduced into $\boldsymbol{S}_{\rm AB}$. Reflection at the dot barrier is 0.7 and that in the reference arm is 0.82. (c) Transmission of (b) is plotted versus k for ϕ/ϕ_0 =0, 0.01, 0.19, 0.29, 0.38, 0.48 from down to up.

discrete states in a QD are assigned to "discrete states." To have interference between incident wave and scattered wave, the incident wave is devided into two, one of which goes through a QD and the ohter directly goes to the outlet. A QD is formed as a one-dimensional double barrier structure. The model is described by S-matrices as in Fig. 8.32(a). That is, the barrier and the dot S-matrices are

$$S_{\rm b} = \begin{pmatrix} \cos\theta & i\sin\theta \\ i\sin\theta & \cos\theta \end{pmatrix}, \quad S_{\rm d} = \begin{pmatrix} 0 & e^{ikd} \\ e^{ikd} & 0 \end{pmatrix}. \tag{8.71}$$

Here k is the wavenumber representing kinetic energy, which corresponds to a gate voltage. In this model, the transmission coefficient and the phase shift are calculated from the composite S-matrix as shown in Fig. 8.32(a), where π phase shift at the resonance peak is clearly observed. This is common for resonance. Resonance is a response of system, in which a pole exists in the region $\mathrm{Re}(z) < 0$ on the complex z-plane. The angle from the pole to a point on the real axis changes from $-\pi$ to 0 with the movement of the point from $-\infty$ to $+\infty$.

As in Sec. 8.5.4, an S-matrix for two junctions with three channels is written as

$$S_t = \begin{pmatrix} 0 & -1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & 1/2 & -1/2 \\ -1/\sqrt{2} & -1/2 & 1/2 \end{pmatrix}. \tag{8.72}$$

Also for the AB phase, we insert an S-matrix

$$S_{\rm AB} = \begin{pmatrix} 0 & e^{i\theta_{\rm AB}} \\ e^{-i\theta_{\rm AB}} & 0 \end{pmatrix}, \qquad \theta \equiv 2\pi\frac{\phi}{\phi_0} = \frac{e}{\hbar}\phi \ \ (\phi \text{ is flux through the ring.}) \tag{8.73}$$

to one of the parallel paths. And the QD represented in (8.71) is inserted into the other path. Though thus obtained analytical form of transmission coefficient is complicated, numerical calculations shows clear Fano effect as in Fig. 8.32(b), (c). The direction of the lineshape distortion (parameter q in (8.70)) changes with the period ϕ_0 as shown in (c). This is natural consequence of the interference and evidences that the distortion comes from the rapid π change in the phase shift appeared in Fig. 8.32(a).

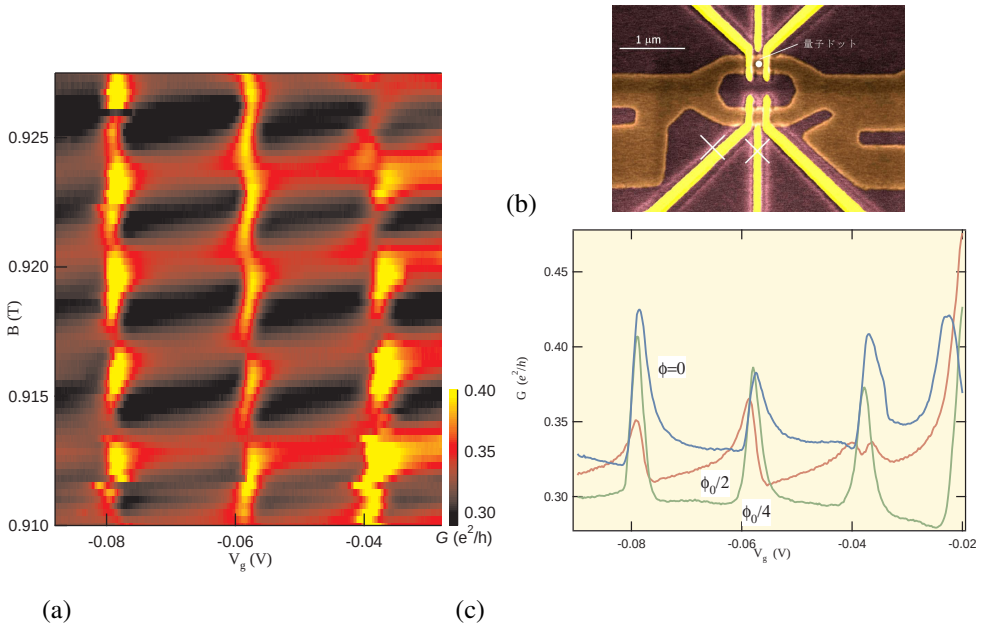


Fig. 8.33 (a) Color plot of conductance of an AB ring with a QD on one of the paths versus gate voltage and magnetic field. (b) Scanning electron micrograph of the sample. The gates with $\times$ mark are not used. (c) Coulomb oscillations at three representative magnetic fields[1].

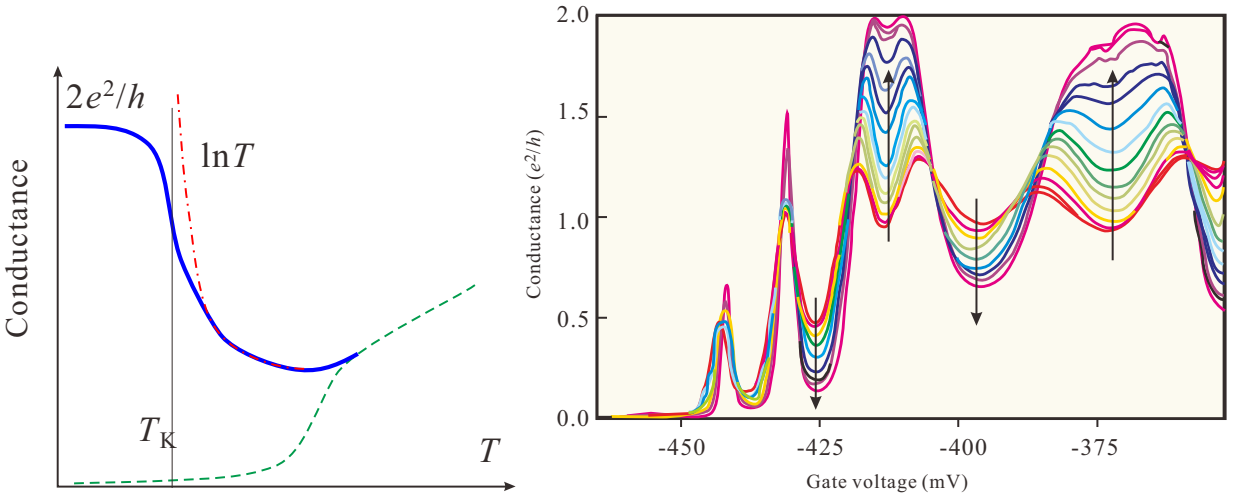
Figure 8.33 shows results of an experiment. The Fano lineshape and Fano parameter vary aginst the flux piercing the AB ring as expected.

8.8.2 Qauntum dot and the Kondo effect

The Kondo effect in QDs originates from the indirectinteraction between electrons in electrodes via localized states. We do not go into the details due to time limitation. In Appendix 8C, very short summary is given. At low temperatures most of the freedoms die out and conductors are like empty cavities (electron "waveguides" are also cavities just like microwave wavefuides). An exceptional case comes from the existence of energy-degenerated freedoms just at the Fermi level. Fermi spheres themselves are such exceptions but if there exists another degenerated freedom exists and the freedom has quantum entanglement with electrons at the Fermi level, the **Kondo effect** appears. The Kramers degeneracy due to time-reversal symmetry, i.e. spin degeneracy is an easiest example of such degenerated freedom. Hence QDs with odd number of electrons are convenient for the experiments because the topmost level is occupied by a single electron and has spin 1/2.

The Kondo effect first appeared as increase of resistance in diluted magnetic alloys with decreasing temperature. Jun Kondo gave theoretical solution to this problem and simultaneously found the divergence in the second order perturbation. This **Kondo problem** became a big problem of physics beyond the frame of solid state physics. For the problem, Anderson impurity model was proposed, **renormalization group theory** was developed. The renormalization group theory was applied to quark confinement problem in particle physics and led to the concept of asymptotic freedom and the establishment of quantum chromodynamics.

As shown in App. 8C, in very short, the Kondo effect in QDs is anomalous enhancement of tunneling probability from Hamiltonian $H_{\rm T}$ by many body resonance between degenerate freedom and the Fermi surface. In the case of QDs, the process expressed by H_T is the transmission of electrons through the QDs. That is, enhancement of H_T means enhancement of conductance. If we consider the anaoly with double barrier resoance, the Kondo many body resonance anomalously enhances the conductance and even when the original conductance is very small due to the



(a) (b)

Fig. 8.34 (a) Illustration of electric conductance through a quantm dot as a function of temperature with the emergence of the Kondo effect. Solid blue line is for odd number of electrons in the QD. In this case the Kondo effect emerges and at high temperatures the conductance is enhanced in proportion to $\ln T$ as the temperature decreases, whereas below $T_{\rm K}$ the enhancement is saturated to reach the unitary limit $2e^2/h$. Green broken line is for even number of electrons and the conductance goes to zero with the Coulomb blockade. (b) The Kondo effect appeared in the conduction experiment of a QD. The parameter is temperature and the arrows indicate the direction of lowering the temperature, with which the conductance increases in the valleys with odd nubmer of electrons and decreases in those with even number of electrons[6].

Coulomb blockade, the final transmission probability should be 1 (unitary). In this case, the conductance is, from the Landauer formula, the universal value $2e^2/h$.

A characteristic feature of the Kondo effect is that it is always in resonance with the Fermi surface. Therefore roughly speaking in the Coulomb valleys with odd number of electrons the conductance is $2e^2/h$ and in those with even number of electrons the conductance is zero. In Fig. 8.34, we show conceptual behavior and actual observation of QD conductance around the temperature characteristic of the Kondo effect (Kondo temperature, $T_{\rm K}$).

Appendix 8C: The Kondo effect

We consider a QD with the impurity Anderson model, in which only one electrons exists in the dot ($n = 1$) as the ground state. The dot has spin 1/2, can be viewed as a kind of magnetic impurity. We write the energy required to add an electron as ΔE^+, that required to extract one as ΔE^-. Then

$$\Delta E^+ = \mu_2 - \mu = \epsilon_0 + U - \mu, \quad \Delta E^- = \mu - \mu_1 = \mu - \epsilon_0. \tag{8C.1}$$

These energies give the state-allowance-times $h/\Delta E^\pm$ from the uncertainty relation for the excited states. There should be, then, second order tunneling processes with H_T by utilizing these excited states as the intermediate states. The probabilities of such processes are

$$\frac{-\gamma_L^*\gamma_R}{\Delta E^-}, \quad \frac{\gamma_L^*\gamma_R}{\Delta E^+}. \tag{8C.2}$$

Such tunnel processes of higher order is called co-tunneling. The Kondo effect can also be regarded as a phenomenon in which the tunnel probability amplitude due to co-tunneling increases anomalously.

First, the Hamiltonian

$$H = H_{\rm leads} + H_{\rm dot} + H_{\rm T} \tag{8C.3}$$

is unitary-transformed as

$$\begin{cases} c_{k\sigma} = (\gamma_L^* c_{L,k\sigma} + \gamma_R^* c_{R,k\sigma})/\gamma \\ \bar{c}_{k\sigma} = (-\gamma_R c_{L,k\sigma} + \gamma_L c_{R,k\sigma})/\gamma \end{cases}, \quad \gamma^2 \equiv \gamma_L^2 + \gamma_R^2 \tag{8C.4}$$

Then the tunnel Hamiltonian is transformed as

$$\begin{aligned} H_T &= \sum_{k,\sigma} [(\gamma_L c_{L,k\sigma}^\dagger + \gamma_R c_{R,k\sigma}^\dagger) d_\sigma + \text{h.c.}] \\ &= \sum_{k,\sigma} [\gamma c_{k\sigma}^\dagger d_\sigma + \text{h.c.}], \end{aligned} \tag{8C.5}$$

in which we can ignore $\bar{c}_{k\sigma}$ because it has nothing to do with the coupling to the dot. This transformation renormalizes the two electrodes model of QD into "a QD and a system with a Fermi surface" model. It formally equalizes a QD for transport experiment with electrodes to a magnetic impurity in a metal [*2].

The transformed Anderson impurity model Hamiltonian is written as

$$H = \sum_{k\sigma} \epsilon_k c_{k\sigma}^\dagger c_{k\sigma} + \sum_\sigma \epsilon_d d_\sigma^\dagger d_\sigma + \sum_{k\sigma} (\gamma c_{k\sigma}^\dagger d_\sigma + \text{h.c.}) + U d_\uparrow^\dagger d_\uparrow d_\downarrow^\dagger d_\downarrow. \tag{8C.6}$$

The condition for having single electron in the ground state of the dot is

$$\epsilon_d < E_\text{F} < \epsilon_d + U. \tag{8C.7}$$

Under the condition, we regard the interaction term (the third term with V_{kd}) of the conduction electron (s-electron) in the electrode and the dot electron (d-electron) as a perturbation. The first order of perturbation does not exist because it changes the number of d-electron, and the leading order is second. This means we need to consider co-tunneling process as of the leading order.

There are following four perturbation processes on the state in which the d-electron has up-spin $\uparrow$. The contraint is that only $\downarrow$-electron is allowed to enter the dot by Pauli principle. We write the unperturbed state as $\psi_\uparrow$.

1) $k\downarrow \to d\downarrow \to k'\downarrow$

2) $k\downarrow \to d\downarrow$, $d\uparrow \to k'\uparrow$ (down-spin electron goes into the dot then up-spin electron goes out)

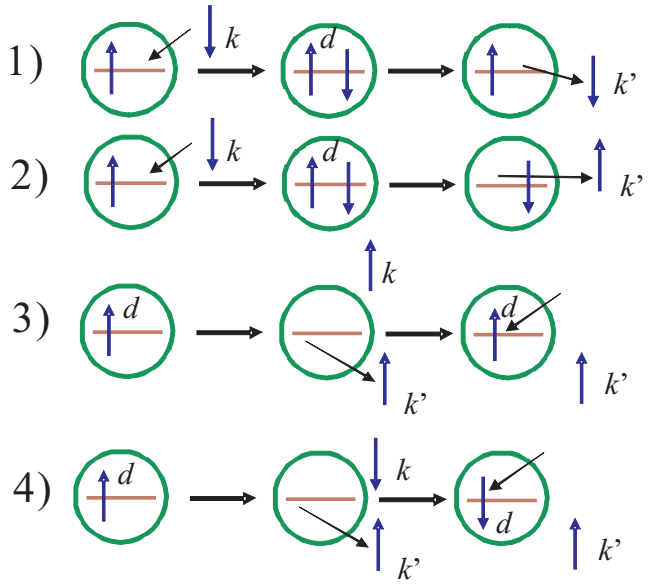


Fig. 8C.1 2nd order possible tunneling processes for the initial states of an up-spin electron inside the dot These 1)∼4) correspond to 1)∼4) in Eq. (8C.8), (8C.9) respectively.

[*2] There is a difference in the physical meaning. In the case of magnetic impurities, $c_{k'\sigma}^\dagger c_{k\sigma}$ means impurity scattering. In the case of QDs, on the other hand, it represents transmission and reflection via co-tunneling. Another difference is that in the case of impurities, k should be three dimensional vectors. This can be, however, transformed to one-dimensional problem with partial wave expansion of scattered wave, and mathematical equivalence is kept. This method is used to find the exact solution based on the Bethe ansatz.

3) $d\uparrow \to k'\uparrow$, $k\uparrow \to d\uparrow$ (up-spin electron goes out then up-spin electron goes into the dot)

4) $d\uparrow \to k'\uparrow$, $k\downarrow \to d\downarrow$ (up-spin electron goes out then down-spin electron goes into the dot)

Effective Hamiltonians for these processes are

$$1) \to -\frac{\gamma^2}{\Delta E^+} c^\dagger_{k'\downarrow} d_\downarrow d^\dagger_\downarrow c_{k\downarrow}, \quad 2) \to -\frac{\gamma^2}{\Delta E^+} c^\dagger_{k'\uparrow} d_\uparrow d^\dagger_\downarrow c_{k\downarrow}, \tag{8C.8}$$

$$3) \to \frac{\gamma^2}{\Delta E^-} d^\dagger_\uparrow c_{k\uparrow} c^\dagger_{k'\uparrow} d^\dagger_\uparrow, \quad 4) \to \frac{\gamma^2}{\Delta E^-} d^\dagger_\downarrow c_{k\downarrow} c^\dagger_{k'\uparrow} d_\uparrow. \tag{8C.9}$$

Just same as above, four perturbation processes exist for the case that $\psi_\downarrow$ is the initial state. The effective Hamiltonians for these processes are obtained by the replacement $\uparrow\downarrow \to \downarrow\uparrow$. These terms are summed up to be

$$\begin{aligned} &\sum_{k\sigma} \frac{\gamma^2}{\Delta E^-} d^\dagger_\sigma d_\sigma + \sum_{kk'\sigma} \frac{\gamma^2}{\Delta E^+} c^\dagger_{k'\sigma} c_{k\sigma} \\ &+ \sum_{kk'} \gamma^2 \left(\frac{1}{\Delta E^+} + \frac{1}{\Delta E^-} \right) (c^\dagger_{k'\uparrow} c_{k\uparrow} d^\dagger_\uparrow d_\uparrow + c^\dagger_{k'\downarrow} c_{k\downarrow} d^\dagger_\downarrow d_\downarrow + c^\dagger_{k'\uparrow} c_{k\downarrow} d^\dagger_\downarrow d_\uparrow + c^\dagger_{k'\downarrow} c_{k\uparrow} d^\dagger_\uparrow d_\downarrow). \end{aligned} \tag{8C.10}$$

The first term represents process 3) for the case of $k = k'$. Because k is outside the Fermi surface, at low temperature under Fermi degeneracy condition, we assume $c_k c^\dagger_k = 1$, $c^\dagger_k c_k = 0$. The second term is for process 1). To obtain this term we use the fact that from $d_\downarrow \psi_\uparrow = 0$, we can write $d_\downarrow d^\dagger_\downarrow = 1$, $d^\dagger_\downarrow d_\downarrow = 0$. The residual part of process 3) and those of 2) and 4) are expressed in the third term.

Here we transform the above to

$$c^\dagger_{k'\uparrow} c_{k\uparrow} d^\dagger_\uparrow d_\uparrow + c^\dagger_{k'\downarrow} c_{k\downarrow} d^\dagger_\downarrow d_\downarrow = \frac{1}{2}(c^\dagger_{k'\uparrow} c_{k\uparrow} - c^\dagger_{k'\downarrow} c_{k\downarrow})(d^\dagger_\uparrow d_\uparrow - d^\dagger_\downarrow d_\downarrow) + \frac{1}{2}(c^\dagger_{k'\uparrow} c_{k\uparrow} + c^\dagger_{k'\downarrow} c_{k\downarrow})(d^\dagger_\uparrow d_\uparrow + d^\dagger_\downarrow d_\downarrow).$$

Because the spin operator of the dot $\hat{S}$ is expressed as

$$\hat{S}_z = \frac{1}{2}(d^\dagger_\uparrow d_\uparrow - d^\dagger_\downarrow d_\downarrow), \quad \hat{S}_+ = d^\dagger_\uparrow d_\downarrow, \quad \hat{S}_- = d^\dagger_\downarrow d_\uparrow,$$

the summation of the second and the third term in (8C.10) is rewritten to the summation of the following two Hamiltonians (H_d,H_{sd}):

$$H_d = \sum_{kk'\sigma} \gamma^2 \left[\frac{1}{\Delta E^+} - \frac{1}{2} \left(\frac{1}{\Delta E^+} + \frac{1}{\Delta E^-} \right) \right] c^\dagger_{k'\sigma} c_{k\sigma}, \tag{8C.11}$$

$$H_{sd} = \sum_{kk'} \gamma^2 \left[\frac{1}{\Delta E^+} + \frac{1}{\Delta E^-} \right] \left[\hat{S}_+ c^\dagger_{k'\downarrow} c_{k\uparrow} + \hat{S}_- c^\dagger_{k'\uparrow} c_{k\downarrow} + \hat{S}_z (c^\dagger_{k'\uparrow} c_{k\uparrow} - c^\dagger_{k'\downarrow} c_{k\downarrow}) \right]. \tag{8C.12}$$

Let us define J as

$$J = \gamma^2 \left(\frac{1}{\Delta E^+} + \frac{1}{\Delta E^-} \right), \tag{8C.13}$$

then

$$H_d = \sum_{kk'} \left(-\frac{J}{2} \right) c^\dagger_{k'\sigma} c_{k\sigma} \tag{8C.14}$$

is ordinary potential scattering, which does not depend on spin. On the other hand,

$$\begin{aligned} H_{sd} &= J \sum_{kk'} \left[\hat{S}_+ c^\dagger_{k'\downarrow} c_{k\uparrow} + \hat{S}_- c^\dagger_{k'\uparrow} c_{k\downarrow} + \hat{S}_z (c^\dagger_{k'\uparrow} c_{k\uparrow} - c^\dagger_{k'\downarrow} c_{k\downarrow}) \right] \\ &= J \sum_j \left[(\hat{S}_x + i\hat{S}_y)(\hat{s}_{xj} - i\hat{s}_{yj}) + (\hat{S}_x - i\hat{S}_y)(\hat{s}_{xj} + i\hat{s}_{yj}) + 2\hat{s}_{zj}\hat{S}_z \right] \\ &= 2J \sum_j \hat{\boldsymbol{s}}_j \cdot \hat{\boldsymbol{S}} \end{aligned} \tag{8C.15}$$

is expressing the exchange interaction between spin of conduction electrons $\boldsymbol{s}_j$ and spin on the dot. This is often called sd-Hamiltonian, which originally expresses interaction between electron spin (s) and localized spin (in many cases, in

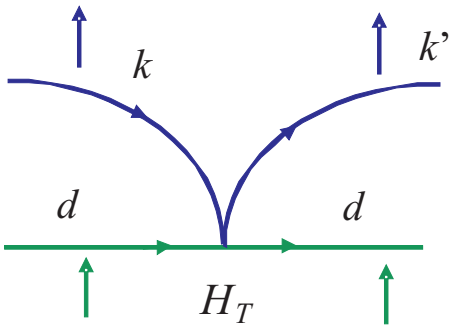


Fig. 8C.2 Diagram representing electron scattering $k \to k'$ by a dot in the first order of H_T. The time flows from the left to the right. d represents the dot and the up/down arrows indicate spins.

d-orbital, then d-spin) in diluted magnetic impurity system. Now with unitary transformation in (8C.4), we can also apply the sd-Hamiltonian to QD-electrode systems.

Then as transmission Hamiltonian H_T we ignore potential scattering H_d and take only H_{sd}. Then with adding the electorons in electrodes, the effective Hamiltonian

$$H_{\text{eff}} = \sum_{k\sigma} \epsilon_k c^\dagger_{k\sigma} c_{k\sigma} + J \sum_{kk'} \left[\hat{S}_+ c^\dagger_{k'\downarrow} c_{k\uparrow} + \hat{S}_- c^\dagger_{k'\uparrow} c_{k\downarrow} + \hat{S}_z (c^\dagger_{k'\uparrow} c_{k\uparrow} - c^\dagger_{k'\downarrow} c_{k\downarrow}) \right]. \tag{8C.16}$$

is obtaiend (Schrieffer-Wolff transformation).

Kondo calculated the scattering amplitude by the effective Hamiltonian (8C.16) to the second order in Born approximation. That is, he treated J in (8C.16) as the parameter of perturbation and calculated up to the quadratic term of J (the forth term of γ). The operator of transition between the left and right electrodes is given as

$$\hat{T} = H_T + H_T \frac{1}{\epsilon - H_0 + i\delta} H_T + \cdots . \tag{8C.17}$$

The tunnel probability of $L \to R$ is formally written as

$$\Gamma_{L\to R} = 2 \sum_{k,k'} \frac{2\pi}{\hbar} \left| \langle Rk' | \hat{T} | Lk \rangle \right|^2 \delta(\epsilon_{Rk'} - \epsilon_{Lk}) f(\epsilon_{Lk} - \mu_L)[1 - f(\epsilon_{Rk'} - \mu_R)]. \tag{8C.18}$$

Let us treat the scattering $|k\uparrow\rangle \to |k'\uparrow\rangle$. Perturbation to the first order of J is expressed in the diagram shown in Fig. 8C.2 and calculated as

$$\langle d\uparrow; k'\uparrow | \hat{T}^{(1)} | d\uparrow; k\uparrow \rangle = J/2. \tag{8C.19}$$

The conduction process H_T requires two consecutive tunnelings and this corresponds to the second order of γ (J is thus proportional to γ^2), and co-tunneling process in (8C.2).

There are three types of processes in the second order of J $\langle d\uparrow; k'\uparrow | \hat{T}^{(2)} | d\uparrow; k\uparrow \rangle$ as shown in Fig. 8C.3 and Fig. 8C.4. The first and second processes are not associated with spin flip and they are distinguished as electron process (Fig. 8C.3(a)) and electron-hole pair process (Fig. 8C.3(b)) for the intermediate propagation process [*3] . The contribution of these two terms is calculated as

$$\begin{aligned}
&\sum_q \left(\frac{J}{2}\right)^2 \frac{1}{\epsilon - \epsilon_q + i\delta}[1 - f(\epsilon_q)] + \sum_q \left(\frac{J}{2}\right)^2 \frac{-1}{\epsilon - (2\epsilon - \epsilon_q) + i\delta} f(\epsilon_q) \\
&= \sum_q \left(\frac{J}{2}\right)^2 \frac{1}{\epsilon - \epsilon_q + i\delta} \\
&= \left(\frac{J}{2}\right)^2 \int_{-D}^{D} d\epsilon' \nu \frac{1}{\epsilon - \epsilon' + i\delta} \qquad \nu : \text{Density of states} \\
&= \left(\frac{J}{2}\right)^2 \nu \left[\ln \left| \frac{D + \epsilon}{D - \epsilon} \right| - i\pi \right].
\end{aligned} \tag{8C.20}$$

[*3] Here "hole" state refers to Fermi liquid lacking single electron. This is largely different from the "hole" state defined as the state created by extracting an electron from valence band (Sec. 3.1.2).

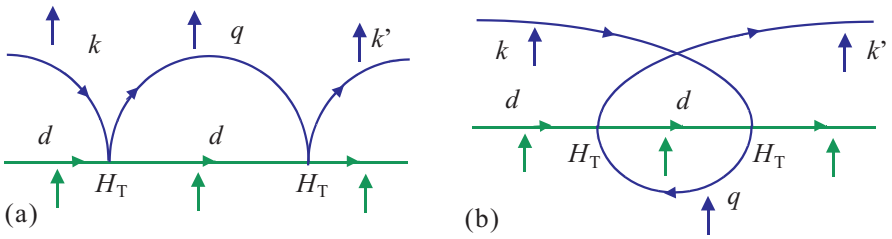


Fig. 8C.3 Non spin-flip processes for 2nd order of H_T. (a) Process with electron excitation as the intermediate state. (b) In the intermedate state of this process, an electron-hole pair propagates. The hole is annihilated by recombination with an electron in the electrode.

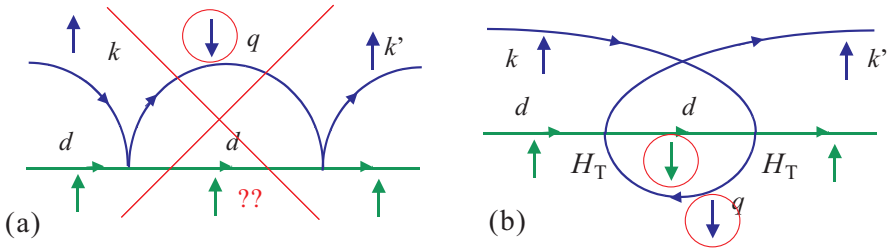


Fig. 8C.4 Processes with spin flipped intermediate stats in the second order of H_T. (a) Process with electron excitation as the intermediate state. This process is absent due to the angular momentum conservation. (b) In the intermedate state of this process, an electron-hole pair propagates.

The result does not show any anomaly as a consequence of electron-hole symmetry. Here as for the electronic states in metal, we have adopted a rough (or abstract) approximation that a band spread over $[-D, D]$ on energy with a uniform density of states ν. Such a "toy" model is often good to see the essence of phenomenon.

In the processes shown in Fig. 8C.4, spin flips occur in the intermediate states. However in (a), the dot spin should be 3/2 for the conservation of angular momentum, and this process is prohibited. In (b), an electron-hole pair propagates in the intermediate state and the contribution is calculated as

$$\sum_q J^2 \frac{1}{\epsilon - \epsilon_q + i\delta} f(\epsilon_q) = J^2 \nu \int_{-D}^{D} \frac{1}{\epsilon - \epsilon' + i\delta} f(\epsilon') d\epsilon$$
$$\approx \begin{cases} -J^2 \nu \ln |\epsilon|/D & |\epsilon| \gg k_{\rm B}T, \\ -J^2 \nu \ln k_{\rm B}T/D & |\epsilon| \ll k_{\rm B}T. \end{cases} \tag{8C.21}$$

This term diverges logarithmically with temperature lowering or smaller ϵ. This is the anomalous term found by Kondo. And various phenomena originate from this anomaly are called the **Kondo effect**.

Let us consider the origin of this term. In the case of non-spin-flip processes, the anomalous terms cancel each other due to the electron-hole symmetry. That is, if we look electrons or holes separately the anomaly exists regardless of spin-flip and the origin is the existence of Fermi surface, which represents huge asymmetry. At absolute zero, states inside a Fermi sphere are fully occupied while those above the Fermi surface are completely empty with almost infinite degeneracy. In the processes without spin flip, the electron-hole symmetry perfectly cancels this huge asymmetry. On the other hand in the processes with spin flip, conservation of spin angular momentum prohibits electron propagation process in the intermediate state [*4], and the asymmetry at the Fermi surface appears as the anomaly.

Because the perturbation leads to the divergence, the perturbative treatment itself is in failure for the condiitons close

[*4] Mathematically this comes from non-commutativity of $\hat{S}_+$ and $\hat{S}_-$ (the commutation relation yields $\hat{S}_z$), hence one can say that this is due to a quantum mechanical effect.

to the divergence. Treatment of this problem thus requires various methods other than simple perturbation. In order for handling this **Kondo problem**, a number of methods including renormalization group, have been developed. Here we do not go into the detail of such methods. Instead we have a look on the results at lowest temperatures obtained from years of research.

In sd Hamiltonian (8C.15), there is antiferromagnetic interaction between the dot spin and the conduction electron spin. That is, the dot spin attracts electrons with anti-parallel spins and repels those with parallel spins through **exchange interaction** arises from co-tunneling. As a result, seen from a distance, a cloud of spin polarization of conduction electrons appears to cling to the dot spin. The Kondo problem indicates the anomalous enhancement of the above effect. The cloud like state of spin polarization created as above is called **Kondo cloud**. On the other hand, the spin polarization surrounding the dot spin reaches complete screening of the dot spin, further polarization stops. This is called **unitary limit**.

This phenomenon can be viewed as **many-body resoance**. A representative of single-body resonance (resonace that comes from potential) is the resonant tunneling through double barrier structures. In the double barrier structure, no matter how high the barrier height is and how small the tunnel probability is, where the energy of the incident wave is in resonance, the reflected and transmitted waves are infinite sum up of coherent reflection and transmission by the two barriers. And finally the reflections cancel each other out, the total transmittance is 1. This resonance energies are close to the bound state energies of an imaginary quantum well composed by making the barrier thicknesses infinite. When the barrier thickneeses are finite, resonance makes average staying time anomalously long and the modes are called quasi-bound states. Even if an electron enters the quasi-bound state, it eventually leaks to the outside, so it is in a resonance state with the free electrons in the electrodes. When the Fermi level hits resonance, the transmission probability reaches a peak value. An example is shown in Fig. 8C.5.

The Kondo effect has many common features with double barrier phenomenon. While single-body resonance is based on infinite number of reflections, the Kondo resonance occurs as a result of infinite degeneracy at Fermi surface. In potential resonance, orbital effect results in non-uniform probability distribution. In the case of the Kondo effect, the force works among spins and no charge inhomogeneity appears. Instead, localized spin polarization occurs as Kondo cloud. The biggest difference is that the Kondo cloud is always in resonance with Fermi surface.

Electrons stay in quasi-bound states for finite times. Let τ_a be the average staying time and the resonance has lifetime broeadning $h/\tau_a(=\hbar\Gamma$, Γ is tunneling frequency). Does the Kondo cloud have width? If it does, how large is that? When the thermal broadening of Fermi surface is larger than the resonance width, the temperature dependence of the contribution of resonance to conduction should be weak. On the contrary, when the thermal width is narrower than the resonance width, the temperature dependence also disappears. The energy scale representing resonance width is, in temperature unit, called **Kondo temperature**, for which symbol $T_{\rm K}$ is often used.

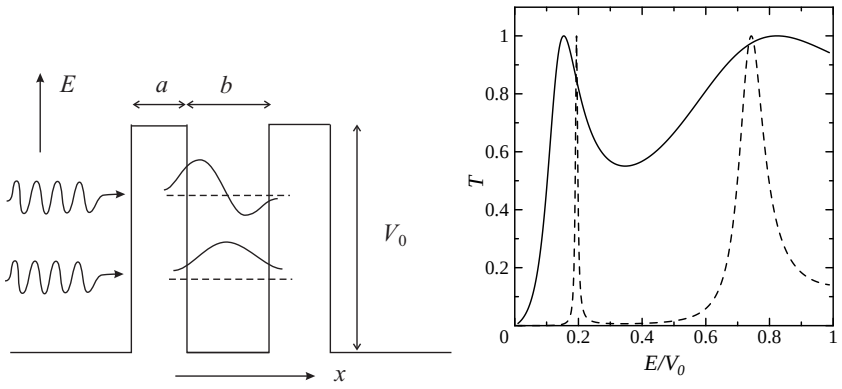


Fig. 8C.5 Left: Schematics of double barrier potential. Vertical axis is energy, horizontal axis is space coordinate. Broken lines indicate the positions of quasi-bound state energies, i.e. positions of resonant tunneling. Real wavelengths of incident waves are much longer than the illustration. Right: Example of transmission probability under condition of $k_0 \equiv \sqrt{2mV_0/\hbar}$ and $k_0 b$=5.0. Solid line is for $k_0 a$=0.5 and broken line is for $k_0 a$=2.0.

For correct estimation of Kondo temperature, perturbation in (8C.21) is not enough and the effect of higher order terms should be taken into account in some way. Though, roughly speaking, T_{K} can be estimated as the temperature-dependent term in (8C.21) is comparable with J at high temperature side. That is, from

$$-J^2\nu \ln k_{\mathrm{B}}T_{\mathrm{K}}/D \sim J,$$

T_{K} is given as

$$k_{\mathrm{B}}T_{\mathrm{K}} \sim De^{-J\nu}.$$

In the above rough estimation, with increasing the anti-ferromagnetic coupling strength J, T_{K} dcreases exponentially. Larger J corresponds to higher barriers in double barrier, narrower lifetime width with decreasing tunneling probability. The same for the density of states ν. On the other hand, widening of band D makes resonance wider, T_{K} higher. Actually in the present simple model, J and D are not independent. But with ignoring that, widening of the resonance by increasing D is interpreted as the increase of contribution from deeper inside the Fermi sphere.

References

[1] K. Kobayashi *et al.*, Phys. Rev. Lett. **88**, 256806 (2002); Phys. Rev. B **68**, 235304 (2003).

[2] 勝本信吾,「半導体量子輸送物性」 (培風館，2014).

[3] 近藤 淳,「金属電子論」(裳華房，1983)，芳田圭,「磁性」(岩波書店，1991)

[4] Yu. V. Nazarov and Y. M. Blanter, “Quantum Transport” (Cambridge, 2009).

[5] S. Tarucha et al., Phys. Rev. Lett. **77**, 3613 (1996).

[6] W. G. van der Wiel, S. De Franceschi, T. Fujisawa, J. M. Elzerman, S. Tarucha, L. P. Kouwenhoven, Science **289**, 2105 (2000).

Chapter 7 The Quantum Hall effect

In many cases, transport in higer dimensions can be understood as that in networks of one-dimensional quantum wires. On the other hand in the case of Landau quantization, mixing of two-dimensional freedoms is important and it is easier to treat the system as continuous two-dimensional space (see Sec.[**?**] for discrete treatment)..

10.1 Two-dimensional electrons under magnetic field

Let us write the Hamiltonian as

$$\mathscr{H} = \frac{m}{2}\boldsymbol{v}^2 = \frac{(\boldsymbol{p}_c + e\boldsymbol{A})^2}{2m} \equiv \frac{\boldsymbol{\pi}^2}{2m} = \frac{\pi_x^2 + \pi_y^2}{2m}. \tag{10.1}$$

$$\boldsymbol{\pi} \equiv \boldsymbol{p}_c + e\boldsymbol{A}, \tag{10.2}$$

where $\boldsymbol{\pi}$ is dynamical momentum, corresponding to real space velocity as $\boldsymbol{v} = \boldsymbol{\pi}/m^*$. $\boldsymbol{\pi}$ has commutation relations among themselves and with space coordinates as

$$[\pi_\alpha, \beta] = -i\hbar\delta_{\alpha\beta}, \quad (\alpha, \beta = x, y), \quad [\pi_x, \pi_y] = -i\frac{\hbar^2}{l^2}. \tag{10.3}$$

We see that x and y components of the momentum do not commute. The fact corresponds to the classical circulating orbits, which mix up the x and y coordinates, in other words they are no longer independent. l is called **magnetic length** defined as

$$l \equiv \sqrt{\frac{\hbar}{eB}} = \sqrt{\frac{1}{2}}\sqrt{\frac{\phi_0}{\pi B}}, \tag{10.4}$$

which is $1/\sqrt{2}$ times the radius of circle for single flux quantum ($\phi_0 \equiv h/e$). l is also called minimum Landau radius. The factor $1/\sqrt{2}$ corresponds to the zero-point energy term $\hbar\omega_c/2$ in Eq. (10.9), which we will see later.

We define the operator $\hat{\boldsymbol{R}}$ of guiding center coordinate (X, Y) as

$$\hat{\boldsymbol{r}} = \hat{\boldsymbol{R}} + \frac{l^2}{\hbar}(\pi_y, -\pi_x), \tag{10.5}$$

where $\hat{\boldsymbol{r}}$ is the real space operator of electrons. The second term in the right hand side is from the classical solution (not in this note). From the commutation relation between π_x and π_y, we get

$$[X, Y] = il^2. \tag{10.6}$$

The Hamiltonian does not depend on (X, Y), thus (X, Y) is a constant of motion while from the commutation relation in (10.6), there is an uncertainty between X and Y. Now we see that as a set of canonically conjugate variables of the system we can take $\boldsymbol{R}$, $\boldsymbol{\pi}$ other than $(\boldsymbol{r}, \boldsymbol{p}_c)$.

10.1.1 Landau quantization

As in (10.1), the Hamiltonian is quadratic for $\boldsymbol{\pi}$ and in the form of harmonic oscillator [*1], by introducing down/up operators as

$$a = \frac{l}{\sqrt{2}\hbar}(\pi_x - i\pi_y), \quad a^\dagger = \frac{l}{\sqrt{2}\hbar}(\pi_x + i\pi_y), \tag{10.7}$$

it can be written as

$$[a, a^\dagger] = 1, \quad \mathscr{H} = \hbar\omega_c \left(a^\dagger a + \frac{1}{2}\right). \tag{10.8}$$

This is in the harmonic form and the engenenergies are given as

$$E_n = \hbar\omega_c \left(n + \frac{1}{2}\right) \quad (n = 0, 1, 2, \cdots). \tag{10.9}$$

This is interpreted as the discretization of (angular) momentum with quantum confinement by magnetic field. Such quantization of orbitals by magnetic field is called **Landau quantization**.

10.1.2 Guiding center

Because $\boldsymbol{R}$ commutes with Hamiltonian (10.1), the eigenenergies in Eq. (10.9) do not depend on $\boldsymbol{R}$, thus they are degenerate as the degree of freedom in $\boldsymbol{R}$. Two dimensional systems under perpendicular uniform magnetic field still keeps spatial translational symmetry. In the set of eigenfunctions which have the guiding center as an index, the translational symmetry is kept through the freedom in $\boldsymbol{R}$. The Landau levels have large degeneracy and the basis can be taken in various form. The uncommutability between the components of $\boldsymbol{R}$ brings large variaty in the outlooks of the basis.

Let us find the basis that diagonalizes X. For that Landau gauge $\boldsymbol{A} = (0, Bx, 0)$ is convenient. Form Eq. (10.1), Schrödinger equation is given by

$$\begin{aligned}\mathscr{H}\psi = \frac{(\boldsymbol{p}_c + e\boldsymbol{A})^2}{2m}\psi &= \frac{-1}{2m}\left[\frac{\hbar^2\partial^2}{\partial x^2} - \left(-i\frac{\hbar\partial}{\partial y} + eBx\right)^2\right]\psi(\boldsymbol{r}) \\ &= \frac{1}{2m}\left[-\hbar^2\nabla^2 - 2i\hbar eBx\frac{\partial}{\partial y} + e^2B^2x^2\right]\psi(\boldsymbol{r}) = E\psi(\boldsymbol{r}).\end{aligned} \tag{10.10}$$

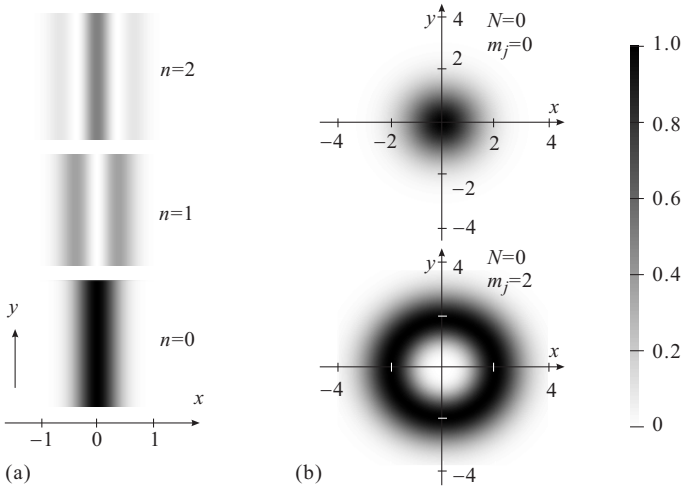


Fig. 10.1 (a) Gray scale plots of probability densities $|\psi_{n\boldsymbol{k}}(\boldsymbol{r})|^2$ in eigenstates (10.12) with three values of n, which diagonalize X. The unit of length is l_B, the width along x is about $\sqrt{2n+1}l_B$. (b) The same for the basis, which diagonalizes $X^2 + Y^2$ (not mentioned in the text). In the case of $N = 0$, the distribution is around the circle with the radius $|\sqrt{2|m_j|}l_B$ at the origin.

[*1] It is written as a sum of π_x^2 and π_y^2. π_x and π_y are canonically conjugate operators.

This Hamiltonian does not contain operator y and y-dependent part of the wavefunction should be a plane wave. Thus we substitute variable separable form $\psi(\boldsymbol{r}) = u(x)\exp(iky)$ into the above equation to obtain

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{(eB)^2}{2m}\left(x + \frac{\hbar}{eB}k\right)^2\right]u(x) = \left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{m\omega_{\rm c}^2}{2}(x + l^2k)^2\right]u(x) = Eu(x). \tag{10.11}$$

This is an equation of a one-dimensional harmonic oscillator that has the center at $x = -l^2k$. The eigenvalues are given in Eq. (10.9), and the eigenfunctions are written as

$$\psi_{nk}(\boldsymbol{r}) \propto H_n\left(\frac{x - x_k}{l}\right)\exp\left(-\frac{(x - x_k)^2}{2l^2}\right)\exp(iky) \quad (x_k \equiv -l^2k), \tag{10.12}$$

where H_n is n-th order Hermite polynomial. In each of these states, X is fixed as $X = x_k = -l^2k = -l^2p_y/\hbar$ while it is a plane wave on y spreading over whole space, namely Y is fully uncertain. In the states given by Eq. (10.12), the energy does not depend on k. Though the states are extended along y, hence the group velocity is zero ($\partial E/\partial k = 0$). On the other hand $\partial X/\partial k$ is not zero, then if some x-dependent potential is added to the system, the states gain a finite group velocity and motions along y.

Figure 10.1(a) shows gray scale plots of probability density in Eq. (10.12). We see they are uniform along y while one-dimensional harmonic oscillators along x. We are not showing the functional form here but the eigenstates can be chosen so as to diagonalize $X^2 + Y^2$. In this case, as shown in Fig. 10.1(b), the probability densities are localized both for x and y. The reason why their outlooks are so different in spite of the fact that they have the same eigenenergy, is of course it has large degeneracy and also the degeneracy comes from the freedom in $\boldsymbol{R}$, which is the freedom in the real space.

10.2 Integer quantum Hall effect

10.2.1 Shubnikov de Haas (SdH) oscillation

Let us consider the process of increasing magnetic field applied perpendicular to a two-dimensional electron system. With Landau quantization (10.9), the energy levels are as in Fig. 10.2, spead radially from the origin (Landau fan, fan diagram). How the electrons occupy those Landau levels when the system is connected to particle reservoirs as in transport experiments? The external particle reserviors make the Fermi level $E_{\rm F}$ constant but if we impose this condition, the origin in Fig. 10.2 should shift with magnetic field. The origin in Fig. 10.2 is defined as the zero-point of kinetic energy in xy-plane, namely energy levels quantized along z-axis. In the simple approximation in Sec. 7.3, the position of $E_{\rm F}$ is determined to screen the electrostatic potential formed by ionized potential with arial density $N_{\rm dep}$. Then with variation in the density of states for kinetic freedom in xy plane, the distribution of occupied states also varies to compensate the potential from the impurities. This leads to the shifts in self-consistent potential and the position of the lowest level (origin). If we look the Landau fan from the coordinate in which the origin is fixed, $E_{\rm F}$ varies with magnetic field. Below, we adopt this coordinate (constant 2DEG arial density n_s).

Let us find the arial density of states per single Ladau level $n_{\rm L}$. For that we count the number of possible wavefunctions in Eq. (10.12) in the area of $W_x \times W_y$ in xy-plane. The function in Eq. (10.12) is a plane wave along y and the "distance" of the states in k-space in $2\pi/W_y$. On the other hand, the section $0 \le X \le W_x$ corresponds to $-W_x/l_B^2 \le k \le 0$ in k-space for the wavefunctions. Hence the number of of states in the area $S = W_xW_y$ is

$$\frac{W_x/l_B^2}{2\pi/W_y} = \frac{S}{2\pi l_B^2} \quad \therefore n_{\rm L} = \frac{1}{2\pi l_B^2} = \frac{eB}{h} = \frac{B}{\phi_0}, \tag{10.13}$$

that is the number of quantum flux in the flux density. The number of Landau levels occupied by electrons is

$$\nu = \frac{\phi_0 n_s}{B}, \tag{10.14}$$

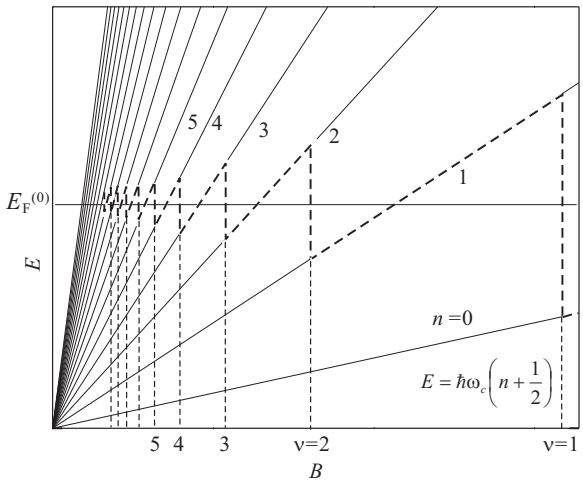


Fig. 10.2 Landau levels in (10.9) as a function of magnetic field. The broken line indicates the potision of $E_{\rm F}$ in this frame under the condition of constant n_s. $E_{\rm F}^{(0)}$ is for zero magnetic field.

which is called **filling factor**

At absolute zero, electrons occupy Landau levels from the lowest one and $E_{\rm F}$ is locked to the highest occupied Landau level. With increasing magnetic field, the changes of "topmost occupied level" take place at the points ν hit integers, where $E_{\rm F}$ shifts from $E = \hbar\omega_c(\nu + 1/2)$ to $\hbar\omega_c(\nu - 1/2)$. To summarize, in Fig. 10.2, $E_{\rm F}$ oscillates as indicated by broken line. This oscillation and the resultant oscillation in the electric resistance is called **Shubnikov-de Haas** (SdH) oscillation.

10.2.2 Localization of wavefuntion

I believe there is no rigorous proof but it is widely believed that in two-dimensional systems with some potential disorder, time-reversal symmetry, no spin-orbit interaction, all the particle states (wavefunctions) localize spatially (Anderson localization). Magnetic field breaks the time reversal smmetry and the Anderson localization is simultaneously broken. However, with further increase in magnetic field, the cyclotron radius becomes shorter than the characteristic length of the potential disorder, localization appears due to a bit different mechanism.

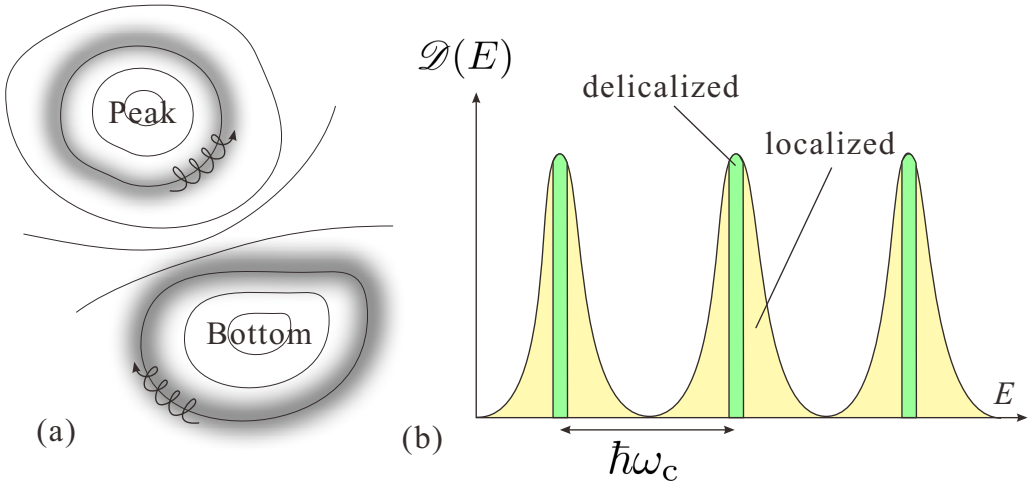


Fig. 10.3 (a) Schematic diagram showing how the Landau level wavefunction is localized by the impurity potential in a strong magnetic field. Wavefunctions in the form of Fig. 10.1(a) are bound on equipotential lines of disordered potential. The lines that depict drifting while rotating are classical orbits. (b) Landau level energies are broadened by disordered potential and localized as shadowed. Delocalized states exist around the centers of Landau "bands," corresponding to the concave-convex transition equipotential lines.

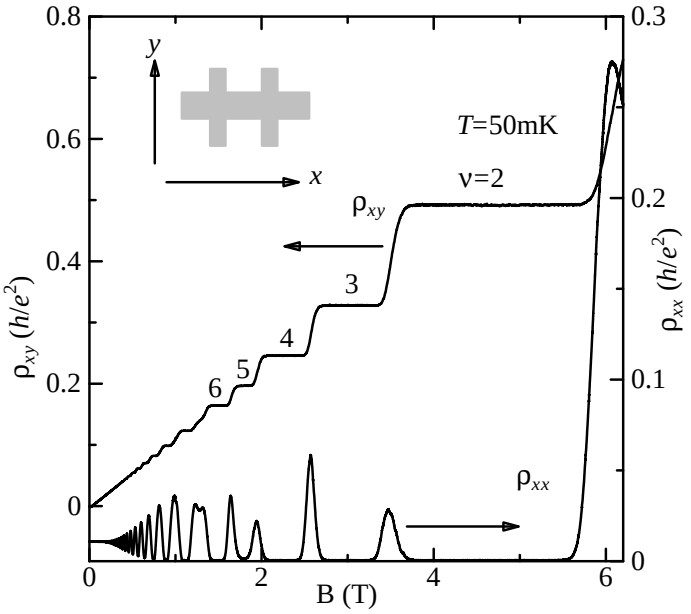


Fig. 10.4 Example of integer quantum Hall effect. 2DEG at an AlGaAs/GaAs interface is fabricated into the shape of Hall bar shown in the inset. A current is applied to the long thick line (x-direction), ρ_{xx} is obtained from the voltage between two probes placed along x while ρ_{xy} is obtained from the probes placed face to face along y.

Such behavior is illustrated in Fig. 10.3(a). Electric field applied to electons in cyclotron motion causes movement perpenducular to the field. In nonuniform potential as shown in the figure, such movement of electrons results in a rounding motion bound on an equipotential line. Then the state as a whole is spatially localized. Such spatial confinement leads to broadening of Landau levels as we have seen in Fock-Darwin state. Then the delta-function density of states of original Ladau level gets broadening as illustrated in Fig. 10.3(b). On the other hand, there are a small number of equipotential lines that do not make a closed loop between potential peaks and dips as in Fig. 10.3(a). Such a state on non-closed equipotential line should be extended and it is known that each broadened Landau "band" has a single such extended state at the center. This is also illustrated in Fig. 10.3(b).

10.2.3 Characteristics of integer quantum Hall effect

In Fig. 10.4 we show an example of measured integer quantum Hall effect (IQHE). Increasing magnetic field perpendicular to two-dimensional plane, the Hall resistance ρ_{xy} deviates from classical linear dependence on magnetic flux density B (Eq. (5.15)) and a clear staircase structure emerges. In the IQHE, the heights of the plateaus are exactly

$$\rho_{xy} = \frac{h}{e^2}\frac{1}{n} = \frac{1}{n}\,(\mathrm{R_K}) \approx \frac{2.5812\times 10^4}{n}\,(\Omega), \quad (n = 1, 2, \dots). \tag{10.15}$$

As can be guessed in Fig. 10.4, in the plateau regions simultaneously $\rho_{xx} = 0$, that is, finite current flows without longitudinal voltage. The current here is, like superconductivity, a kind of supercurrent without energy dissipation.

10.2.4 Explanation based on edge mode transport

Comparing the experiment shown in Fig. 10.4 and the localization/delocalization diagram in Fig. 10.3(b), we see that the supercurrent which causes $\rho_{xx} = 0$ flows and ρ_{xy} is quantized when E_F does not exist in the regions of delocalized states. To put this the other way around, ρ_{xy} is in a transient region between quantized plateaus when E_F exists in the retions of delocalization. Namely the quantization and supercurrent take place when two-dimensional electrons are insulating in the bulk.

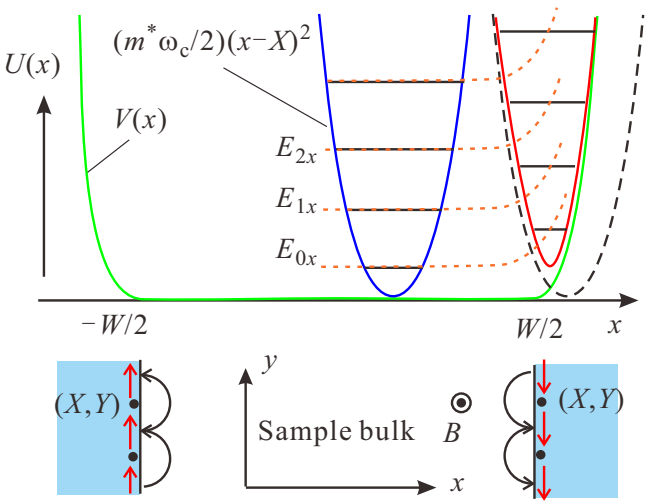


Fig. 10.5 Two-dimensional electrons under strong magnetic field are confined by "gutter-like" potential $V(x)$. The effective potential including the effect of magnetic field is expressed as $U(x)$ (sum of $V(x)$ and magnetic confinement potential). Formation of edge states is indicated by broken lines. The lower panel illustrates classical skipping orbits.

The "edge state model" explains the phenomenon based on edge mode transport. A sample with a finite width as illustrated in Fig. 10.4 inevitably has edge states[*2]. To model that, we consider a two-dimensional electron gas confined in x-direction by a well-like potential $V(x)$ with width W, spreading over y. In this model the current is applied in y-direction (for convenience the coordinate is rotated by $\pi/2$).

$V(x)$ is added to Eq. (10.10) for the wave equation. Figure 10.5 illustrates the situation, in which the gutter-like potential and the harmonic potential by magnetic field co-exist. $V(x) = 0$ deep inside the bulk and ordinary Landau quantization takes place while in the vicinity of edges, $V(x)$ makes the effective harmonic potential narrower, i.e. effective ω_c larger, hence Landau levels go up with approaching the edges. The increase of n-th Landau level begins where X-coordinate of guiding center is in the width of wavefunction $\sqrt{2n+1}l_B/2$ to the edge. In the region of such level increase,

$$\langle v_y \rangle = dE/\hbar dk = -(l_B^2/\hbar)dE/dX \tag{10.16}$$

becomes finite, giving spatial motion to Landau quantized electrons. Such mobile states correspond to classical skipping orbits, which consist of cyclotron motions and collisions to an edge as illustrated in the lower panel of Fig. 10.5. They are called **edge states**. In the edge states the direction of electron motion is determined by the sign of magnetic field.

Normalizing the edge mode wavefunction in the length L_y along y, the current brought by the mode is $j = (e/L_y)\langle v_y \rangle$. A single mode at one-side edge is occupied up to the electrochemical potential μ. We take a base energy E_0 lower than μ and higher than the bulk Landau level with the same Landau index as the edge mode. The current brought by the electrons occupying the states from E_0 to μ in this edge state is obtained from (10.13) and (10.16) as

$$J = \int_{X_0}^{X_\mu} \frac{L_y dX}{2\pi l_B^2} \frac{e}{L_y} \langle v_y \rangle = \frac{e}{h} \int \mathrm{d}X \frac{\mathrm{d}E}{\mathrm{d}X} = \frac{e}{h}(\mu - E_0). \tag{10.17}$$

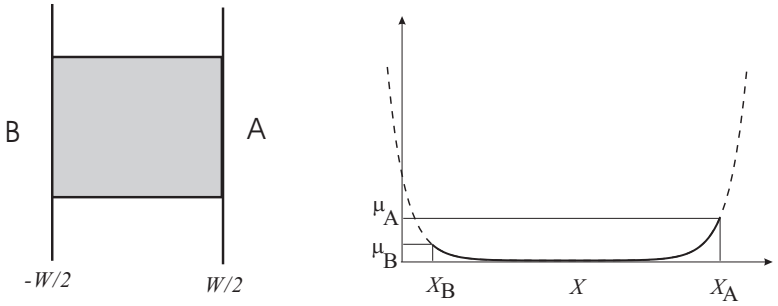


Fig. 10.6 Schematic drawing of a Landau level with edge modes. Finite net current is flowing perpendicular to the figure (y-direction) and consequently finite gradient in x-direction is given (exaggerated).

[*2] There is no edge state along the current if the two edges are connected. Such a structure in a plane is called "Corbino disk."

When $E_{\rm F}$ is in regions of localization, in equilibrium the chemical potential is uniform over the sample and the edges opposite to each other have counter-flowing currents with the same amount, the bulk states are localized and the net current is zero (circular equilibrium current is flowing at the edge). Now we apply the boundary condition that the net current J_y flows along y. As in Fig. 10.6, J_y is the difference between currents $J_{\rm A}$ and $J_{\rm B}$ at edges A and B respectively. Hence from Eq. (10.17), there should be a difference between $\mu_{\rm A}$ and $\mu_{\rm B}$, which leads to the Hall voltage. Then

$$\sigma_{xy} = \frac{J_y}{V_x} = \frac{e(J_{\rm A} - J_{\rm B})}{\mu_{\rm A} - \mu_{\rm B}} = \frac{e^2}{h}. \tag{10.18}$$

This is the conductivity for single Landau level, and for ν levels σ_{xy} is ν times of this value, thus the IQHE is explained.

The above derivation is the same as that of the Landauer formula other than crossing of x and y. The quantization is not so pricise for QPC conductance while surprizingly high precision is achieved for IQHD because of the chirality and the geometrical effect in the edge modes. In the case of QPC, conductance channels with opposite direction are spatially overlapped and backscattering of electrons can easily occurs. On the other hand in the case of IQHE, there is a macroscopic spatial distance between counter-flowing edge states and the probability of backscattering is astronomically low, and the transmission coefficient is exactly one. Therefore, the quantization of IQHE should be inaccurate if the sample width is narrowed and scattering between the edge states is likely to occur, which has been confirmed by experiments. In the above simple model, we ignore the Hall electric field inside the sample. dE/dX caused by the Hall electric field leads to finite bulk current though they cancel each other, does not contribute to J_y and the above discussion still holds.

10.3 Explanation based on topological invariant

We continue theoretica explanation for IQHE. In this section, we need to introduce several new concepts. Below we continue along Ref. [4].

10.3.1 Bloch electrons in magnetic field

We expand the concept of Bloch electron to two-dimensional electrons in magnetic field. This way of treatment is close to tight-binding model while that in Sec. 10.1 is based on two-dimensional free electron. In a two-dimensional square lattice, we write the translational operator by lattice vector $\boldsymbol{R}$ as $T_{\boldsymbol{R}}$.

$$T_{\boldsymbol{R}} f(\boldsymbol{r}) = f(\boldsymbol{r} + \boldsymbol{R}).$$

By expanding $f(\boldsymbol{r})$ with plane wave $e^{i\boldsymbol{k}\boldsymbol{r}}$, from $T_{\boldsymbol{R}} e^{i\boldsymbol{k}\boldsymbol{r}} = e^{i\boldsymbol{k}(\boldsymbol{r}+\boldsymbol{R})} = e^{i\boldsymbol{k}\boldsymbol{R}} e^{i\boldsymbol{k}\boldsymbol{r}}$, $T_{\boldsymbol{R}}$ is written as

$$T_{\boldsymbol{R}} = \exp\left(\frac{i}{\hbar}\boldsymbol{R}\cdot\boldsymbol{p}\right). \tag{10.19}$$

$T_{\boldsymbol{R}}$ commutes with $\mathscr{H}_0$ (lattice Hamiltonian for zero magnetic field) and the Bloch states are defined as the eigenstates that diagonalize the two operators simultaneously.

We then proceed to treat a system under a uniform magnetic field.

$$\mathscr{H} = \frac{1}{2m}(\boldsymbol{p} + e\boldsymbol{A})^2 + V(\boldsymbol{r}). \tag{10.20}$$

The lattice potential $V(\boldsymbol{r})$ is invariant for the operation of $T_{\boldsymbol{R}}$ though the vector potential $\boldsymbol{A}$ is not. Generally

$$\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{A}(\boldsymbol{r} + \boldsymbol{R}) + \boldsymbol{\nabla} g(\boldsymbol{r}).$$

The loss of translational symmetry due to the cyclotron motion, which does not conserve momentum. Now we consider modification of the translation operator. We define **magnetic translation operator** by replacing $\boldsymbol{p}$ with $\boldsymbol{p} + e\boldsymbol{A}$ in Eq. (10.19). Under symmetric gage $\boldsymbol{A} = \boldsymbol{B} \times \boldsymbol{r}/2$, the magnetic translation operator $T_{B\boldsymbol{R}}$ is given by

$$T_{B\boldsymbol{R}} \equiv \exp\left\{\frac{i}{\hbar}\boldsymbol{R}\cdot\left[\boldsymbol{p} + \frac{e}{2}(\boldsymbol{r}\times\boldsymbol{B})\right]\right\} = T_{\boldsymbol{R}}\exp\left[\frac{ie}{\hbar}(\boldsymbol{B}\times\boldsymbol{R})\cdot\frac{\boldsymbol{r}}{2}\right]. \tag{10.21}$$

$T_{B\boldsymbol{R}}$ commutes with $\mathscr{H}$, and there exists a basis which diagonalizes the two operators simultaneously. Care should be taken that the magnetic translational operators do not commute each other generally just like the guiding center coordinates of Landau levels do not. The commutation relation can be represented as a phase factor in

$$T_{B\boldsymbol{R}a}T_{B\boldsymbol{R}b} = \exp(2\pi i\phi)T_{B\boldsymbol{R}b}T_{B\boldsymbol{R}a}, \quad \phi = \frac{eB}{h}ab, \tag{10.22}$$

where a and b are the lengths of unit vectors. Hence ϕ is the magnetic flux piercing a unit cell in the unit of flux quantum h/e. When ϕ is a rational number p/q, commutable set of magnetic translational opertors can be prepared as a lattice limits translational vectors into discrete lattice vectors. To have simpler view, we consider a **magnetic unit cell**, which is defined from magnetic unit vectors $q\boldsymbol{a}$, $\boldsymbol{b}$ corresponding to original unit vectors $\boldsymbol{a}$, $\boldsymbol{b}$. A magnetic lattice vector $\boldsymbol{R}'$ is expressed as

$$\boldsymbol{R}' = n(q\boldsymbol{a}) + m\boldsymbol{b}. \tag{10.23}$$

Then the flux piercing the magnetic unit cell is p (integer) times a flux quantum and the magnetic translational operators $T_{B\boldsymbol{R}'}$ commute each other.

Now we take ψ as a common eigenstate of $\mathscr{H}$ and $T_{B\boldsymbol{R}'}$. Let T_{qa} and T_b (we do not write $B\boldsymbol{R}'$ for simplicity) be elements of the set of $T_{B\boldsymbol{R}'}$, then the eigenvalues are written as

$$T_{qa}\psi = e^{ik_1qa}\psi, \tag{10.24a}$$

$$T_b\psi = e^{ik_2b}\psi, \tag{10.24b}$$

where k_1, k_2 are generalized crystal momenta. In reduced zone representation, k_1, k_2 can be limited in the first **magnetic Brillouin zone** $0 \le k_1 < 2\pi/qa$, $0 \le k_2 < 2\pi/b$. The magnetic eigenstates is written in the Bloch form

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{k}\boldsymbol{r}}u_{n\boldsymbol{k}}(\boldsymbol{r}), \tag{10.25}$$

where n is a band index, $\boldsymbol{k}$ is a generalized momentum. The conditions for $u_{n\boldsymbol{k}}(\boldsymbol{r})$ are as follows.

$$u_{n\boldsymbol{k}}(x+qa, y) = \exp\left(i\frac{\pi py}{b}\right)u_{n\boldsymbol{k}}(x,y), \tag{10.26a}$$

$$u_{n\boldsymbol{k}}(x, y+b) = \exp\left(-i\frac{\pi px}{qa}\right)u_{n\boldsymbol{k}}(x,y). \tag{10.26b}$$

Then if we write $u_{n\boldsymbol{k}}(\boldsymbol{r})$ in the amplitude-phase factor form as $u_{n\boldsymbol{k}}(\boldsymbol{r}) = |u_{n\boldsymbol{k}(\boldsymbol{r})}|\exp[i\theta_{\boldsymbol{k}}(\boldsymbol{r})]$,

$$p\ \text{(integer)} = -\frac{1}{2\pi}\oint d\boldsymbol{l}\cdot\frac{\partial\theta_{\boldsymbol{k}}(\boldsymbol{r})}{\partial\boldsymbol{l}}, \tag{10.27}$$

where the integral route is taken counter clock direction along the edge of magnetic unit cell.

10.3.2 Hall conductivity from linear response theory

In the k·p perturbation for the band calculation, by renormalizing the plane wave part of wavefunction into the Hamiltonian we obtain the equation for the lattice periodic part $u_{n\boldsymbol{k}}(\boldsymbol{r})$. We can go the same way for the tight-binding model in strong magnetic field. Operation of the Hamiltonian in (10.20) on the magnetic Bloch function in (10.25) can be calculated from $\boldsymbol{p}e^{i\boldsymbol{k}\boldsymbol{r}} = e^{i\boldsymbol{k}\boldsymbol{r}}(\hbar\boldsymbol{k} + \boldsymbol{p})$ as

$$(\boldsymbol{p} + e\boldsymbol{A})^2e^{i\boldsymbol{k}\boldsymbol{r}}u_{n\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{k}\boldsymbol{r}}(\hbar\boldsymbol{k} + \boldsymbol{p} + e\boldsymbol{A})^2u_{n\boldsymbol{k}}(\boldsymbol{r}).$$

We can rewrite the Schrödinger equation as

$$\mathscr{H}_{\boldsymbol{k}} u_{n\boldsymbol{k}}(\boldsymbol{r}) = E_{n\boldsymbol{k}} u_{n\boldsymbol{k}}(\boldsymbol{r}), \quad \mathscr{H}_{\boldsymbol{k}} = \frac{1}{2m}(-i\hbar\boldsymbol{\nabla} + \hbar\boldsymbol{k} + e\boldsymbol{A})^2 + V(\boldsymbol{r}). \tag{10.28}$$

Now we utilize Kubo formula for Hall conductivity in (9B.3). We take the basis as magnetic Bloch funtions and state indices are taken as $(n, \boldsymbol{k})$. Velocity operator $\boldsymbol{v}$ can be written as $\boldsymbol{v} = (-i\hbar\boldsymbol{\nabla} + e\boldsymbol{A})/m$, and for the integration in the numerator we write the matrix element of the operator by using braket representation $u_{n\boldsymbol{k}}(\boldsymbol{r}) \to |n, \boldsymbol{k}\rangle$ as

$$\langle n, \boldsymbol{k}|\boldsymbol{v}|m, \boldsymbol{k}'\rangle = \delta_{\boldsymbol{k}\boldsymbol{k}'} \int_0^{qa} dx \int_0^b dy u^*_{n\boldsymbol{k}} \boldsymbol{v} u_{m\boldsymbol{k}'} \equiv \delta_{\boldsymbol{k}\boldsymbol{k}'} \langle n|m\rangle. \tag{10.29}$$

From the periodicity in $\boldsymbol{k}$ space, the integration just on the magnetic unit cell is enough. The normalization should be

$$\int_0^{qa} dx \int_0^b dy |u_{n\boldsymbol{k}}(\boldsymbol{r})|^2 = 1.$$

By using $\boldsymbol{k}$-dependent Hamiltonian in (10.28), we can write down the matrix elements as

$$\langle n|v_x|m\rangle = \frac{1}{\hbar}\left\langle n \left| \frac{\partial \mathscr{H}_{\boldsymbol{k}}}{\partial k_x} \right| m \right\rangle, \tag{10.30a}$$

$$\langle n|v_y|m\rangle = \frac{1}{\hbar}\left\langle n \left| \frac{\partial \mathscr{H}_{\boldsymbol{k}}}{\partial k_y} \right| m \right\rangle, \tag{10.30b}$$

where $\boldsymbol{k} = (k_x, k_y)$. These are further calculated as

$$\left\langle n \left| \frac{\partial \mathscr{H}_{\boldsymbol{k}}}{\partial k_j} \right| m \right\rangle = (E_m - E_n)\left\langle n \left| \frac{\partial u_m}{\partial k_j} \right\rangle = -(E_m - E_n)\left\langle \frac{\partial u_n}{\partial k_j} \right| m \right\rangle, \quad j = x, y. \tag{10.31}$$

Substtuting the above to the Kubo formula (9B.3) to obtain

$$\begin{aligned}
\sigma_{xy} &= -i\frac{e^2}{\hbar}\sum_{\boldsymbol{k}}\sum_n f(E_{n\boldsymbol{k}}) \sum_{m(\neq n)} \left[\frac{\langle n\boldsymbol{k}|\partial\mathscr{H}_{\boldsymbol{k}}/\partial k_x|m\boldsymbol{k}\rangle\langle m\boldsymbol{k}|\partial\mathscr{H}_{\boldsymbol{k}}/\partial k_y|n\boldsymbol{k}\rangle}{(E_{n\boldsymbol{k}} - E_{m\boldsymbol{k}})^2} - \text{c.c.} \right] \\
&= -i\frac{e^2}{\hbar}\sum_{\boldsymbol{k}}\sum_n f(E_{n\boldsymbol{k}}) \sum_{m(\neq n)} \left[\left\langle \frac{\partial u_n}{\partial k_x} \right| m \right\rangle \left\langle m \left| \frac{\partial u_n}{\partial k_y} \right\rangle - \left\langle \frac{\partial u_n}{\partial k_y} \right| m \right\rangle \left\langle m \left| \frac{\partial u_n}{\partial k_x} \right\rangle \right] \\
&= \frac{e^2}{h}\frac{2\pi}{i}\sum_{\boldsymbol{k}}\sum_n f(E_{n\boldsymbol{k}}) \left[\left\langle \frac{\partial u_n}{\partial k_x} \right| \frac{\partial u_n}{\partial k_y} \right\rangle - \left\langle \frac{\partial u_n}{\partial k_y} \right| \frac{\partial u_n}{\partial k_x} \right\rangle \Big].
\end{aligned} \tag{10.32}$$

Now we define a vector field $\boldsymbol{A}_{n\boldsymbol{k}}$ with

$$\boldsymbol{A}_{n\boldsymbol{k}} = \int d^2\boldsymbol{r} u^*_{n\boldsymbol{k}} \boldsymbol{\nabla}_{\boldsymbol{k}} u_{n\boldsymbol{k}} = \langle u_{n\boldsymbol{k}}|\boldsymbol{\nabla}_{\boldsymbol{k}}|u_{n\boldsymbol{k}}\rangle. \tag{10.33}$$

We assume $T = 0$ and that E_{F} is in the localized region. Writing the summation on $\boldsymbol{k}$ as the form of integration, σ_{xy} is given by

$$\sigma_{xy} = \frac{e^2}{h}\frac{1}{2\pi i}\sum_{E_n < E_{\mathrm{F}}} \int_{\mathrm{MBZ}} d^2k [\boldsymbol{\nabla}_{\boldsymbol{k}} \times \boldsymbol{A}_{n\boldsymbol{k}}]_{k_z} = \frac{e^2}{h}\frac{1}{2\pi i}\sum_{E_n < E_{\mathrm{F}}} \int_{\mathrm{MBZ}} d^2k [\mathrm{rot}_{\boldsymbol{k}} \boldsymbol{A}_{n\boldsymbol{k}}]_{k_z}. \tag{10.34}$$

The integration is over the magnetic Brillouin zone.

Because at the edges of a magnetic Brillouin zone, $k_x = 0$ and $k_x = 2\pi/qa$, $k_y = 0$ and $k_y = 2\pi/b$ are the same points, topologically the zone is two-dimensional torus $T^2 = S^1 \times S^1$. When $\boldsymbol{A}_{n\boldsymbol{k}}$ is single-valued on this torus, σ_{xy} calculated from (10.34) is zero as known from the Stokes theorem. That is, for $\sigma_{xy} \neq 0$, $\boldsymbol{A}_{n\boldsymbol{k}}$ should have non-trivial topology. Here it is important that the magnetic Brillouin zone is a torus, which cannot be squeezed continuously to single point. If such squeezing is possible, $\boldsymbol{A}_{n\boldsymbol{k}}$ defined on the manifold cannot have non-trivial topology.

To see the topology of $\boldsymbol{A}_{n\boldsymbol{k}}$, we consider **local gauge transformation**. A solution of Schrödinger equation (10.28) $u_{\boldsymbol{k}}(\boldsymbol{r})$ can be transformed with an arbitrary continuous function $f(\boldsymbol{k})$ to another solution

$$u'_{\boldsymbol{k}}(\boldsymbol{r}) = \exp[if(\boldsymbol{k})]u_{\boldsymbol{k}}(\boldsymbol{r}). \tag{10.35}$$

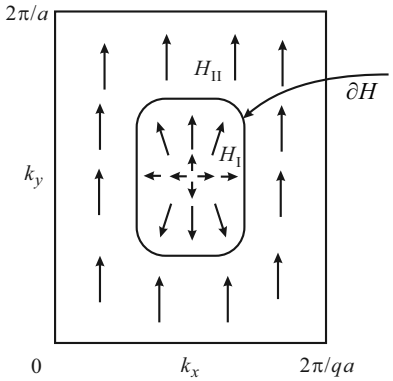


Fig. 10.7 Illustration of phase of wavefunction when it has zero in the magnetic Brillouin zone.

u and u' are physically the same. From the definition in Eq. (10.33), this transformation corresponds to

$$\boldsymbol{A}'_{n\boldsymbol{k}} = \boldsymbol{A}_{n\boldsymbol{k}} + i\boldsymbol{\nabla}_{\boldsymbol{k}} f(\boldsymbol{k}). \tag{10.36}$$

To eliminate the above uncertainty originated from the gauge undertainty, we assume fixing the phase of $u_{n\boldsymbol{k}}(\boldsymbol{r})$ at one point. With this, though, we cannot fix the entire phase over the whole magnetic Brillouin zone. We assume $u_{n\boldsymbol{k}}(\boldsymbol{r})$ is zero at a point $\boldsymbol{k}_0$. As shown in Fig. 10.7, the magnetic Brillouin zone is devided into retion $H_{\rm I}$ that contains $\boldsymbol{k}_0$ and residual region $H_{\rm II}$. If $H_{\rm I}$ contains a zero, the phase must "rotate" around the zero as in the figure. On the other hand $H_{\rm II}$ should be connected at the edges as a torus and the wavefunction should have different structure in phase. Hence we need to take different gauges in the two regions.

For simplicity we consider the contribution of band n only and n can be omitted. The integrals in (10.34) are, by applying Stokes' theorem to the two regions, given by

$$I = \frac{1}{2\pi i}\left[\int_{\rm I} d^2k[{\rm rot}\boldsymbol{A}]_{k_z} + \int_{\rm II} d^2k[{\rm rot}\boldsymbol{A}]_{k_z}\right] = \oint_{\partial H} (\boldsymbol{A}^{\rm II} - \boldsymbol{A}^{\rm I})\cdot\frac{d\boldsymbol{k}}{2\pi i}. \tag{10.37}$$

The integral over circumference of region II cancels out due to the torus boundary condition (equivalent to "back and forth" integration over a single line). On the boundary ∂H, with gauge transformation the relation of the wavefunction is expressed as

$$u^{\rm I}_{\boldsymbol{k}} = u^{\rm II}_{\boldsymbol{k}} e^{i\theta(\boldsymbol{k})}. \tag{10.38}$$

From the definition (10.33), the integral should be

$$I = \oint_{\partial H}\left[\langle u^{\rm II}_{\boldsymbol{k}}|\boldsymbol{\nabla}_{\boldsymbol{k}}|u^{\rm II}_{\boldsymbol{k}}\rangle + (i\boldsymbol{\nabla}_{\boldsymbol{k}}\theta)\langle u^{\rm II}_{\boldsymbol{k}}|u^{\rm II}_{\boldsymbol{k}}\rangle - \langle u^{\rm II}_{\boldsymbol{k}}|\boldsymbol{\nabla}_{\boldsymbol{k}}|u^{\rm II}_{\boldsymbol{k}}\rangle\right]\cdot\frac{d\boldsymbol{k}}{2\pi i} = \frac{\Delta_{\partial H}\theta}{2\pi}. \tag{10.39}$$

The phase evolution over single circulation on the boundary $\Delta_{\partial H}\theta$ should be an integer times 2π and I is limited to an integer. Let $\nu_{\rm C}$ be that integer. And let n_B be the number of bands lower or at the same level as $E_{\rm F}$, we find

$$\sigma_{xy} = n_B\nu_{\rm C}\frac{e^2}{h}, \tag{10.40}$$

which tells that the Hall conductance should be an integer times e^2/h. Equation (10.40) is called **Thouless-Kohmoto-Nightingale-Nijs** (**TKNN**) formula[5]. $\nu_{\rm C}$ is called **Chern number** and known to be 1 for the Landau bands. The above gives the same result as Eq. (10.18).

Chern number is the number of anomalies in the phase of wavefunction, equivalently the number of zeros. It is a kind of **topological invariant**. The origin of Chern number is in the topological property of energy bands. In order to turn a torus into a sphere, we should once tear up the surface around the hole then sew the surfaces together and finally erase the hole. Similarly to change the band structure into the one with different topology (Chern number), we need to crush the band gap once. For this reason, the Hall conductance found in TKNN formula is stable and precise regardless of the variety of sample properties.

Here $\boldsymbol{A}_{\boldsymbol{k}}$ is a Berry connection, ${\rm rot}\boldsymbol{A}_{\boldsymbol{k}}$ is a Berry curvature in Appendix 9A. We will revisit them in the section of topological insulator.

10.3.3 Laughlin's gedankenexperiment

Robert Laughlin considered a sample in which a 2DEG is rolled into a cylinder with a radius of R and a circular electrode is attached to the end of the cylinder (Fig. 10.8)[6]. The magnetic field is emitted outward from the core of the cylinder and is applied perpendicularly to the 2DEG. x and y axes are taken as in the figure. There is no edge because the current is applied along x and the sample is closed along y. Further, a thin, long solenoid is placed at the core and an applied current creates a magnetic flux Φ through it. The flux does not touch the 2DEG directly but gives an AB phase on orbits going around the cylinder. The vector potential for the perpendicular magnetic field and that for the field by the solenoid are in Ladau gauge

$$\boldsymbol{A} = (0, Bx), \quad \boldsymbol{A}_\Phi = \left(0, -\frac{\Phi}{2\pi R}\right). \tag{10.41}$$

We write down the wavefunction in the form of Eq. (10.12). Because the system is circular in y-directoin, the wavefunction should go around the circle.

The current in y-direction is

$$j_y = \frac{1}{L_x}\frac{\partial \mathscr{E}_t}{\partial \Phi}. \tag{10.42}$$

$\mathscr{E}_t$ is the total energy on the cylinder per the normalization length L_x [*3]. The vector potential in Landau gauge is $\boldsymbol{a} = (0, Bx - \Phi/L_y, 0)$. We take the unperturbed Hamiltonian $\mathscr{H}_0$ as the one of 2DEG under magnetic field. Then the effect of solenoid flux is taken into account by the transformation

$$k_y \to k_y - \frac{2\pi}{L_y}\frac{\Phi}{\phi_0}, \quad \left(\phi_0 \equiv \frac{h}{e}\right) \tag{10.43}$$

in the Hamiltonian. This transformation corresponds to the variation in X-coordinate of the guiding center as

$$X \to X + \left(\frac{\Phi}{\phi_0}\right)\frac{L_x}{N_\phi}. \tag{10.44}$$

Then the variation in penetration magnetic flux $\Phi \to \Phi + \Delta\Phi$ appears as that in X $\Delta X = (L_x/N_\phi)\Delta\Phi/\phi_0$. If $\Delta\Phi$ is an integer (q) times ϕ_0, then $\Delta X = qL_x/N_\phi = 2q\pi l_B^2/L_y$, which is q times the distance in x between the eigenstates. Namely the states shift to q-th next eigenstates and the variation in Φ is absorbed into the phase of wavefunction. When

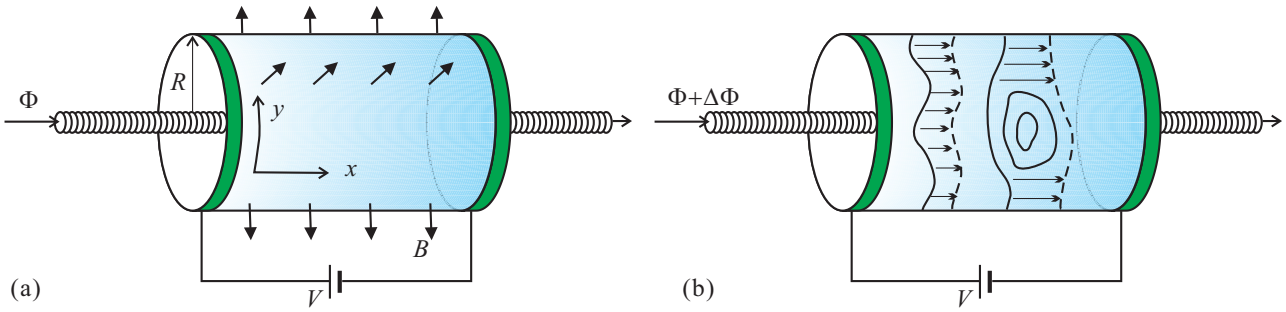


Fig. 10.8 (a) Rolled up two-dimensional system used in Laughlin's gedankenexperiment. The magnetic field B is emitted outward from the core of the cylinder and is applied perpendicularly to the 2DEG. A thin, long solenoid is running at the core of the cylinder giving an AB phase. The electric field is applied in x-direction. (b) Schematic drawing of the variation in wavefunction when the flux by the solenoid is increased from Φ to $\Delta\Phi$.

[*3] Here we prove the equation simply as follows. Let $\mathscr{L}$ be the inductance of the cylinder. A state with current J has the magnetic energy $\mathscr{E}_H = \mathscr{L}J^2/2 = \Phi^2/2\mathscr{L}$, which means $\partial\mathscr{E}_H/\partial\Phi = \Phi/\mathscr{L} = J$. Let L_x be the normalization length. From $J = L_x j_y$ we reach the equation.

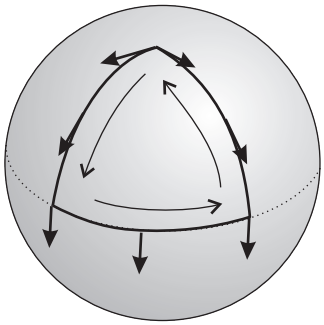
N Landau levels are completely occupied, that is $\nu = N$, and there is an electric field of E in x, the variation of $\Delta\Phi = \phi_0$ causes a variation in the energy of $\Delta\mathscr{E} = -eE\Delta X N_e$ ($N_e = \nu N_\phi = N N_\phi$).

Here we assume as follows. In the quantum Hall state, the current does not depend on the boundary condition in y-direction. In other words the current does not depend on the absolute value of Φ. Then we replace the derivative in (10.42) with the finite differece to find

$$j_y = \frac{1}{L_x}\frac{\partial \mathscr{E}_t}{\partial \Phi} = \frac{1}{L_x}\frac{\Delta \mathscr{E}_t}{\Delta \Phi} = \frac{1}{L_x}\left(-eE\frac{L_x}{N_\phi}\right)\frac{N_e}{\phi_0} = N\frac{e^2}{h}E. \tag{10.45}$$

That is the Hall conductance $\sigma_{xy} = j_y/E_x$ is quantized as an integer times e^2/h. Because e^2/h is the conductance for single band, this is an indirect proof that the Chern number of single Landau level is one.

Appendix 9A: The Berry phase



A common classic example of the Berry phase is the parallel displacement of a vector on a sphere as shown on the left. When the vector is translated in three-dimensional space on an appropriate trajectory and return to the original point, the vector does not change. However if we apply a constraint that the vector should be in-plane during the "parallel displacement" (or the vector should be in the tangent plane of the sphere), then as in the left figure, the direction of the vector generally changes when the vector origin comes back to the starting point. The angle between the starting vector and the returning vector corresponds to the Berry phase.

Let α be the angle of variation in the direction of the vector, C be the tranjectory, then α can be expressed as a line integral on C of a vector $\boldsymbol{A}$. This $\boldsymbol{A}$ is called **Berry connection** (Berry connection depends on the constraints on the vector movement). From Stokes' theorem α can also be written as the integration over an area S rimmed by C as

$$\alpha = \oint_C \boldsymbol{A}\cdot d\boldsymbol{s} = \int_S \mathrm{rot}\boldsymbol{A}\cdot d\boldsymbol{\sigma},$$

where $\mathrm{rot}\boldsymbol{A}$ is called **Berry curvature**.

Let us go to quantum mechanics. We consider a time-dependent Hamiltonian $H(t)$ and write the eigenvalue equation as

$$H(t)|n(t)\rangle = E_n(t)|n(t)\rangle. \tag{9A.1}$$

Taking time-derivatibe and operating the eigenfunction $\langle k|$ from left we get

$$\langle k(t)|\partial|n(t)\rangle/\partial t \equiv \langle k(t)|\dot{n}(t)\rangle = \frac{1}{E_n(t)-E_k(t)}\left\langle k(t)\left|\frac{\partial H}{\partial t}\right|n(t)\right\rangle. \tag{9A.2}$$

$$\therefore \langle \dot{n}|n\rangle + \langle n|\dot{n}\rangle = 0 \quad \therefore \mathrm{Re}(\langle n|\dot{n}\rangle) = 0. \tag{9A.3}$$

Let $\psi(t)$ be a solution of the Schrödinger equation composed of $H(t)$. $\psi(t)$ is expanded by $|n(t)\rangle$ as

$$|\psi(t)\rangle = \sum_n c_n(t)|n(t)\rangle \exp\left(-\frac{i}{\hbar}\int_0^t E'_n(t')dt'\right), \quad (E'_n(t) \equiv E_n(t) - \hbar\eta_n(t), \quad \eta_n(t) = i\langle n|\dot{n}\rangle). \tag{9A.4}$$

Substituting this into the Schrödinger equation we find

$$\sum_n i\hbar\left(\dot{c}_n|n\rangle + c_n|\dot{n}\rangle - \frac{i}{\hbar}E'_n c_n|n\rangle\right)\exp\left[-\frac{i}{\hbar}\int_0^t E'_n(t')dt'\right] = \sum_n c_n H|n\rangle \exp\left[-\frac{i}{\hbar}\int_0^t E'_n(t')dt'\right]. \tag{9A.5}$$

Operating $\langle k|$ from the left, from Eq. (9A.2) we obtain

$$\frac{dc_k}{dt} = \sum_{n\neq k} \frac{\langle k|\partial H/\partial t|n\rangle}{E_k - E_n} \exp\left[\frac{i}{\hbar}\int_0^t (E'_k(t') - E'_n(t'))dt'\right] c_n. \tag{9A.6}$$

We consider variation of $H(t)$ slow enough for the variation of the wavefunction to be adiabatic. We take the starting point of the wavefunction $\psi(0) = |m(0)\rangle$ and the adiabatic change means $|\psi(t)\rangle = |m(t)\rangle$ with no mixing of other eigenstates. Let us express the time evolution of H as that in a set of parameters $\{R_i(t)\}$, which can be written in the vector form $\boldsymbol{R}(t)$. We consider a loop trajectory in $\boldsymbol{R}$-space starting $\boldsymbol{R}(0)$ at $t = 0$ and coming back to $\boldsymbol{R}(0)$ at time T.

$$|\psi(t) = |m(\boldsymbol{R}(t))\rangle \exp\left[-\frac{i}{\hbar}\int_0^t E'_m(t')dt'\right] = |m(\boldsymbol{R}(t))\rangle \exp\left[-\frac{i}{\hbar}\int_0^t E_m(t')dt'\right] e^{i\gamma_m(t)}, \tag{9A.7}$$

$$\text{where} \quad \gamma_m(t) = \int_0^t \eta_m(t')dt' = i\int_0^t \langle m(\boldsymbol{R}(t'))|\dot{m}(\boldsymbol{R}(t'))\rangle dt'. \tag{9A.8}$$

As known from (9A.3), γ_m is a real number. For a loop trajectory, with variable transformation $t \to \boldsymbol{R}$,

$$\gamma_m(T) = i\int_0^T \langle m(\boldsymbol{R}(t))|\nabla_R m(\boldsymbol{R}(t))\rangle \cdot \dot{\boldsymbol{R}}(t)dt = i\oint_C \langle m(\boldsymbol{R}(t))|\nabla_R m(\boldsymbol{R}(t))\rangle \cdot d\boldsymbol{R}(t))\rangle d\boldsymbol{R} = \gamma_m(C). \tag{9A.9}$$

∇_R is the gradient operator in $\boldsymbol{R}$-space. Below we omit the subscript R. The above equation means with a loop variation of Hamiltonian associated with adiabatic transition of the state, **Berry phase** $\gamma_m(C)$ is added to the wavefunction. Further by using Stokes' theorem,

$$\gamma_m(C) = -\mathrm{Im}\oint_c \langle m(\boldsymbol{R})|\nabla m(\boldsymbol{R})\rangle \cdot d\boldsymbol{R} = -\mathrm{Im}\int_S [\nabla \times \langle m(\boldsymbol{R})|\nabla m(\boldsymbol{R})\rangle] \cdot d\boldsymbol{S} \tag{9A.10}$$

is obtained.

Appendix 9B: Kubo formula for Hall conductivity

The bf Kubo formula is the ultimate form of linear response theory developed from the first half to the middle of the 20th century. There are various mathematically equivalent expressions in the Kubo formula, but here we introduce what is called Nakano-Kubo formula. We consider a two-dimensonal electrons under perturbation eEy of electric field E in y-direction. First order perturbed states $|\alpha'\rangle$ are written by unperturbed eigenstates $|\alpha\rangle$ as

$$|\alpha'\rangle = |\alpha\rangle + \sum_{\beta\neq\alpha} \frac{\langle\beta|eEy|\alpha\rangle}{E_\alpha - E_\beta}|\beta\rangle. \tag{9B.1}$$

To consider the Hall conductance we need to sum up the contributions from each $|\alpha'\rangle$ to the current along x-direction. Then the current density in x-direction to the first order of perturbation is written as

$$j_x = \frac{1}{L^2}\sum_\alpha f(E_{\alpha'})\langle\alpha'|\hat{j}_x|\alpha'\rangle = \frac{1}{L^2}\sum_\alpha f(E_\alpha)\sum_{\beta\neq\alpha} \frac{\langle\alpha|(-ev_x)|\beta\rangle\langle\beta|eEy|\alpha\rangle}{E_\alpha - E_\beta} + \text{c.c.}, \tag{9B.2}$$

where $f(E)$ is the Fermi distribution function, L^2 is the area of normalization. Because the perturbation term is odd function, there is no first order energy correcltion, and $E_{\alpha'} = E_\alpha$. From

$$\langle\beta|v_y|\alpha\rangle = \langle\beta|\dot{y}|\alpha\rangle = -\frac{i}{\hbar}\langle\beta|[y,\mathscr{H}]|\alpha\rangle = -\frac{i}{\hbar}(E_\alpha - E_\beta)\langle\beta|y|\alpha\rangle,$$

this $\langle\beta|y|\alpha\rangle$ is substituted into Eq. 9B.2 to obtain

$$\sigma_{xy} = \frac{j_x}{E} = \frac{e^2\hbar}{iL^2}\sum_\alpha f(E_\alpha)\sum_\beta \frac{\langle\alpha|v_x|\beta\rangle\langle\beta|v_y|\alpha\rangle}{(E_\alpha - E_\beta)^2} + \text{c.c.}. \tag{9B.3}$$

Appendix 9C: Fractional quantum Hall effects

In the quantum Hall effect, various novel phenomena and ideas have been found. Among them we have a very short look at the fractional quantum Hall effect.

9C.1 Experiment on fractional quantum Hall effects

Fractional Quantum Hall Effect (FQHE) was found in transport experiment in a high-mobility 2DEG. In IQHE, the Hall conductance plateaus appear at $\sigma_{xy} = nG_q$ (n is an integer) while in FQHE the conductance plateaus appear at

$$\sigma_{xy} = fG_q, \quad f = \frac{m}{n} \ \ (n : \text{odd integer}, \ \ m : \text{integer}). \tag{9C.1}$$

Figure 9C.1 shows a representative mesurement of FQHE. The result contains IQHE though the widths of the plateaus are not prominent and rather the behavior is on the classical line. And at the positions in (9C.1), narrow plateaus are observed. On the other hand, the behavior of ρ_{xx} against the magnetic field is dramatic. Even for narrow plateaus at positions (9C.1), ρ_{xx} goes to zero or becomes very small. Hence fine and steep oscillation is observed. Even in the high magnetic field region where no IQHE is observed (filling factor $\nu < 1$), fine oscillation is observed. In particular an oscillation symmetric to $\nu = 1/2$ is observed.

FQHE is very sensitive to the electron mobility, cannot be observed in low mobility samples. In comparison with IQHE, FQHE is observed at lower temperatures with activation energy of a few K. Generally FQHE is easier to be observed at higher magnetic field.

Before going into the physics, we have a short look at the mutual electron interaction and the localization. As we saw in Sec. 10.2.2, when a 2DEG is under a strong magnetic field, the electrons at the Fermi level are in the edge mode at the equipotential lines of impurity potential. The localized state are the states going around the closed equipotential lines. The electron-electron interaction gives some fluctuation to the impurity potential and there is a possibility to lift the localization.

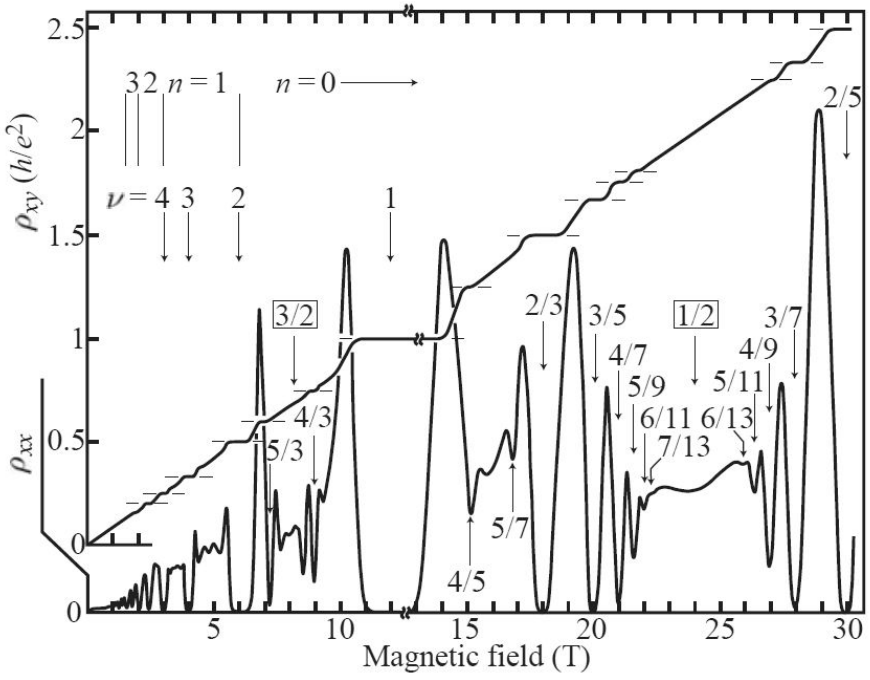


Fig. 9C.1 Representative example of FQHE measurement.

9C.2 Laughlin state

It has been clarified by long-term researches that the electronic states causing FQHE is a kind of electron liquid, in which the electron mutual interaction is dominating the many-body state. The **Laughlin state**, in spite of its simpleness, has been proven to be a good approximation to such electron liquids. This is a big event in many-body physics since the BCS theory.

We again consider a two-dimensional electron system on xy plane in the magnetic field of flux density B. Here for convenience we take the symmetric gauge $\boldsymbol{A} = (-By/2, Bx/2)$. xy-plane can be expressed as a complex plane. The spatial length is measured by the magnetic length. That is, a point on the 2DEG plane can be represented as a complex number $z = (x - iy)/l$. The Hamiltonian with the electron-electron interaction is written as

$$\mathscr{H} = \sum_j \left[\frac{1}{2m}(-i\hbar\nabla + e\boldsymbol{A})^2 + V(z) \right] + \sum_{j<k} \frac{e^2}{|z_j - z_k|}. \tag{9C.2}$$

First we make a many-body wavefunction from single-body wavefunctions at the lowest Landau level without potential and the Coulomb interaction. Then the detail of the many-body wavefunction is determined to minimize the electron interaction energy. The wavefunction which diagonalizes $X^2 + Y^2$, thus the angualr momentum is written as

$$\phi(z) = p(z) \exp\left(-\frac{|z|^2}{l^2} \right), \tag{9C.3}$$

where $p(z)$ is a polynomial of z. Let N_e be the number of electrons and the many-body wavefunction can be written as

$$\psi(z_1, \cdots, z_{N_e}) = f(z_1, \cdots, z_{N_e}) \exp\left(-\sum_i \frac{|z_i|^2}{4} \right), \tag{9C.4}$$

where a polynomial f should be anti-symmetric for the exchange in $(1, \cdots, N_e)$ due to the Pauli principle.

The general form of the terms in f is (coefficient)$\times \prod_i z_i^{m_i}$. This mathematical form indicates that in the state this term represents, the i-th electron is occupying the state with angular momentum $m_i\hbar$. Hence the total angular momentum $\hat{M}$ in this term is $\sum_i m_i\hbar$, and $\hat{M}$ commutes with $\mathscr{H}$. Because $\hat{M}$ represents a conserved quantity, it is desirable to take ψ as to diagonalize $\mathscr{H}$ and $\hat{M}$ simultaneously. For that f should be a homogeneous polynomial.

Further, to make the interaction energy smaller, we consider two-body correlation. The distance between two electrons i and j is $|z_i - z_j|$. Then we try a functional form that f is given by a product of functions g that only depend on $z_i - z_j$, that is

$$f(z_i, \cdots, z_{N_e}) = \prod_{i>j} g(z_i - z_j). \tag{9C.5}$$

From the anti-symmetric property of f, $g(z) = z^q$ and q should be an odd number. The above consideration is summarized into

$$\psi_q(z_1, \cdots, z_{N_e}) = \prod_{i>j} (z_i - z_j)^q \exp\left(-\sum_i \frac{|z_i|^2}{4} \right), \tag{9C.6}$$

which is called **Laughlin state**.

It has been clarified that various ground states exist in a two-dimensional electron system under a strong magnetic field due to strong electron-electron correlation. The Laughlin state is proposed to explain FQHE. As we can guess from the functional form, it is composed to electron interaction energy. It is known that it is close to the exact solution in the finite system obtained by using the exact diagonalization.

9C.3 Filling factor of Laughlin states

In the Laughlin state (9C.6), let us consider the polynomial in front of the exponential. The electron coordinate z_i has a maximum power of $M = (N_e - 1)$. This term of maximum power represents the state, in which the electron indexed i has the maximum angular momentum $M\hbar$. The orbit of this state spreads by l on the circle with a radius $\sqrt{2M}l$. The area corresponding to N_e Landau levels is $2\pi l^2 N_e$ and the filling factor of the state represented by the term is

$$\nu = \frac{2\pi l^2 N_e}{\pi \times 2Ml^2} = \frac{N_e}{M} = \frac{N_e}{(N_e - 1)q} \approx \frac{1}{q}. \tag{9C.7}$$

Among many terms in the polynomial, the ones with largest orbital radius are that gives the largest angular momentum to single electron. Hence the filling factor of this term becomes the filling factor of ψ_q itself. In other words, the filling factor determines q of the corresponding Laughlin state.

9C.4 Excited states

Next we consider the excitation from Laughlin state (9C.6). For that we write the state with increased angular momentum by one for each electron as $\prod_i z_i \psi_q$.

$$\prod_i z_i \psi_q = \prod_i z_i \sum A_{m_1,m_2,\cdots} z_1^{m_1} z_2^{m_2} \cdots z_{N_e}^{m_{N_e}} \exp\left(-\sum_j \frac{|z_j|^2}{4}\right) \tag{9C.8}$$

$$= \sum A_{m_1,m_2,\cdots} z_1^{m_1+1} z_2^{m_2+1} \cdots z_{N_e}^{m_{N_e}+1} \exp\left(-\sum_j \frac{|z_j|^2}{4}\right). \tag{9C.9}$$

The operation of taking the product with $\prod_i z_i$ increases the angular momentum of each electron and at the same time introduces a zero at the origin [*4]. Around the zero, the amplitude of the wavefunction is small with the scale of l and the negative charge density decreases, which can be viewed as a positive charge around the zero. This can be treated as a **quasiparticle**.

We first take the product with $\prod_k (z_k - z_0)^q$, which introduces q quasiparticles at a point z_0. Now we put an electron with spatial size of l at z_0. Then the wavefunction is

$$\prod_k (z_k - z_0)^q \prod_{i<j} (z_i - z_j)^q \exp\left(-\sum_l \frac{|z_l|^2}{4} - \frac{|z_0|^2}{4}\right). \tag{9C.10}$$

This is nothing but a uniform Laughlin state with the electron number increased by one. The above operation means q quasiparticles with a positive charge and an electron with the charge $-e$ are canceled out. This indicates we can cosider that the charge of a quasiparticle is e/q.

9C.5 Composite fermion picture

In a Laughlin state ($\nu = 1/q$), the electrons avoid each other and if we keep our eyes on a single electron, it looks as if a single electron is in a uniform magnetic field. In the $\nu = 1$ Landau level, single a quantum flux Φ_0 is going through the area of a single electron. In the case of Laughlin state the number of quantum magnetic flux per an electron is q. Let us consider such an electron as a "particle" with an even number ($2k$) of quantum flux. Such a particle obeys, if one goes back to Laughlin wavefunction, the Fermi statistices, hence they are called **composite fermion** (CF)[8]. The magnetic field such CFs feels is that of $q - 2k$ times quantum flux.

[*4] With $\prod_i (z_i - z_0)$ zero can be introduced any point z_0.

That is, the field of $q - 2k = 1$ can be seen as IQHE state of $n = 1$ for CFs, where n is the filling factor of CFs. Similarly in the case of $1/(q - 2k) = n > 1$, IQHE of CFs appears for integer n. Because they express extended state of CFs, the electron wavefunction is also extended. There, the filling factor ν of the electrons is

$$\nu = \frac{1}{q} = \frac{1}{2k + 1/n} = \frac{n}{2kn + 1}. \tag{9C.11}$$

For $k = 1$, this gives an FQHE series of 2/5, 3/7, 4/9, $\cdots$, which is comparatively easy to be observed. Taking these states as the starting states, we can explain the next generation of FQH states. The above indicates that the FQHE of electron can be interpreted as IQHE of CFs. ρ_{xx} looks symmetric to $\nu = 1/2$ and the oscillation can be interpreted as SdH oscillation of CFs.

References

[1] K. von Klitzing, G. Dorda, and M. Pepper, Phys. Rev. Lett. **45**, 494 (1980).

[2] Z. F. Ezawa, "Quantum Hall Effects: Recent Theoretical and Experimental Developements", (World Scientifric, 2013).

[3] M. Büttiker, Phys. Rev. B **38**, 9375 (1988).

[4] M. Kohmoto, Ann. Phys. **160**, 343 (1985).

[5] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs, Phys. Rev. Lett. **49**, 405 (1982).

[6] R. B. Laughlin, Phys. Rev. B **23**, 5632 (1981).

[7] D. J. Thouless, "Topological Quantum Number in Nonrelativistic Physics" (World Scientific, 1998).

[8] J. Jain, "Composite Fermions" (Cambridge, 2007).

Lecture note on Physics of Semiconductors (14)
14th July (2021) Shingo Katsumoto, Institute for Solid State Physics, University of Tokyo

Chapter 10 Spintronics

This is the final lecture of "Semiconductors." I would like to close the lecture with a talk on semiconductor spintronics. Though the lecture time is limited, beginning with semi-classical spintronics, spin-orbit stpintronics then finally we would like to consider the spin Hall effect and the topological insulators.

10.1 Classical treatment of spin transport

10.1.1 What is spintronics?

So far we have seen tranport phenomena in semiconductors from classical to quantum mechanical. An electron has a charge and simultaneously a spin, associated magnetic moment and angular momentum. Hence motion of an electron is associated not only with the charge but with the spin angular momentum. However, due to Kramers degeneracy, spin angular momentums cancel out each other and the microscopic transport of spin angular momentum does not appear in the net current. Since electrons are charge monopoles, electronics have made great progress by controlling their flow and accumulation, and semiconductors, which make this possible, have played a central role. However, now that the down-sizing, speeding up, and lowering of energy consumption are reaching their limits, spintronics is to use spin, which is the internal degree of freedom of electrons, for information storage and manipulation[1]. For a long time, magnetic disks and tapes, which are examples of spins frozen by the many-body effect, have been used for information storage. Giant magnetoresistance (GMR) devices, which utilize magnetic multilayers or spin-valve structure have been used for the readout of such information from 1990's. Since then, the word "spintronics" has been gradually used.

The reason why semiconductors were the center of electronics is, paradoxically, because they are insulators. It is most important that there are no electrons in an undisturbed crystalline state, and it is several times more difficult to cut off current in a metal with good control than to introduce conduction in an insulator, and in metal electronics. The role of metals in electronics is limited to wiring [*1] . On the other hand in spintronics, there is no net spin in normal metals in equilibrium. They can be considered as "vacuum" for spins. In a sense, normal metals are semiconductors in spintronics, substances to be treated in semiconductor physics.

The above is by no means a play of words. In electronics, there is almost no electric field, that is, a slope of chemical potential in metals due to its short screening lengths. On the other hand, when an electric current is passed through an interface between a normal metal and a ferromagnet, the chemical potential is separated by spin. This is similar to the situation that in semiconductors under minority carrier injection in that different quasi Fermi levels are associated with electrons and holes. Also, even inside metals, the spin current causes gradients in chemical potentials that can be detected by external circuits. Therefore various physics we have seen in semiconductors may be observed as spintronics in normal metals in modified forms.

[*1] However, the establishment of the concept of electronic circuits itself depends on the existence of a substance called metal. For the detail, see the lecture note of the present lecturer on electronics (in Japanese).

10.1.2 Two-current model

In the **two-current model** proposed by Nevil Mott, the net electric current by electrons is devided into the portions by up-spin (↑) and down-spin (↓) electrons. The model holds when the scattering time of spin flip is sufficiently longer than those by various mechanisms (Sec. 5.1.5) dominating conduction. The difference in the resisitivities for ↑ and ↓ is due to the difference in parameters like density of states, $k_{\rm F}$, etc. The total resistivity ρ is expressed by those for spin subband channels $\rho_\uparrow$, $\rho_\downarrow$, as $1/\rho = 1/\rho_\uparrow + 1/\rho_\downarrow$. In diffusive conductors, regardless of magnetic or non-magnetic, the spin diffusion length is generally longer than the mean free path ($\lambda_{\rm F} \gg l_{\rm F}$), hence the two current model is considered to work well. On the other hand, in ballistic transport, particularly in the presence of strong spin-orbit interaction, the two-current model meets difficulty. In this section we treat classical transport with no or weak spin-orbit interaction. Hence the discussion is on the two-current model.

Assuming metallic conductors and the two-current model, we apply the Drude conductance to each spin-subband, which is given by $\sigma_s = e^2 n_s \tau_s / m_s^*$ ($s = \uparrow, \downarrow$). The net electric current density $\boldsymbol{j}_{\rm c}$ is given by $\boldsymbol{j}_\uparrow + \boldsymbol{j}_\downarrow$ while the spin-polarized portion is given by the difference $\boldsymbol{j}_{{\rm p}\uparrow} = \boldsymbol{j}_\uparrow - \boldsymbol{j}_\downarrow$. The polarization of **spin-polarized current** density is definced as

$$P_{\rm c} = \frac{|\boldsymbol{j}_\uparrow - \boldsymbol{j}_\downarrow|}{|\boldsymbol{j}_\uparrow + \boldsymbol{j}_\downarrow|} = \frac{j_{{\rm p}\uparrow(\downarrow)}}{j_{\rm c}}. \tag{10.1}$$

Each component of the current is further devided into the drift term and the diffusion term as

$$\boldsymbol{j}_{{\rm p}s} = \boldsymbol{\sigma}_s \boldsymbol{E} - eD_s(-\nabla \delta n_s). \tag{10.2}$$

10.1.3 Spin-dependent electrochemical potential

Even when there is nonequilibrium between spin subbands by spin injection(or emission), providing that intra-subband scattering is sufficiently frequent and local equilibrium inside each spin band is kept, we can define local Fermi energy ϵ_s and shift from the equilibrium $\delta\epsilon_s$ for each spin-subband. For simplicity, a scaler σ_s is assumed for conductance tensor $\boldsymbol{\sigma}_s$. We write the electrostatic potential as ϕ ($\boldsymbol{E} = -\nabla\phi$). The Einstein relation $\sigma_s = e^2 N_s(E_{\rm F}) D_s$ (this experssion is low-temperature, metallic version of Eq. (5.13)), $\delta n_s = N_s(E_{\rm F})\delta\epsilon_s$ gives the following.

$$\boldsymbol{j}_{{\rm p}s} = -\frac{\sigma_s}{e}\left[e\nabla\phi - \frac{D_s \nabla \delta n_s}{\sigma_s}\right] = \frac{\sigma_s}{e}[-e\nabla\phi + \nabla\delta\epsilon_s]. \tag{10.3}$$

In the two-current model, the local electrochemical potential for each spin can be defined as

$$\mu_s = -e\phi + \epsilon_s, \tag{10.4}$$

which leads to the expression of current density in each spin-subband

$$\boldsymbol{j}_{{\rm p}s} = -\frac{\sigma_s}{-e}\nabla\mu_s. \tag{10.5}$$

Below we write electrochemical potential simply as “chemical potential.”

10.1.4 Spin current

There are several ways to defince **spin current**, namely flow of spin angular mometum. A representative one is

$$\boldsymbol{j}^s(\boldsymbol{r}, t) = \frac{\hbar}{2(-e)}(\boldsymbol{j}_\uparrow - \boldsymbol{j}_\downarrow). \tag{10.6}$$

A spin current generally is a tensor formed by local spin density vector and flow vector. In the above for simplicity, we consider a spin current as a flow of z(direction of magnetization)-component of spin angular momentum. Also more generally, there is a type of spin current in which a flow of spin angular momentum is mediated by exchange interaction (e.g. by spin-wave).

With writing local spin angular momentum density as $\boldsymbol{s}(\boldsymbol{r},t)$, and the z component as s_z, spin angular momentum conservation law is written by

$$\frac{\partial s_z}{\partial t} + \mathrm{div}\, \boldsymbol{j}^s = 0. \tag{10.7}$$

In the presence of spin relaxation, we need to consider the relaxation term in the right hand side of Eq. (10.7). Within the relaxation time approximation, we can write

$$\frac{\partial s_z}{\partial t} + \mathrm{div}\, \boldsymbol{j}^s = \frac{\partial s_z}{\partial t} + \frac{\hbar}{2(-e)} \nabla \cdot (\boldsymbol{j}_\uparrow - \boldsymbol{j}_\downarrow) = \frac{\hbar}{2} \left(\frac{\delta n_\uparrow}{\tau_\uparrow} - \frac{\delta n_\downarrow}{\tau_\downarrow} \right). \tag{10.8}$$

On the other hand, the charge (ρ) conservation law is given by

$$\frac{\partial \rho}{\partial t} + \mathrm{div} \boldsymbol{j} = \frac{\partial \rho}{\partial t} + \nabla \cdot (\boldsymbol{j}_\uparrow + \boldsymbol{j}_\downarrow) = 0. \tag{10.9}$$

In a steady state, $\partial\rho/\partial t = \partial s_z/\partial t = 0$. From the constraint that there should be no total spin flip in the whole system the relaxation times $\tau_\uparrow$ and $\tau_\downarrow$ should fulfill the detailed balance condition:

$$N_\uparrow \tau_\downarrow = N_\downarrow \tau_\uparrow, \tag{10.10}$$

where $N_{\uparrow,\downarrow}$ are the spin-dependent density of states at the Fermi level. The above and Eqs. (10.8), (10.9) lead to

$$\nabla^2 (\sigma_\uparrow \mu_\uparrow + \sigma_\downarrow \mu_\downarrow) = 0, \tag{10.11a}$$

$$\nabla^2 (\mu_\uparrow - \mu_\downarrow) = \frac{1}{(\lambda_{\mathrm{sf}}^{\mathrm{F}})^2} (\mu_\uparrow - \mu_\downarrow). \tag{10.11b}$$

Averaged spin diffusion length $\lambda_{\mathrm{sf}}^{\mathrm{F}}$ is defined from the Matthiessen's law as $(\lambda_{\mathrm{sf}}^{\mathrm{F}})^{-2} = (\lambda_\uparrow^{\mathrm{F}})^{-2} + (\lambda_\downarrow^{\mathrm{F}})^{-2}$, where $\lambda_\uparrow^{\mathrm{F}}$, $\lambda_\downarrow^{\mathrm{F}}$ are spin diffusion lengths for up and down spins respectively. Equation (10.11b) takes the form of diffusion equation, thus called **spin diffusion equation**.

10.2 Spin injection and relaxation

There are various methods of injecting spins into a paramagnetic material, similar to injecting minority carriers into a semiconductor by irradiating light or applying a forward bias to a pn junction. Here, in particular, spin injection from a ferromagnet to a paramagnetic material is described. As with minority carriers, spin injection occurs at the interface and spreads into the bulk and disappears (relaxes).

Let us consider an interface between a ferromagnet (FM) and a normal metal (NM) with a current j_{c} perpendicular to the interface. We write down the spin dependent local chemical potentials (Eq. (10.5)) for FM and NM regions in the following form:

$$\mu_s^{\mathrm{M}} = a^{\mathrm{M}} + b^{\mathrm{M}} x \pm \frac{c^{\mathrm{M}}}{\sigma_s^{\mathrm{M}}} \exp\left(\frac{x}{\lambda_{\mathrm{sf}}^{\mathrm{M}}} \right) \pm \frac{d^{\mathrm{M}}}{\sigma_s^{\mathrm{M}}} \exp\left(-\frac{x}{\lambda_{\mathrm{sf}}^{\mathrm{M}}} \right), \tag{10.12}$$

where M is F(ferromagnet) or N(normal metal). x-axis is taken to be perpendicular to the interface in the direction to the normal metal, and the origin is taken at the interface. In the double sign $\pm$, $+$ corresponds to $\uparrow$, $-$ does $\downarrow$. The sum of the first two terms in the r.h.s. is written as μ_0, which does not depend on spin. The function form of the third and the fourth terms comes from the fact that the spin-dependent parts should obey the diffusion equation (10.11). It is straightforward to confirm that the expression in Eq. (10.12) satisfies Eq. (10.11).

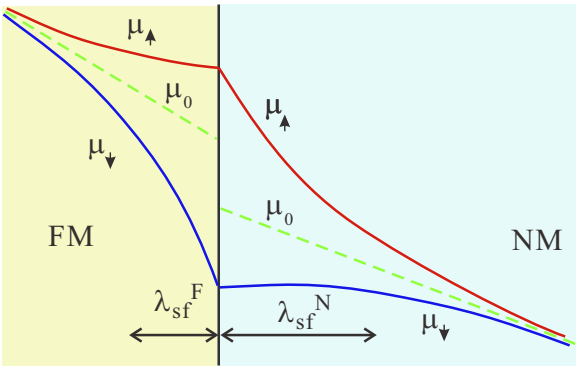


Fig. 10.1 Sketch of spatial variation of spin-dependent chemical potential at an FM-NM interface, through which electrons flow from the FM side.

Coefficient $a \sim d$ is determined as follows. Under the assumption of two-current model, the local chemical potential μ_s for each spin subband must be continuous at the interface, i.e. $\mu_s^{\mathrm{F}}(-0) = \mu_s^{\mathrm{N}}(+0)$. μ_0 can be discontinuous at the interface for non-equilibrium states while in $|x| \to \infty$ both in FM and NM the difference between $\mu_\uparrow$ and $\mu_\downarrow$ approaches zero. This means $d^{\mathrm{F}} = 0$ and $c^{\mathrm{N}} = 0$. Also the sum of current densities in spin-subband should be the total current density j_c.

From the above, the density of states and the spin polarization in F P_{F}, μ_s^{M} is given by

$$\mu_s^{\mathrm{F}} = \frac{(-e)j_c}{\sigma^{\mathrm{F}}}x \mp \frac{(-e)j_c P_{\mathrm{F}} \lambda_{\mathrm{sf}}^{\mathrm{N}}(1-P_{\mathrm{F}}^2)\sigma^{\mathrm{F}}}{2\sigma_s^{\mathrm{F}}\sigma^{\mathrm{N}}\left[1+(1-P_{\mathrm{F}}^2)\dfrac{\sigma^{\mathrm{F}}\lambda_{\mathrm{sf}}^{\mathrm{N}}}{\sigma^{\mathrm{N}}\lambda_{\mathrm{sf}}^{\mathrm{F}}}\right]} \exp\left(\frac{x}{\lambda_{\mathrm{sf}}^{\mathrm{F}}}\right), \tag{10.13a}$$

$$\mu_s^{\mathrm{N}} = \frac{(-e)j_c}{\sigma^{\mathrm{N}}}x + \frac{(-e)j_c P_{\mathrm{F}} \lambda_{\mathrm{sf}}^{\mathrm{N}}}{\sigma^{\mathrm{N}}\left[1+(1-P_{\mathrm{F}}^2)\dfrac{\sigma^{\mathrm{F}}\lambda_{\mathrm{sf}}^{\mathrm{N}}}{\sigma^{\mathrm{N}}\lambda_{\mathrm{sf}}^{\mathrm{F}}}\right]} \left[1 \mp \exp\left(-\frac{x}{\lambda_{\mathrm{sf}}^{\mathrm{N}}}\right)\right]. \tag{10.13b}$$

In the comlex symbol $+$, $-$ correspond to $\uparrow$, $\downarrow$ respectively. The origin of energy is taken to the chemical potential in equilibrium ($j_c = 0$). A schematic diagram is given in Fig. 10.1(a).

10.2.1 Spin injection and detection

Semiconductors are appropriate for the control of electric conduction because they are insulators in their intrisic states. Similarly, non-magnetic materials are appropriate for the control and operation of spin current. For that, however, spin injection into non-magnetic materials just like minority carrier injection in pn junctions, etc. In the configuration in

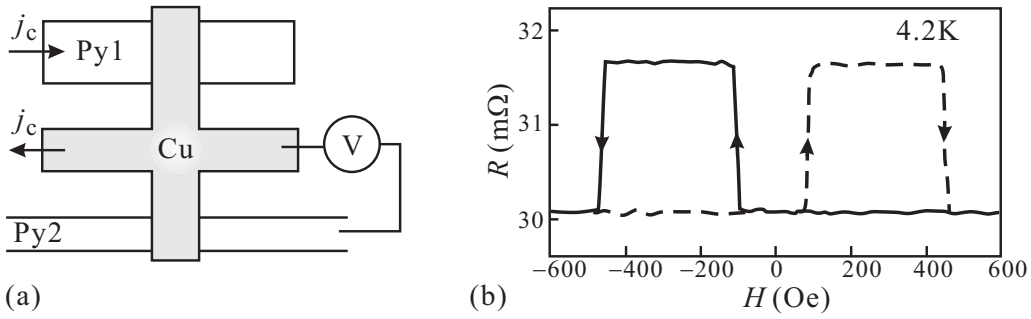


Fig. 10.2 (a) Configuration of probes and circuit for non-local detection of spin injection. The electric current j_c goes through the center of the cross made of Cu to the left terminal while the spin current injected from Py1 reaches Py2, causes separation of $\mu_\uparrow$ and $\mu_\downarrow$ hence a step in μ_0 at the N-F (Cu-Py2) interface, which is detected as a voltage V. (b) Thus measured non-local resistance. The spin-valve like magnetoresistance comes from the difference in the coercive force between Py1 and Py2 due to the difference in the shape. The data are from [2].

Fig. 10.1, spin current and electric current are overlapped and the electric separation of the two effects is difficult. Therefore in many of experiments on spin injection, non-local configuration of electrodes is adopted.

An example is shown in Fig. 10.2. Current j_c through permalloy (alloy of Fe-Ni, Py)1 and the Cu sample causes separation of $\mu_\uparrow$ and $\mu_\downarrow$, which means a spin accumulation at the interfaces. Though no electric current flows between Py2 and Cu, the spin diffusion occurs independently. The spin current thus also flows to Py2 and there causes difference in chemical potential, which is detected as a voltage. From Eq. (10.13) and the spin diffusion equation (10.11b), the detected voltage is given by

$$V = \pm\frac{1}{2} e j_c P_{\rm Py}^2 \frac{\rho^{\rm Py}\rho^{\rm Cu}}{\rho^{\rm Py}+\rho^{\rm Cu}} \exp\left(-\frac{L}{\lambda_{\rm sf}^{\rm Cu}}\right). \tag{10.14}$$

Figure Fig. 10.2(b) shows the result of non-local measurement. Due to the difference in widths of Py1 and Py2, the coercive forces for the magnetic field along the strips are different, which results in the spin-valve like non-local magnetoresistance. The analysis of experimental results with various parameters by Eq. (10.14), material parameters like $P_{\rm F}$ and $\lambda_{\rm sf}$ can be obtained.

Figure Fig. 10.3(a) shows a schematic view of non-local four-terminal probe configuration, which is often adopted for the detection of spin injection into semiconductors. An electric current is applied between the left two electrodes including a ferromagnet, and the difference in the electrochemical potentials of the two electrodes also including a ferromagnet in the right is measured as a voltage. In such needle-shaped thin film electrodes, due to magnetic anisotropy from shape, the magnetic field is applied along the needles. On the other hand, a magnetic field perpendicular to the injected spins causes precession of spin magnetic moment (Appendix 10A). If the electron spins rotates in perfect coherence and the starting angle is also synchronized, the detected voltage oscillates reflecting the precession. In diffusion process, the distance of migration largely distributes and in real samples the region of injection has a finite width and the oscillation decays with the progress of precession. This is called Hanle effect.

If the problem is restricted to one dimensional spin diffusion along x-axis in a normal metal, spin dependence in σ_s, D_s and τ_s can be dropped in Sec. 10.1. In Eq. (10.2), since only the diffusion term is effective for non-local effect, the drift term is dropped. Applying the relaxation time approximation (10.8), we get the spin diffusion equation:

$$\frac{\partial s_z}{\partial t} = D\frac{\partial^2 s_z}{\partial x^2} - \frac{s_z}{\tau_{\rm sf}}. \tag{10.15}$$

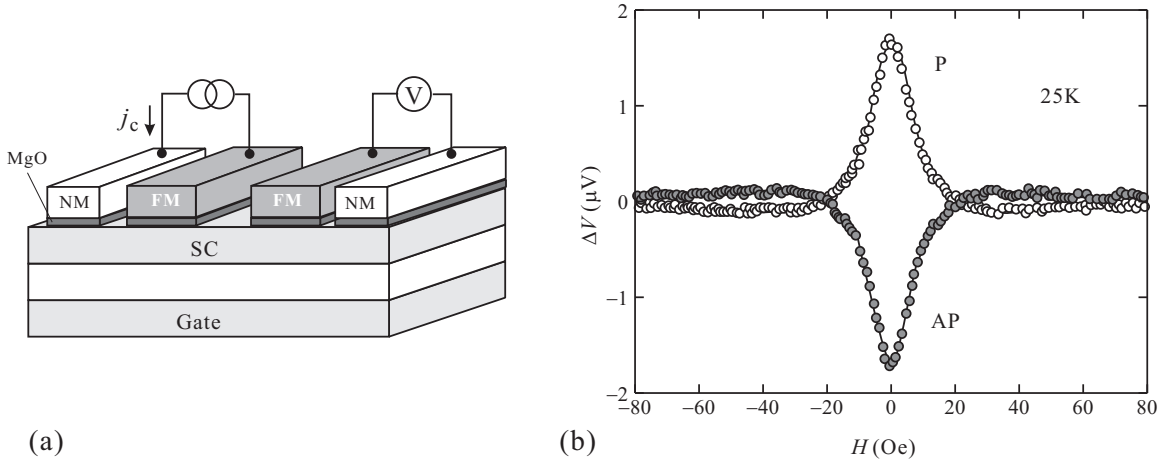


Fig. 10.3 (a) Four terminal probe configuration for detection of spin injection into semiconductors (SCs) with non-local resistance. Current j_c between left pair of ferromagnet (FM) and normal metal (NM) causes spin current to the right, which is detected as the chemical potential difference (voltage). MgO for the potential barrier is used for spin-injection with high efficiency. (b) Hanle signal measured in a similar structure as in (a). The SC here is Si and the spin rotation is caused by the magnetic field perpendicular to the current plane. From Ref. [3].

This leads to the expression of Hanle signal:

$$\begin{aligned} \Delta V &= \pm \frac{j_c P_j^2}{e^2 N_{\rm SC}} \int_0^\infty dt \varphi(t) \cos \omega t, \\ \varphi(t) &= \frac{1}{\sqrt{4\pi D t}} \exp\left(-\frac{d^2}{4Dt}\right) \exp\left(-\frac{t}{\tau_{\rm sf}}\right), \end{aligned} \tag{10.16}$$

where d is the distance between the injection and detection electrodes, P_j is the spin polarization just below the injection electrode, $\omega = g\mu_{\rm B}B/\hbar$ is the Larmor frequency (Appendix 10A).

Figure Fig. 10.3(b) shows a Hanle signal measured in an experiment, in which spins were injected from an Fe electrode into Si. The signal can be fitted by (10.16). The fitting provides the parameters $\lambda_{\rm sf}$, etc.

10.3 Spin-orbit interaction

Henceforth we go into semiconductor spintronics. Here "Semiconductor" refers to materials defined by the charge degrees of freedom that we have seen in the previous chapters. The spin-orbit interaction (SOI) has been already introduced when we had a look on k·p perturbattion perticularly for fcc-type semiconductors. In spintronics, the SOI is very important since it connects spin and orbital degree of freedoms. Below we see an example that the SOI becomes important even in the conduction bands through the mixing with the valence bands, in which the SOI is strong in intrinsic bulk states. More specifically, we consider two type of spin-orbit interactions in two-dimensional electron systems.

10.3.1 Spin-orbit splitting due to bulk and structural inversion assymetries

In crystals when the lattice has the spatial inversion symmetry, the states of $\boldsymbol{k}$ and $-\boldsymbol{k}$ are degenerate. For example in the case of $\uparrow$-spin states, $E(\boldsymbol{k},\uparrow) = E(-\boldsymbol{k},\uparrow)$. The time-inversion operation causes $-\boldsymbol{k} \to \boldsymbol{k}$ and simultaneously inversion in spin. When the crystals have time-inversion symmetry not having magnetism and the spatial inversion symmetry, $E(\boldsymbol{k},\uparrow) = E(\boldsymbol{k},\downarrow)$. In other words, in order for a system to have finite spin-splitting at finite $\boldsymbol{k}$, the system should have some inversion asymmetry.

In the primitive cell of zinc-blende type crystals, it is apparent that there is spatial inversion asymmetry along [111], which gives spin-splittings in energy dispersion due to the SOI. Such inversion asymmetries arising from crystal structures is called **bulk inversion asymmetry** (**BIA**). SOIs caused by BIA is called **Dresselhaus spin-orbit interaction**. The Dresselhaus SOI is obtained from $k \cdot p$ perturbation that takes BIA into account[4, 5]. BIA is some variance for $\boldsymbol{k} \to -\boldsymbol{k}$, and then the interaction term should be odd order in $\boldsymbol{k}$. In the caes of Dresselhaus interaction in three-dimensional systems is in 3rd order. The form of the Hamiltonian is

$$\mathscr{H}_{\rm DSO}^{\rm 3d} = \gamma\hbar^2[k_x(k_y^2 - k_z^2)\sigma_x + k_y(k_z^2 - k_x^2)\sigma_y + k_z(k_x^2 - k_y^2)\sigma_z], \tag{10.17}$$

where xyz is [100], [010], and [001]. When a two dimensional electron system (2DES) is formed on (001) surface, due to the averaging of kinetic freedom in z-direction ([001]), a k-linear term appears.

$$\begin{aligned} \mathscr{H}_{\rm DSO}^{\rm 2d} &= \gamma\hbar^2[k_x(k_y^2 - \langle k_z^2\rangle)\sigma_x + k_y(\langle k_z^2\rangle - k_x^2)\sigma_y] \\ &= \beta(k_y\sigma_y - k_x\sigma_x) + \gamma\hbar^2(k_x k_y^2\sigma_x - k_y k_x^2\sigma_y). \end{aligned} \tag{10.18}$$

While the BIA is from the crystal structure, the inversion symmetry is broken by introducing some structures that break the periodicity like a hetro-interface. This is called **structure inversion asymmetry** (**SIA**). The SOI introcuded from the SIA at an interface is called **Rashba spin-orbit interaction** [6, 7]. The Rashba-type interaction Hamiltonican is written in the form:

$$\mathscr{H}_{\rm RSO} = \alpha\boldsymbol{\sigma} \cdot (\boldsymbol{k} \times \boldsymbol{e}_z) = \alpha(k_y\sigma_x - k_x\sigma_y). \tag{10.19}$$

This comes from the expression Eq. (2.56). Here ∇V (in (2.56)) is an electric-field like term perpendicular to the 2DES plane (z-direction) [*2]. We need to be careful for "electric-field like term ∇V." Let us assume that this $\nabla V/e$ is really an electric field for a while. Here, ∇V is the total (including the one from the band discontinuity) force, which confines conduction electrons into the two-dimensional plane. The fact that the electrons are confined into the two-dimensional plane means that the averaged expectation value of ∇V is zero due to the Ehrenfest theorem on the motion in z-direction, i.e. $\langle \nabla V \rangle = 0$[8]. Therefore, the Rashba interaction cannot be introduced from a real electric field $\nabla V/e$.

As ssen in the k·p approximation (Sec. 2.2.6-7), the SOI comes from the mixing of valence-band wavefunction into the conduction-band wavefunction at finite $\boldsymbol{k}$. There, we introduced a potential V (Eq. (2.56)) common for conduction and valence bands. If there is a difference in their band discontinuities, we need to consider different V's for them, and if $\langle (\nabla V)_z \rangle = 0$ in the conduction band, $(\nabla V)_z \rangle \neq 0$ in the valence band, then the Rashba interaction survives with that. Along the above line, we write the potentials in conduction and valence bands as V_c and V_v respectively and derive the expression for the SOI. Because $\langle (\nabla V_c)_z \rangle = 0$, ∇V term is replaced by ∇V_v. In experiments, many phenomena characteristic to the Rashba SOI have been found, e.g. in a 2DES of a narrow gap semiconductor InGaAs. They are considered to arise from the valence band.

We consider a 2DES which only has k-linear terms in the Rashba SOI (10.19) and the Dresselhaus SOI (10.18). We take a plane wave with wavenumber $\boldsymbol{k} = (k\cos\varphi, k\sin\varphi)$ as the orbital part to write the SOI Hamiltonian as follows.

$$\begin{aligned}\mathscr{H}_{\rm SO} &= \alpha \begin{pmatrix} 0 & -i\hat{k}_x + \hat{k}_y \\ i\hat{k}_x + \hat{k}_y & 0 \end{pmatrix} + \beta \begin{pmatrix} 0 & -\hat{k}_x - i\hat{k}_y \\ -\hat{k}_x + i\hat{k}_y & 0 \end{pmatrix} \\ &= \alpha k \begin{pmatrix} 0 & ie^{-i\varphi} \\ -ie^{i\varphi} & 0 \end{pmatrix} - \beta k \begin{pmatrix} 0 & e^{i\varphi} \\ e^{-i\varphi} & 0 \end{pmatrix} .\end{aligned}$$

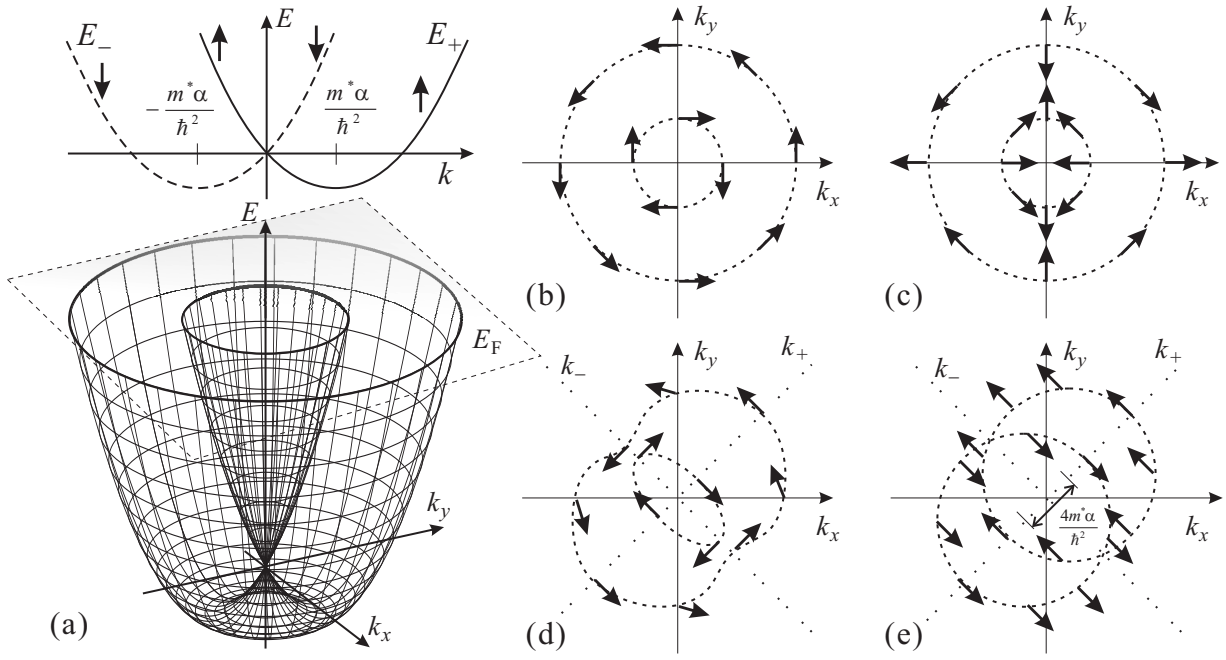


Fig. 10.4 (a) Upper: Energy dispersion relation (10.20) in the presence of the Rashba SOI. The spin at $+\pi/2$ is taken as "up." Lower: Three dimensional wireframe expression of the energy dispersion on $k_x - k_y$ plane. (b)∼(e) are cross sections at $E = E_{\rm F}$. (b) Two Fermi circles and the direction of effective field (spins) in the case of $\beta = 0$ (Rashba model). (c) The same as (b) in the case of $\alpha = 0$ (Dresselhaus model). (d) α and β are finite, but $\alpha \neq \beta$. (e) The case of $\alpha = \beta$.

[*2] From this expression, one often falls into the following incorrect explanation. – The existence of electric field means creation of charges both upper and lower sides of 2DES. From an electron running in 2DES those charges cause a loop current enclosing the 2DES. The loop current results in a magnetic field, which is nothing but the Rashba effective field. – For the reason why this is wrong, see the text.

In the case of $\beta = 0$ (Rashba model), writing the spin part as ${}^t(1, e^{i\phi})/\sqrt{2}$, we get $\phi = \varphi \pm \pi/2$ from the condition of eigenfunction. Namely the wavevector and the spin are orthogonal. The eigenenergy $E_\pm$ corresponding to $\pm\pi/2$ are obtained in the effective mass approximation as

$$E_\pm = \frac{\hbar^2 k^2}{2m^*} \mp \alpha k = \frac{\hbar^2}{2m^*}\left(k \mp \frac{m^*\alpha}{\hbar^2}\right)^2 - \frac{m^*}{2\hbar^2}\alpha^2. \tag{10.20}$$

Equation (10.20) indicates that the dispersion shifts in k-space depending on the direction of spin. The dispersion is described as two spin-dependent parabolas as drawn in Fig. 10.4(a), and in Fig. 10.4(b) in a three-dimensional view. If we cut the dispersion at $E = E_\mathrm{F}$, two cocentric Fermi circles appear. The direction of spin on the Fermi circles indicated in Fig. 10.4(b) rotates inversely to each other. Similarly for $\alpha = 0$, $\phi = -\varphi,\ -\varphi + \pi$ and the energy dispertion is in the same form as Eq. (10.20) but with replacing α with β. However, a rotation of $\boldsymbol{k}$ on the Fermi circles causes an inverse rotation of spin as sketched in Fig. 10.4(c).

Under the coexistence of α and β, generally the spin and the dispersion show complicated forms as in Fig. 10.4(d). In the special case of $\alpha = \beta$, the plane wave and the spin are separated as $\mathscr{H}_\mathrm{SO} = \alpha(\hat{k}_x + \hat{k}_y)(\sigma_x - \sigma_y)$. With rotating the wave-vector part and the spin part as $k_\pm = \frac{k_y \pm k_x}{\sqrt{2}}$, $\chi_\pm = \pm{}^t(1/\sqrt{2}, (i-1)/2)$, the Hamiltonian is expressed as

$$\mathscr{H} = \frac{\hbar^2}{2m^*}(\hat{k}_+^2 + \hat{k}_-^2) - 2\alpha\hat{k}_+\sigma_z'. \tag{10.21}$$

σ_z' is a Pauli matrix on the basis of $\chi_\pm$. Since the wave-vectors and the spins are separated, once the eigenfunction of the spin part $\chi_\pm$ is determined, the dispersion is ordinary parabola whoes center shifts by $\pm 2m^*\alpha/\hbar^2$ depending on χ_+, χ_- Therefore, as in Fig. 10.4(e), the centers of the spin-dependent two parabolas shift to each other. The two center-shifted Fermi circles have a partial overlap.

10.3.2 Spin-orbit interaction and SdH oscillation

In a 2DES where the Rashba interaction is strong and β can be ignored, two Fermi cirlcles with different k_F exist as in Fig. 10.4(b). This leads to spin-dependent 2DES sheet density $n_{s\sigma} = k_{\mathrm{F}\sigma}^2/(4\pi)$. The SdH oscillation in such a system should have two different periods in $1/B$ plot as in Eq. (9.14) resulting in the beating among the two frequencies. The difference in the size of the Fermi circles is proportional to α, thus to $\langle \nabla V_{vz} \rangle$. Therefore the frequency of beatng should change with applying the external electric field to cause the change in $\langle V_v \rangle$.

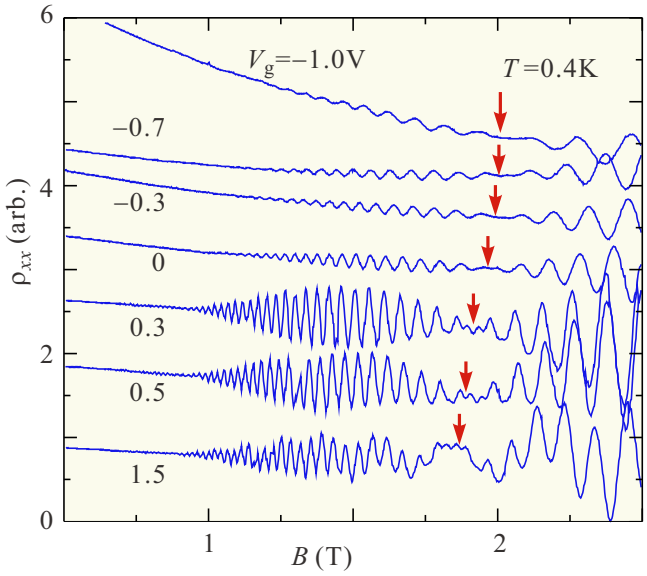


Fig. 10.5 SdH oscillation observed in a 2DES at a quantum well of $In_{0.53}Ga_{0.47}As$. The parameter is the gate voltage V_g applied onto the surface of 2DES. The arrows indicate the position of nodes in the beats. From [9].

Figure 10.5(a) shows SdH oscillations measured in 2DES at a quantum well of (In,Ga)As grown on an InP substrate. Although there is a large lattice mismatch between InAs (narrow gap) and GaAs (relatively large gap), lattice matching is performed on the InP substrate by mixing crystals and setting the In composition to 0.53. In addition, heterojunctions can be formed by adding Al to the mixed crystal. A clear beat appears in the SdH oscillation, and the position of the node indicated by the arrow shifts due to the gate voltage, which is expected for the Rashba SOI.

10.4 Spin Hall effect and topological insulator

10.4.1 Spin Hall effect

When an electric field is applied to an electron system with SOI, a spin current is drived in the direction perpendicular to the field. This phenomenon is called spin Hall effect. Let J_{ij} be the spin current tensor with spin coordinate index i, flow coordinate index j. For the external electric field $\boldsymbol{E}$, J_{ij} is written as

$$J_{ij} = \sigma_s \sum_k \epsilon_{ijk} E_k, \tag{10.22}$$

where ϵ_{ijk} is the completely antisymmatric tensor indicating the mutual orthogonality of spin, flow vector of spin current and electric field. σ_s is called spin Hall conductance. The spin Hall effect arises from impurity scattering, or SOI from special orbital motion due to the band structure. The former is called extrinsic spin Hall effect while the latter is called intrinsic spin Hall effect.

Due to space limitation here I just introduce an example of experiment. Figure 10.6 shows an experiment, in which a spin accumulation at edges of an n-type GaAs sample is detected by the difference in chemical potentials of ferromagnetic electrodes. A clear spin signal which reverts with current reversal is detected (no signal for anti-parallel magnetization configuration). The signal is due to the spin Hall effect. From the temperature dependence, it is concluded that the origin is the extrinsic effect.

10.4.2 Anomalous velocity and spin Hall effect

Let us consider the motion of a wavepacket in a crystal. The wavepacket is expanded by Bloch functions and the effect of external force $\boldsymbol{F} = -e\mathcal{E}$ from electric field $\mathcal{E}$ on each Bloch component is examined. We introduce "Bloch

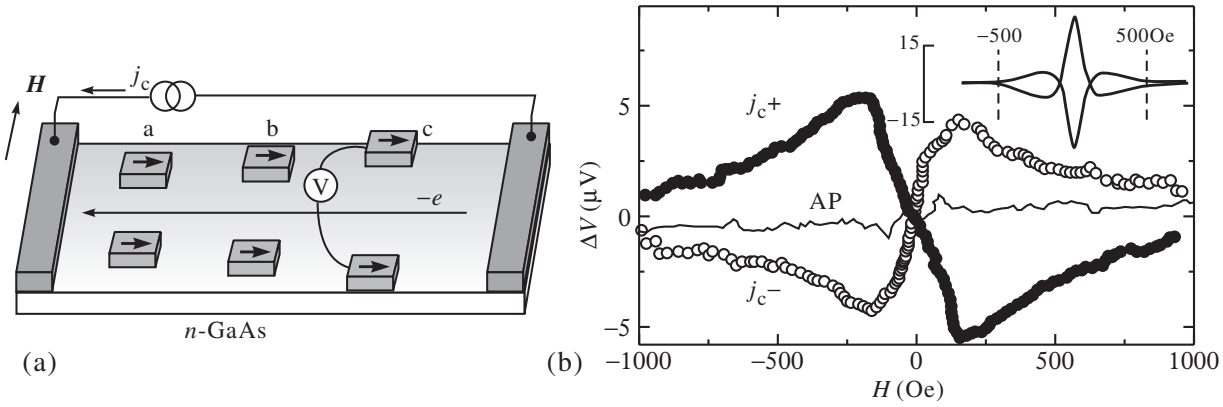


Fig. 10.6 (a) Probe configuration of an n-type GaAs sample. Spin accumulation at the sample edges is detected by the Fe electrode pairs placed perpendicular to the current. (b) Spin Hall signal from electrodes at 2 μm from the sample edges. The current density is 5.7 $\times 10^3$A/cm^2. The solid and open circles are the results for reversed currents in parallel magnetization. The solid line is for anti-parallel magnetization. The inset shows the Hanle effect between the electrodes. The temperature is 30 K.

Hamiltonian" $\mathscr{H}_{\mathrm{B}}(\boldsymbol{k}) = e^{-i\boldsymbol{k}\boldsymbol{r}}\mathscr{H}_0 e^{i\boldsymbol{k}\boldsymbol{r}}$, where $\mathscr{H}_0$ is the crystal Hamiltonian, and the wave vector $\boldsymbol{k}$ is treated as a parameter. And the eigenfunction is a lattice periodic function $u_{n\boldsymbol{k}}(\boldsymbol{r})$.

From the space of $u_{n\boldsymbol{k}}(\boldsymbol{r})$, a variation in $\boldsymbol{k}$ can be viewed as that in the Hamiltonian. We have considered such situation in the introduction of Berry phase, in that the crystal wavenumber $\boldsymbol{k}$ is taken as the set of parameters $\boldsymbol{R}$ for adiabatic transition. Then the Berry connection and the Berry curvature in this case are

$$\boldsymbol{A}_n = i\left\langle u_{n\boldsymbol{k}} \middle| \frac{\partial u_{n\boldsymbol{k}}}{\partial \boldsymbol{k}} \right\rangle, \quad \boldsymbol{B}_n(\boldsymbol{k}) = i\left\langle \frac{\partial u_{n\boldsymbol{k}}}{\partial \boldsymbol{k}} \middle| \times \middle| \frac{\partial u_{n\boldsymbol{k}}}{\partial \boldsymbol{k}} \right\rangle, \tag{10.23}$$

respectively.

Let $|n\boldsymbol{k}\rangle$ be a Bloch function, and we drop the band index n limiting the band to n (single band). We assume the quantities in Eq. (10.23) are not zero, $\boldsymbol{k}$-representation of the coordinate operator $\hat{\boldsymbol{r}}$ is

$$\langle \boldsymbol{k}|\hat{\boldsymbol{r}}|\boldsymbol{k}'\rangle = (i\nabla_{\boldsymbol{k}} + \boldsymbol{A})\,\delta(\boldsymbol{k}-\boldsymbol{k}'),$$

where $\nabla_{\boldsymbol{k}} = \partial/\partial\boldsymbol{k}$. This corresponds to that the dynamic momentum in a magnetic field is written as $-i\hbar\nabla + e\boldsymbol{A}$ in the coordinate representation. Then we obtain

$$\langle \boldsymbol{k}|[\hat{x},\hat{y}]|\boldsymbol{k}'\rangle = (i\nabla_{\boldsymbol{k}} \times \boldsymbol{A})_z\,\delta(\boldsymbol{k}-\boldsymbol{k}') = iB_z\delta(\boldsymbol{k}-\boldsymbol{k}').$$

From the Heisenberg equation $d\hat{q}/dt = [\hat{q}, \mathscr{H}_0 - \boldsymbol{F}\cdot\hat{\boldsymbol{r}}]/i\hbar$, we write the time evolution of operators $\hat{x}$, $\hat{k}_x$ by $\boldsymbol{F}$ as

$$\left\langle \boldsymbol{k}\middle|\frac{d\hat{x}}{dt}\middle|\boldsymbol{k}'\right\rangle = \left[\frac{\partial E}{\partial k_x} - (\boldsymbol{F}\times\boldsymbol{B})_x\right]\frac{\delta(\boldsymbol{k}-\boldsymbol{k}')}{\hbar}, \quad \left\langle \boldsymbol{k}\middle|\frac{d\hat{k}_x}{dt}\middle|\boldsymbol{k}\right\rangle = F_x\frac{\delta(\boldsymbol{k}-\boldsymbol{k}')}{\hbar}.$$

Under the above conditions, a wavepacket f is expanded by Bloch functions as $f = \sum_{\boldsymbol{k}} a_{\boldsymbol{k}}|\boldsymbol{k}\rangle$ ($a_{\boldsymbol{k}} = \langle\boldsymbol{k}|f\rangle$). The time evolution of the averaged values of f in real and wavenumber spaces $\boldsymbol{r}_0$, $\boldsymbol{k}_0$ are

$$\frac{d\boldsymbol{r}_0}{dt} = \boldsymbol{v} = \left\langle f\middle|\frac{d\hat{\boldsymbol{r}}}{dt}\middle|f\right\rangle = \sum_{\boldsymbol{k}}\frac{\langle f|\boldsymbol{k}\rangle}{\hbar}\left(\nabla_{\boldsymbol{k}}E - \boldsymbol{F}\times\boldsymbol{B}\right)\langle\boldsymbol{k}|f\rangle \approx \frac{1}{\hbar}\left(\nabla_{\boldsymbol{k}}E - \boldsymbol{F}\times\boldsymbol{B}\right)|_{\boldsymbol{k}=\boldsymbol{k}_0}, \tag{10.24a}$$

$$d\boldsymbol{k}_0/dt = \boldsymbol{F}/\hbar, \tag{10.24b}$$

respectively. In (10.24a), the average over the wavepacket is replaced with the expectiation value on $\boldsymbol{k}$. The second term in (10.24a) is the difference from the effective mass approximation due to the Berry curvature. This is called **anomalous velocity**.

When the Fermi level is within a band gap (i.e. the system is a band insulator) and the anomalous velocity exists, the Hall conductance is quantized as $\sigma_{xy} = \nu e^2/h$ from the TKNN formula. However in ordinary situation with time-reversal symmetry, the Berry curvature $\boldsymbol{B}$ is zero, no anomalous velocity exists, the Hall conductance disappears. As in the two current model the system is devided into $\uparrow, \downarrow$ and we consider $\sigma_{xy}^{\uparrow\downarrow}$ in each system. From the definition (10.6), $\boldsymbol{j}^s = (\hbar/(-2e))(\sigma_{xy}^{\uparrow} - \sigma_{xy}^{\downarrow})E$. Therefore

$$\sigma_{xy}^s = \frac{\hbar}{-2e}(\sigma_{xy}^{\uparrow} - \sigma_{xy}^{\downarrow}) = \frac{-e}{4\pi}(\nu^{\uparrow} - \nu^{\downarrow}) = \frac{-e}{4\pi}\nu_{\mathrm{s}}. \tag{10.25}$$

Here $\nu^{\uparrow,\downarrow}$ is the Chern number for each spin subband and the difference between the two ν_{s} is called spin Chern number. For the spin Chern number to be finite, there should be an effect that gives the same action as the magnetic field and the action should be reversed with spin reversal because in (10.25) the difference between the spin-subband is taken. Systems with a k-linear SOI just like Rashba model apparently fulfill the condition, which results in the appearance of the spin Hall effect.

10.4.3 Quantum spin Hall effect

The above discussion is for an insulator and we need to be careful about "electric curren of each spin" [10]. Even when the total net current is zero, current in each spin subband may be finite and in such a case, the Hall conductance is quantized by e^2/h in each subband. Let us consider the case the region $y < 0$ in xy-plane is occupied with such a two-dimensional insulator. With Heaviside function $\Theta(x)$, the current components are $j_x^\chi = \Theta(y)\sigma_{xy}^\chi E_y$, $j_y^\chi = -\Theta(y)\sigma_{xy}^\chi E_x$, where χ is $\uparrow$ or $\downarrow$. Then the charge conservation is written as

$$\frac{d\rho^\chi}{dt} + \nabla \cdot \boldsymbol{j}^\chi = \frac{d\rho^\chi}{dt} - \delta(y)\sigma_{xy}\chi E_x = \frac{d\rho^\chi}{dt} - \delta(y)\nu^\chi \frac{e^2}{h} E_x = 0.$$

Taking the difference between the two equations for $\chi = \uparrow, \downarrow$, and integrating on the entire space of the system, we get

$$\frac{dS_z}{dt} = L\frac{-e}{2\pi}\nu_{\rm s} E_x,$$

where S_z is the z-component of total spin of the system, L is the length at the edge $y = 0$. The result tells that we need some anomaly at the edge to conserve S_z.

This leads to the idea of the edge states as in the quantum Hall effect, but in the present case there should be no net electric current. Then we consider two edge states (**helical edge states**) with opposite charge velocity and spin at the bounday. We write their dispersions as $E_k^{\uparrow\downarrow} = \pm v(\delta k_x - eE_x t)$ ($\uparrow$: $+$, $\downarrow$: $-$, $\delta k_x = k_x - k_{\rm F}$). Because the variation in the number of particles in the unit time is $\delta N_{\uparrow\downarrow} = \pm eE_x L/2\pi$

$$\frac{dS_z}{dt} = \frac{1}{2}(\delta N_\uparrow - \delta N_\downarrow) = L\frac{e}{2\pi}E_x.$$

With comparison of the two equations, from the condition $dS_z/dt = 0$ in total, the number of helical edge states should be the spin Chern number.

Such an insulator is called quantum spin Hall insulator or **topological insulator**. Since spin Chern number is an integer, from Eq. (10.25), the spin Hall conductance of a topological insulator is quantized by $e/4\pi$.

10.4.4 Quantum well of a topological insulator

We will conclude this lecture by introducing an experiment that verified a topological insulator with quantum spin Hall effect for the first time. After this experiment, many topological insulators were discovered in a dozen years, and not only topological insulators but also Dilac semimetals and Weyl semimetals were found, and a wide range of topological materials such as magnetic ones were found. The research is now widely going on. In addition, the reason why we named

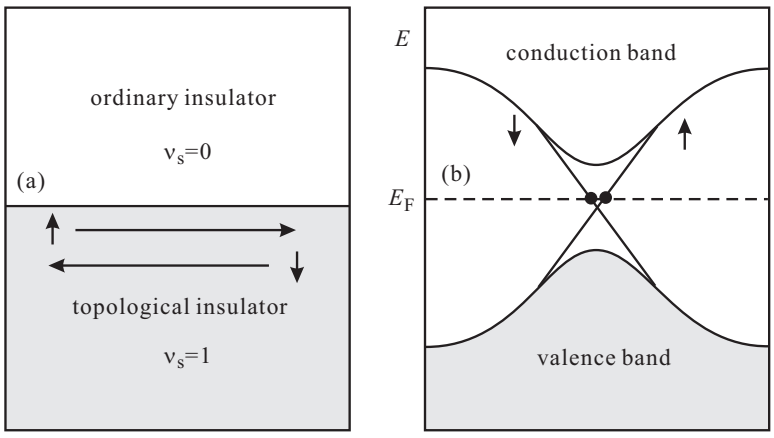


Fig. 10.7 Concept of the topological insulator. (a) The region $y < 0$ is a 2-dimensional topological insulator and the rest is a vacuum, which is an ordinary insulator. A pair of helical edge states exists at the boundary. The spin Chern number in the topological insulator is 1 corresponding to the nubmer of edge states. (b) The energy dispersion diagram. The helical edge states have a linear dispersion relation[10].

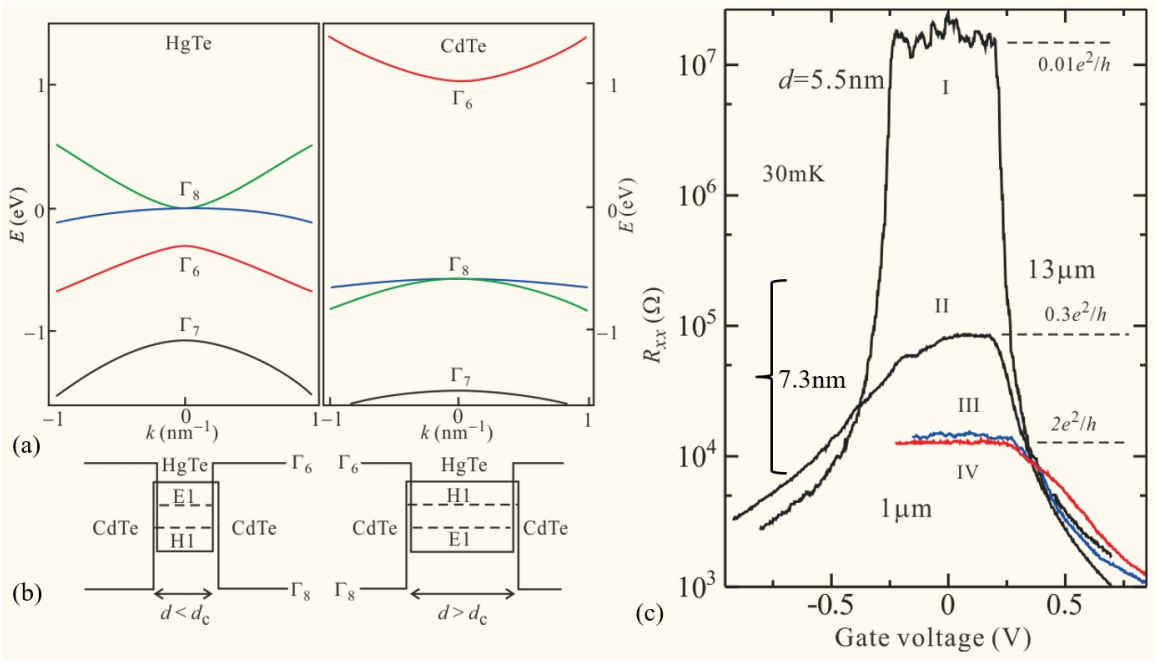


Fig. 10.8 (a) Band structure around Γ-point of HgTe and CdTe. Γ_6, Γ_7, Γ_8 are Koster symbols for space group elements, which indicate the symmetries of the bands. (b) Schematic view of the positions of E1 and H1 subbands. (c) Longitudinal resistances of CdTe/HgTe/CdTe quantum wells. The gate voltage is measured from the position at which the Fermi level comes to the center of the gap. Well widths, I: 5.5 nm; II, III, IV: 7.3 nm. The distance between the electrodes are 13 μm for I and II, 1 μm for III and IV.

the "quantum spin Hall effect" is that quantum hole insulators, which have a long history, are also considered to be a type of topological insulator, so the first topological insulator discovered by humans is called the quantum hall insulator. The reason why I've added "quantum spin Hall effect" is that quantum Hall insulators, which have a long history, are also considered to be a type of topological insulator, so the first topological insulator discovered by humans should be the quantum Hall insulator and I think we need to mention that clearly.

Figure 10.8 shows the setup of the experiment and the results. HgTe is used as the topological insulator. A thin film of HgTe is inserted between two CdTe films, which work as barrier layers and form a quantum well. Figure 10.8(a) shows the band diagrams of HgTe and CdTe zinc blende crystals calculated by 8-band k·p model with SOI. In CdTe, just like GaAs, the Γ_6 ($J = \pm 1/2$) conduction band mainly comes from s-orbital while the Γ_8 ($J = \pm 1/2,\ \pm 3/2$) valence band plus the Γ_7 spin-split-off band come from p-orbitals. On the other hand in HgTe, the strong SOI causes a band inversion, that is the Γ_8 band floats above the Γ_6 band. In a HgTe quantum well, the quantum confinement modifies the band structure. We name quantum-confined level from the electron-like dispersion band Γ_6 as E1, and that from the hole-like dispersion band Γ_8 as H1.

A theoretical model named Bernevig-Hughes-Zhang(BHZ) model was proposed for the HgTe quantum well[11]. According to the model, as long as the order of E1 and H1 bands on the energy axis keeps the inversion ($E_{\mathrm{H1}} > E_{\mathrm{E1}}$) the Chern number ($Z_2$ topological number) is 1 namely the HgTe quantum well is a two-dimensional topological insulator. The quantum well potential is symmetric for the center of well and there is no SIA hence no Rashba SOI. Hence the well structure does not give important change in the SOI. On the other hand, the quantum confinement enhances E1 level and lowers H1 level. With decreasing the well width, then, the order in the level is transformed into the ordinary one at the critical width and the system goes into an ordinary insulator.

In Fig. 10.8(c) the four-terminal resistance of quantum wells of CdTe/HgTe/CdTe, in which the E1-H1 crossing critical width is d_c=6.3 nm. The well widths is 5.5 nm (less than d_c) for I and 7.3 nm for others. When the well width is less

than d_c, the resistance is very high around $V_g = 0$ indicating that the system is an ordinary insulator. On the other hand when the width is wider than d_c and H1 level places above E1, a topological insulator is realized and a helical edge state appears at the sample edge. The electric conductance through the helical edge state should be $2e^2/h$ from the Landauer formula (**??**). Actually in the samples III (sample width 1 μm), IV (width 0.5 μm), the conductances are $2e^2/h$ around the gate voltage for the insulating phase, which fact indicates the realization of the topological insulating (quantum spin Hall) phase.

References

[1] S. Maekawa ed. *Concepts in Spin Electronics*, (Oxford, 2005).

[2] F. Jedema, A. Filip, B. Van Wees, Nature **410**, 345–348(2001).

[3] T. Sasaki *et al*., Applied Physics Express **2**, 053003(2009).

[4] R. Winkler, *Spin-Orbit Coupling Effects in Two-Dimensional Electron and Hole Systems*, (Springer, 2003).

[5] G. Dresselhaus, Phys. Rev. **100**, 580–586(1955).

[6] E. Rashba, Soviet Physics-Solid State **2**, 1109–1122(1960).

[7] Y. A. Bychkov, E. I. Rashba, Journal of physics C: Solid state physics **17**, 6039(1984).

[8] A. Därr, J. Kotthaus, T. Ando, Proc. 13th Int. Conf. Phys. Semidond. p.774 (1976) .

[9] J. Nitta *et al*., Phys. Rev. Lett. **78**, 1335–1338(1997).

[10] M. Z. Hasan, C. L. Kane, Rev. Mod. Phys. **82**, 3045–3067(2010).

[11] A. Bernevig, T. L. Hughes, and S.-C. Zhang, Science **314**, 1757 (2006).

[12] M. König, S. Wiedmann, C. Brüne1, A. Roth, H. Buhmann, L. W. Molenkamp, X.-L. Qi, S.-C. Zhang, Science **318**, 766 (2007).

[13] S.-Q. Shen, "Topological Insulators" (Springer, 2012).

Appendix 10A: Motion of small magnetic moment

10A.1 Electron spin in a magnetic field

Let us consider a single electron spin s in static magnetic field B_0 along z-direction. We consider only the Zeeman energy:

$$\mathscr{H} = (e\hbar/2m_0)gB_0\hat{s}_z = g\mu_{\rm B}B_0\hat{s}_z,$$

where $\mu_{\rm B}$ is the **Bohr magneton**. From the commutatio relatoin of spin operators $[\hat{s}_j, \hat{s}_k] = i\hat{s}_l/2$ ((j,k,l) are cyclic replacement of (x,y,z)),

$$[\mathscr{H}, \hat{s}_x] = ig\mu_{\rm B}B_0\hat{s}_y, \quad [\mathscr{H}, \hat{s}_y] = -ig\mu_{\rm B}B_0\hat{s}_x, \quad [\mathscr{H}, \hat{s}_z] = 0.$$

Hence the Heisenberg equation of motion tells.

$$\frac{\partial\langle s_x\rangle}{\partial t} = -\frac{g\mu_{\rm B}}{\hbar}B_0\langle s_y\rangle, \quad \frac{\partial\langle s_y\rangle}{\partial t} = \frac{g\mu_{\rm B}}{\hbar}B_0\langle s_x\rangle, \quad \frac{\partial\langle s_z\rangle}{\partial t} = 0. \tag{10A.1}$$

$$\therefore\ \langle s_x\rangle = A\cos\omega_0 t, \ \ \langle s_y\rangle = A\sin\omega_0 t, \ \ \langle s_z\rangle = C, \ \ \omega_0 = \frac{eg}{2m_0}B_0, \tag{10A.2}$$

where $A^2 + C^2 = s^2$. Equation (10A.2) is representing precession around z-axis with the **Larmor frequency** ω_0.

Next we add a rotating magnetic field $B_1(\boldsymbol{e}_x\cos\omega t + \boldsymbol{e}_y\sin\omega t)$ in xy-plane. The time-dependent Hamiltonian with that is written as

$$\mathscr{H}(t) = g\mu_{\rm B}(B_1\cos\omega t\hat{s}_x + B_1\sin\omega t\hat{s}_y + B_0\hat{s}_z).$$

The time evolution of spin-wavefunction $\chi(t) = u(t)|\uparrow\rangle + d(t)|\downarrow\rangle$ is represented as

$$i\hbar\frac{\partial}{\partial t}\begin{pmatrix} u \\ d \end{pmatrix} = -g\mu_{\rm B}\begin{pmatrix} B_0 & B_1 e^{-i\omega t} \\ B_1 e^{i\omega t} & -B_0 \end{pmatrix}\begin{pmatrix} u \\ d \end{pmatrix}.$$

The solutions of the above simultaneous differential equations are expressed as

$$u(t) = C(\Omega \mp \omega_0 \pm \omega)e^{i(\pm\Omega-\omega)t/2}, \tag{10A.3a}$$

$$v(t) = \pm C\omega_c e^{i(\pm\Omega+\omega)t/2}, \tag{10A.3b}$$

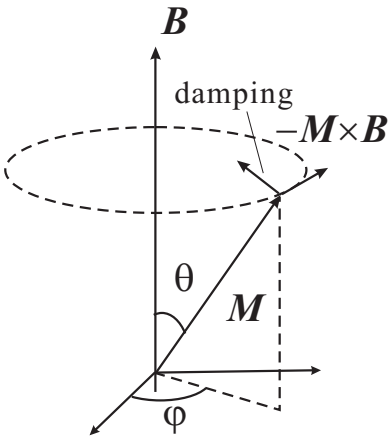
where C is an integration constant, $\omega_c = eB_1/m_0$, and $\Omega = \sqrt{(\omega-\omega_0)^2+\omega_c^2}$. Taking the initial condiction as $u(t) = 1$, $d(t) = 0$, the solutions are

$$u(t) = \sqrt{2 - \frac{\omega_c^2}{\Omega^2}}\sin\left(\frac{\Omega t}{2} + \alpha\right)e^{-i\omega t/2}, \quad v(t) = \frac{\omega_c}{\Omega}\sin\frac{\Omega t}{2}e^{i\omega t/2},$$

where $\alpha = \arctan(\Omega/(\omega-\omega_0))$. Then we obtain

$$|d(t)|^2 = \frac{\omega_c^2}{(\omega-\omega_0)^2+\omega_c^2}\sin^2\frac{\Omega t}{2}, \tag{10A.4}$$

which indicates a Lorentz type resonance and the osicllation with the frequency $\Omega(=\omega_c)$ at $\omega = \omega_0$.



10A.2 LLG equation

Appling the equation of motion (10A.1) to a general macroscopic magnetic moment $\boldsymbol{M}$, we obtain the Landau-Lifshitz equation

$$\frac{\partial \boldsymbol{M}}{\partial t} = -\frac{g\mu_{\rm B}}{\hbar}\boldsymbol{M}\times\boldsymbol{B}. \tag{10A.5}$$

The addition of the relaxation $\boldsymbol{\mathcal{R}}$ of $\boldsymbol{M}$ gives

$$\frac{\partial \boldsymbol{M}}{\partial t} = -\frac{g\mu_{\rm B}}{\hbar}\boldsymbol{M}\times\boldsymbol{B} + \mathcal{R}. \tag{10A.6}$$

What should be the mathematical form of $\boldsymbol{\mathcal{R}}$? Since $\boldsymbol{M}$ has the lowest energy at the direction of $\boldsymbol{B}$, the relaxation should be a force to this direction as shown in the left. The force is perpendicular to $-\boldsymbol{M}\times\boldsymbol{B}$ and $\boldsymbol{M}$. It is natural to infer that

$$\mathcal{R}_{\rm LL} = -\lambda\frac{\boldsymbol{M}}{|\boldsymbol{M}|}\times(\boldsymbol{M}\times\boldsymbol{B}), \tag{10A.7}$$

where λ is a constant. This is called the Landau-Lifshitz damping term.

Another idea is that the relaxation rate should be proportional to time-variation of $\partial\boldsymbol{M}/\partial t$, which has the same direction as $-\boldsymbol{M}\times\boldsymbol{B}$. In this idea the damping term can be written with a constant α as

$$\mathcal{R}_{\rm G} = \alpha\frac{\boldsymbol{M}}{|\boldsymbol{M}|}\times\frac{\partial\boldsymbol{M}}{\partial t}, \tag{10A.8}$$

which is called Gilbert damping term. If we substitute the equation of motion (10A.5) into this $\partial\boldsymbol{M}/\partial t$, the term is the same as the Landau-Lifshitz term. Adopting $\mathcal{R}_{\rm G}$ for the damping, we reach the **Landau-Lifshitz-Gilbert** (**LLG**) equation:

$$\frac{\partial \boldsymbol{M}}{\partial t} = -\frac{g\mu_{\rm B}}{\hbar}\boldsymbol{M}\times\boldsymbol{B} + \alpha\frac{\boldsymbol{M}}{|\boldsymbol{M}|}\times\frac{\partial\boldsymbol{M}}{\partial t}, \tag{10A.9}$$

which is often used to describe motions of magnetization phenomenologically.