1. Introduction

We have seen that when the electrons in two hydrogen atoms interact, their energy levels will split, i.e., the degeneracy associated with two identical isolated systems is removed. In the case of a crystalline solid, there are $\sim 10^{23}$ atoms, each with its contingent of electrons and the electrons of all these atoms will interact to varying degrees, dependent on their separation. Thus, a large degree of degeneracy ($\sim 10^{23}$) will be removed. However, there are so many states that even though the overall splitting of a state may be a few eV wide, the individual split states will be very close together. (There will be $\sim 10^{23}$ different states with an energy range of a few eV.) Thus, these split states form bands, where a band is a set of states that are very close together in energy, which span a certain energy width. The band theory has been extremely successful in explaining the properties of solids.

2. The Kronig-Penney Model

Crystal lattices are periodic and so the potential experienced by an electron will be periodic. In the Kronig-Penney (KP) model, positive ions are placed at the lattice positions in a one-dimensional crystal. The potential energy of an electron is shown in part (a) of the figure below. (The potential energy approaches minus infinity as the distance between the electron and an ion approaches zero.) In the KP model, the potential is treated as a periodic array of square wells. The depth of a well is $V_0$ and the width of the barrier between wells is $b$. (See part (b) of the figure below.)

To solve the TISE for this model, we first take note of Bloch’s theorem. Bloch’s theorem states that if the lattice (and hence the potential) has period $a$, then the eigenfunctions that are solutions to the TISE must take the form

$$\psi(x) = u_k(x)e^{ikx},$$  \hspace{1cm} (5.1)
where \( k \) is the wave vector in the periodic potential and the function \( u_0(x) \) has the periodicity of the lattice, i.e., \( u_0(x) = u_0(x+a) \). Bloch’s theorem may be written in another convenient form by noting that 

\[
\psi(x + a) = u_k(x + a)e^{ik(x+a)} = e^{ika}u_k(x)e^{ikx} = e^{ika}\psi(x).
\]

Thus, the bands become narrower. In the limit \( E > V_0 \) or \( E < V_0 \). In the barrier region, the solutions are traveling waves or exponentially decaying depending on whether \( E > V_0 \) or \( E < V_0 \). The continuity boundary conditions have to be satisfied as well as periodicity (for both \( \psi(x) \) and its first derivative). Assuming \( E < V_0 \) and slogging though the algebra gives

\[
\left[ \frac{\alpha^2 - \beta^2}{2\alpha\beta} \right] \sin\alpha\beta a + \cos\alpha\beta \cos \beta a = \cos k(a + b),
\]

where \( \alpha = \sqrt{2m(V_0 - E)}/\hbar \) and \( \beta = \sqrt{2mE}/\hbar \). In Eq. (5.3), the LHS depends only on the energy \( E \) and the RHS depends only on the wave vector \( k \). Thus, this equation gives us the allowed values of energy corresponding to a given value of \( k \). Note that \( k \) is not the usual quantity proportional to the electron's momentum. (In this case, that "usual quantity" is \( \beta \).) Rather, \( k \) describes the state of the crystal lattice as a whole (electron and ions). It is very much like a quantum number, though its value is not discrete. \( \hbar k \) is called the crystal momentum.

To simplify the interpretation of Eq. (5.3), we take the limits \( V_0 \to \infty \) and \( b \to 0 \) such that the quantity \( P = \alpha^2 ba/2 \) remains finite. Then Eq. (5.3) becomes

\[
(P/\beta a)\sin \beta a + \cos \beta a = \cos ka.
\]

The RHS of Eq. (5.4) must lie between \(-1\) and \(+1\) and therefore, so must the LHS. The plot shows that Eq. (5.4) is only satisfied for only certain ranges of values of \( \beta a \) and hence only certain ranges of values of the energy. The regions of allowed energy, called bands, are separated by regions where there no allowed energies (forbidden regions) called band gaps.

As \( P \) gets larger, the height of the curve becomes very large and so the slope becomes steeper. Thus, the bands become narrower. In the limit \( P \to \infty \), each band shrinks to a single quantized value of energy given by \( \sin \beta a = 0 \), i.e., \( E = n^2\hbar^2/2ma^2 \), which are the energies for an infinite square well. (See ppt images.) In the limit \( P \to 0 \), one obtains \( \cos \beta a = \cos ka \), i.e., \( k = \beta \). Thus,
\( E = \frac{\hbar^2 k^2}{2m} \), which is the relation for a free electron for which all energies are allowed. At the edge of allowed band, \( \cos ka = \pm 1 \), i.e.,

\[
    k = n\frac{\pi}{a}, \quad n = \pm 1, \pm 2, ..
\]

The energy is discontinuous at those values of \( k \), i.e., there is an *energy gap*.

2. Origin of the Energy Gap

Consider an electron wave in a periodic potential (Bloch wave) moving in the positive \( x \) direction. The propagating part of the wave function is \( \psi_{pp,+} = e^{ikx} \). As the electron propagates, it will undergo partial reflection due to the periodic array of atoms, producing a wave moving in the opposite direction: \( \psi_{pp,-} = e^{-ikx} \). In x-ray diffraction, Bragg reflection occurs when there is constructive interference between the incident and reflected waves. The Bragg condition is \( 2a \sin \theta = n\lambda \) or \( ka \sin \theta = n\pi \), where \( k = 2\pi / \lambda \) and \( n \) is an integer. Here, \( a \) is the separation of the atomic planes. If we take \( \theta = 90^\circ \) (one dimension or normal incidence), we see that \( k = n\pi / a \). The Bragg condition occurs when there is complete constructive interference between incident and reflected waves, i.e., there is no "transmitted" wave; only incident and reflected. The electron wave undergoes precisely this process when \( k = n\pi / a \), i.e., at \( k \) values corresponding to the edge of the allowed bands. This reflection produces standing waves \( e^{ikx} + e^{-ikx} \) or \( e^{ikx} - e^{-ikx} \), i.e., \( \cos kx \) or \( \sin kx \) (omitting a factor of 2). Now these standing waves have their antinodes (and nodes) at different points and so they pile up charge at different points. One will pile up charge at the location of the positive ion cores, which will lower the potential energy (negative and positive right on top of each other) and the other will pile up charge between the ions thereby raising the potential energy. [Insert picture from Kittel.] These two distinct energies for the same \( k \) value give rise to the energy gap. [See Phet “Band Structure” simulation.]

3. Coupled Modes

Consider a one-dimensional chain of \( N \) atoms. As argued before, the \( N \)-fold degeneracy that exists in the absence of interactions will be split when the atoms are close enough to interact.
This splitting causes the allowed and forbidden bands. In our discussion, we will assume that only nearest-neighbor atoms interact (a common starting-point assumption in condensed-matter and materials physics).

Consider the \(j\)th atom in the chain. It will interact with atoms \(j-1\) and \(j+1\). Assuming \(E_1\) is the atomic energy level in the absence of interaction and the coupling constants are \(-A\) [e.g., \(H_{j,j+1} = -AT_{j+1}\)]. Proceeding according to the Feynman model, one obtains for atom \(j\)

\[
i \hbar \frac{dT_j}{dt} = E_1 T_j - AT_{j-1} - AT_{j+1}. \tag{5.6}
\]

Let the position of the \(j\)th atom be \(x_j\). Then the position of atom \(j-1\) is \(x_j-a\) and that for atom \(j+1\) is \(x_j+a\). We substitute the trial wave function \(T_j = e^{ikx_j - \frac{i\pi}{a}}\). This yields

\[
E = E_1 - 2A \cos ka. \tag{5.7}
\]

For \(-\pi/a \leq k \leq \pi/a\), we obtain the range of energy values \(E_1 - 2A \leq E \leq E_1 + 2A\) within the allowed band. In the Feynman model, we see that all necessary \(k\) values fall within the region \(-\pi/a \leq k \leq \pi/a\). Values of \(k\) outside of this range simply duplicate other solutions. [However, in other models such as KP, using values of \(k\) outside this range can be useful as seen before (see \(E\) vs. \(k\) plot above.)] These \(k\) ranges associated with the allowed bands are called Brillouin zones.

In the coupled-mode model, other energy bands can be obtained by going to other uncoupled energy levels, say, \(E_2\). In two or three dimensions, the conclusions are much the same as those we arrived at with simpler one-dimensional models.

4. Effective mass

In a periodic potential, how does an electron respond to a force, e.g., that exerted by an electric field? To answer this question, we go back to the concept of group velocity, which represents the velocity of a wave packet. It is the velocity with which the envelope encompassing the packet moves. For a wave packet with center frequency \(\omega\) and center wave number \(k\), the group velocity is

\[
v_g = \frac{d\omega}{dk}. \tag{5.8}
\]

For a quantum particle, \(E = \hbar \omega\), so

\[
v_g = \frac{1}{\hbar} \frac{dE}{dk}. \tag{5.9}
\]
If a force $F$ is exerted on the electron, energy is transferred to the system in the amount of the work done:

$$dE = Fdx = F\nu_g dt.$$  

Also, $dE = \hbar d\omega = \hbar \frac{d\omega}{dk}dk = \hbar \nu_g dk = \hbar \nu_g \frac{dk}{dt}dt$. Comparing the two expressions for $dE$ gives

$$F = \hbar \frac{dk}{dt}. \quad (5.10)$$

Taking the time derivative of Eq. (5.9) gives

$$\frac{dv_g}{dt} = \frac{d}{dk} \left( \frac{1}{\hbar} \frac{dE}{dk} \right) \frac{dk}{dt} = \frac{1}{\hbar} \frac{d^2E}{dk^2} dt.$$

Using Eq. (5.10), this gives

$$\frac{dv_g}{dt} = \frac{1}{\hbar^2} \frac{d^2E}{dk^2} F.$$

This looks like Newton’s second law $a = F/m$. Thus, we define the effective mass as follows:

$$m^* = h^2 \left( \frac{d^2E}{dk^2} \right)^{-1}. \quad (5.11)$$

Thus, the effective mass of an electron is inversely proportional to the curvature of the $E(k)$ function. For a free electron, $E = \hbar^2 k^2 / 2m$. We quickly see that in this case, $m^* = m$. Using Eq. (5.7) for a one-dimensional lattice, we find that

$$m^* = \frac{h^2}{2Aa^2} \sec ka. \quad (5.12)$$

Close to the bottom of a band, e.g., near $k = 0$ in the first Brillouin zone, $E(k)$ is nearly parabolic and the effective mass is free-electron-like. Away from $k = 0$ within the range $-\pi / a \leq k \leq \pi / a$, $m^*$ goes to infinity and becomes negative. [Show plots in Solymar.] This behavior can be understood on the basis of the increased reflection of the electron Bloch wave as $|k| \rightarrow \pi / a$. For example, when $m^*$ goes to infinity, the forward momentum change due to the applied force is balanced by the backward momentum change due to reflection. Hence, there is no net change in electron momentum despite the application of a force. As $|k|$ increases toward the top of the band, the greater amount of reflection actually causes a net change in momentum in the backward direction, giving a negative effective mass.

The propagation of an electron in the backward direction is equivalent to the propagation of a positive particle in the forward direction (i.e., in the direction of the applied force). Thus, at the top of a band where electron effective masses tend to be negative, the concept of a hole, i.e., a positive particle propagating in the direction of the applied force, is useful.
5. Effective Number of Free Electrons

When an electric field is applied to a material, for the electrons to gain energy and produce a current, there must be nearby higher-energy empty states that the electrons can occupy. If a band is full, then there are no empty states and so there will be no net electron motion, i.e., there will be no current when an electric field is applied. In this case, the effective number of free electrons is zero. [Note that in a periodic potential, the electrons are never free. We are simply making an analogy with a free-electron system.] The effective number of free electrons for a one-dimensional band filled up to $k = k_a$ at $T = 0$ K is given by

$$N_{\text{eff}} = \frac{2 m}{\pi \hbar^2} \frac{dE}{dk} \bigg|_{k=k_a} .$$

(5.13)

Thus, the effective number of free electrons that can contribute to electric conduction depends on the slope of the $E(k)$ function (proportional to $v_g$) at the highest occupied energy level. Note that at the top of a band, the slope is zero, so when the band is full, $N_{\text{eff}} = 0$, as noted above.

6. More About Holes

In a full band, the sum of all the $k$ values is equal to zero because for every $k$, there is a $-k$. [Bold type denotes a vector, i.e., the space is 2D or 3D.] Thus,

$$\sum_j k_j = 0 .$$

(5.14)

If an electron an electron is removed, e.g., one near the top of the band, then an excitation is created since energy is needed to remove the electron. The resulting unoccupied state is a hole. If the $l^{th}$ electron is removed, then we have

$$\sum_{j \neq l} k_j = \sum_j k_j - k_l = -k_l .$$

(5.15)

This is the $k$ value of the hole, i.e.,

$$k_h = -k_l .$$

(5.16)

Let the energy at the top of the band be zero. When a hole is created, the energy of this state is the negative of the energy of the state that was previously occupied by the electron. Thus,

$$E_h = -E(k_l) .$$

(5.17)

Using the definition of effective mass [Eq. (5.11)], we find that

$$\frac{1}{h^2} \left( \frac{d}{dk_h} \left( \frac{dE_h}{dk_h} \right) \right) = \frac{1}{h^2} \left( \frac{d}{d(-k_l)} \left( \frac{d(-E_l)}{d(-k_l)} \right) \right) = \frac{1}{h^2} \left( \frac{d}{dk_l} \left( \frac{d(-E_l)}{dk_l} \right) \right) = - \frac{1}{h^2} \left( \frac{d}{dk_l} \right) \left( \frac{dE_l}{dk_l} \right) ,$$

i.e.,

$$m_h^* = -m_l^* .$$

(5.18)

(We have used 1D quantities for simplicity.) Thus, the effective mass of the hole is the negative of the effective mass of the absent electron as we saw before. Finally, we note that a full band carries no current:

$$\sum_j (-e)v_j = 0 .$$

(5.19)

Let us examine the group velocity. Using 1D quantities again, we find that

$$v_l = \frac{1}{h} \frac{dE_l}{dk_l} = \frac{1}{h} \frac{d(-E_h)}{d(-k_l)} = \frac{1}{h} \frac{dE_h}{dk_h} = v_h .$$

(5.20)
If the $j^{th}$ electron is removed, then there will be an associated current:

$$\sum_{j \neq l} (-e)v_j = \sum_{j} (-e)v_j - (-e)v_l = (+e)v_h,$$

where we have used the result $v_h = v_i$. We see that the hole acts as a positively charged particle.

7. Metals and Insulators

A single atomic level for a crystal of $N$ atoms broadens into a band of $2N$ states because of interaction (or tunneling of electrons between atoms). The factor of two arises because of spin. If the basis atom has one valence electron, this is just enough to half-fill the uppermost occupied band. Since there are plenty of empty states available, the material will have high conductivity and will be a metal. If the atom has an even number of valence electrons, then the occupied bands will be full and the crystal should be an insulator. While this is true for many solids, there are also metals whose atoms have two or four valence electrons. How is this possible?

8. Divalent Metals

The existence of divalent metals cannot be understood within a one-dimensional model. So we do the next best thing and go to two dimensions. Within the coupled-mode approach, one obtains

$$E = E_1 - 2A_x \cos k_x a - 2A_y \cos k_y b,$$

where $a$ and $b$ are the lattice spacings in the $x$ and $y$ directions, respectively, and $A_x$ and $A_y$ are the corresponding coupling constants. Let us take $a = b$, $A_x = A_y = \frac{1}{4}$, and $E_1 = 1$. The constant-energy contours in the first Brillouin zone for this 2D crystal are shown in the figure to the right. At the zone boundaries $(k_x, k_y = \pm \frac{\pi}{a})$, there will be an energy gap, with a discontinuous jump in energy. Thus at $B$, the energy is 1 unit whereas at $C$, just outside the zone boundary, the energy is higher, say, 1.2 units. Note that at the corners of the zone, the energies are significantly higher than in the central region because both $k_x a$ and $k_y a$ are near $\pm \pi$ so the cosines are both near -1. Thus, rather than filling those states, electrons will go to the next higher band (next zone) because the energies at the bottom of the next band will be lower than
those close to the zone corners. Thus, if the energy gap is relatively small, the upper and lower bands will not be full and the crystal will be a conductor. Note that the lower band will have holes as the carriers (band almost full) and the upper band will have electrons as the carriers. This is precisely what experiments show for some divalent metals: The current carriers have both positive and negative sign. Clearly, band theory is crucial in explaining this phenomenon. [Show pictures of band filling from Solymar.] Since the basis atoms are divalent (or 4 or 6 valence electrons), then there will be as many electrons as there are states in the band. If the energy gap is large, the band will be full and the upper band will be empty. We will have an insulator.

9. Non-zero Temperatures

Our discussion thus far has been conducted at $T = 0$ K since we have not accounted for the effects of temperature. What if $T > 0$ K? For metals in which the uppermost band is not full, electrons would be able to occupy nearby higher energy levels but the band would still not be full. Hence, a metal will remain a metal. For an insulator with a full band at $T = 0$ K, the Fermi level must lie somewhere in the gap between the uppermost two bands (since the upper band is empty). The lower of these bands is called the valence band and the upper one is the conduction band. For temperatures above zero, the FD distribution is no longer equal to one for all states below $E_F$ and no longer zero for all states above $E_F$. Thus, some electrons will occupy the conduction band. If the temperature is high enough, then a significant number of electrons will occupy the conduction band, especially when the thermal energy $k_B T$ becomes comparable to the energy gap. The material therefore becomes a conductor. The conductivity depends sensitively on the energy gap because for $E >> E_F$, the FD function varies exponentially.