UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE FÍSICA
CAIXA POSTAL 20516
01498 - SÃO PAULO - SP
BRASIL

PUBLICAÇÕES

INTRODUCTION TO THE THEORY OF ELECTRON STATES IN SEMICONDUCTORS

Marília Junqueira Caldas
Instituto de Física, Universidade de São Paulo

Julho/1990

INTRODUCTION TO THE THEORY OF ELECTRON STATES IN SEMICONDUCTORS

Marília Junqueira Caldas

Instituto de Física, Universidade de São Paulo

C.P. 20516, 01498 São Paulo, S.P., Brazil

These lectures were given at the 4th Brazilian School of Semiconductor Physics, and were directed at an audience of graduate students, some just starting research work, others almost at the end of doctorate programs, most of them involved with experimental research. I planned to present the theory side of research in semiconductors at an introductory level, preparing for the specialized lectures they would hear during the following week of the School. As such, I started from the Hartree–Fock and Local–Density approximations and worked through some of the basic descriptions of bulk crystalline semiconductors; I then gave a very brief view of the theory of defects in semiconductors, with emphasis on the interpretation of theoretical results, and on the comparability with experiment.

I. INTRODUCTION

We are interested in describing the electron states in a semiconductor, which amounts to describing the behavior of $\sim 10^{23}$ electrons each in the field of the other plus all the atomic nuclei, if no other external field is included. As usual, the stationary wave function $\Psi(\{\vec{x}_i\})$, where $\vec{x} = (\vec{r}, \vec{\xi})$ accounts for space and spin coordinates, is to be obtained from a hamiltonian in $[\mathcal{H} - E]\, \Psi(\{\vec{x}_i\}) = 0$ at temperature $T = 0$. For such a huge number of particles no useful information could be gained from writing Ψ out explicity, and it is in fact impossible to do that. Our task of describing the electron states is, however, feasible if (i) we adopt the independent particle or mean–field approximation, in which one electron moves in the average effective potential of all the others (departures from mean–field considered for special cases only); (ii) we assume that the lattice ions are rigidly fixed at equilibrium positions in which (iii) the external potential on the electrons due to the nuclei is periodic.

We write the periodicity conditions (ii) and (iii) in terms of primitive lattice vectors $\vec{t}_\nu$ ($\nu = 1,2,3$) as

$$V^{ext}(\vec{r} + \vec{t}_\nu) = V^{ext}(\vec{r}) \tag{I.1}$$

Adopting the independent particle model (i) each particle is described by a single–particle state $u_i(\vec{x})$ associated with a single–particle energy ε_i. These energies will be grouped around certain values according to the character of $u_i(\vec{x})$: there will be deeply bound core states highly localized around the nuclei, and the valence electrons with ionization energies closer to zero, say centered around 5 to 10 eV. We will focus on these valence electrons: as they interact throughout the solid, the degeneracy of the single particle energies in the free atoms will be lifted and (due to the enormous number of interacting particles) the energies will be distributed in a quasi–continuous fashion in some energy range (fig. 1). The electrons will "occupy" these states with a lower–to–higher ordering. This is a reasonable

picture of any arrangement of a great number of atoms: the special characteristic (I–1) of
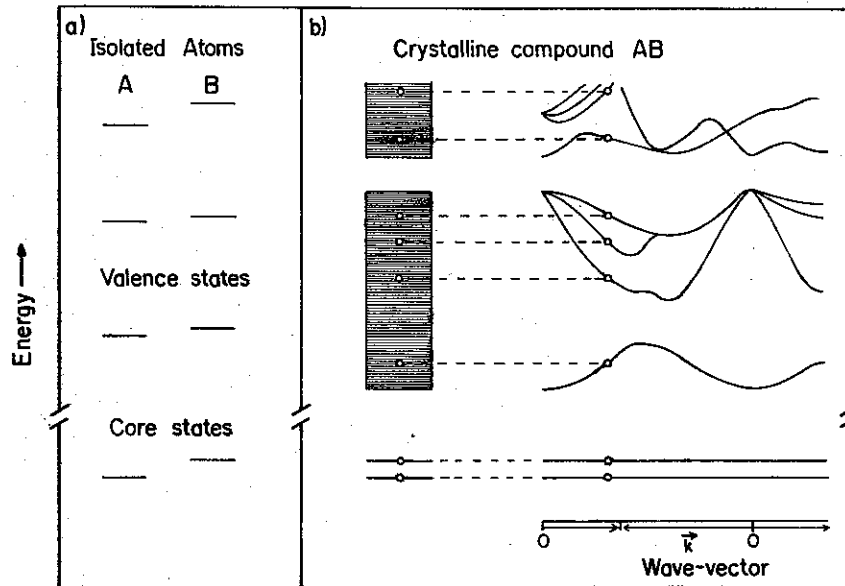

Figure 1— Schematic representation of the energy distribution of single–particle states in a semiconductor a) single–particle energies in the isolated atoms A and B b) compound AB, showing the formation of regions quasi–continuously filled accessible states and forbidden regions–gaps; also usual band picture of compound AB, showing the distribution of accessible states for each $\vec{k}$–vector in the Brillouin Zone along some chosen directions. Note that the extremely localized "core states" do not show energy–dependence with $\vec{k}$.

crystalline solids is reflected on the single–particle states $u_i(\vec{x})$ in the form of Bloch's theorem

$$T_\nu u\,(\vec{r},\vec{\xi}) = u\,(\vec{r} + \vec{t}_\nu,\vec{\xi}) = \exp\{i\vec{k}\cdot\vec{t}_\nu\}\,u\,(\vec{r},\vec{\xi}) \qquad (I-2)$$

or, for every state that satisfies the equation of motion there is a vector $\vec{k}$ such that translation by $\vec{t}_\nu$ is equivalent to straightforward multiplication by $\exp\{-i\vec{k}\cdot\vec{t}_\nu\}$.

In writing (I–1) and (I–2) we are assuming either an infinite solid or, as we do here, periodic boundary conditions imposed on a finite crystal with a very large number $N_c = N_1 N_2 N_3$ of unit cells of volume $\Omega_c = \vec{t}_1\cdot(\vec{t}_2\times\vec{t}_3)$, each containing N_A atoms and N_v valence electrons. Thus, the number of valence electrons in the crystal is large but finite $N = N_v N_c$. Each lattice ion can be reached from some fixed origin in the central unit cell by translation through a vector $\vec{R} = \vec{t} + \vec{\delta}$,

$$\vec{t} = n_1\vec{t}_1 + n_2\vec{t}_2 + n_3\vec{t}_3 \qquad (I-3)$$

where the n_ν are integers $n_\nu = 1,2,...N_\nu$, and $\vec{\delta}$ is a vector locating the ion from an (equivalent) origin in the unit cell. The reciprocal lattice is then defined in terms of primitive vectors

$$\vec{g}_1 = \Omega_c^{-1}\,2\pi\,(\vec{t}_2\times\vec{t}_3)\ ;\ \vec{g}_2 = \Omega_c^{-1}\,2\pi\,(\vec{t}_3\times\vec{t}_1)\ ;\ \vec{g}_3 = \Omega_c^{-1}\,2\pi\,(\vec{t}_1\times\vec{t}_2)\ . \qquad (I-4)$$

Equation (I–2) is a statement about the symmetry of the system, and of state u_i in particular. According to these translation symmetry requirements the valence states $u_i(\vec{x})$ only interact within restricted groups, which may be classified in terms of the $\vec{k}$–vectors: only interact states i, i' for which $\vec{k}' = \vec{k} + \vec{g}$ ($\vec{g} = \sum_\nu n_\nu \vec{g}_\nu$). For each of these sets we choose a representative vector with the smallest possible norm $|\vec{k}|^2$, and thus define a Wigner–Seitz cell in reciprocal space – the first Brillouin Zone (BZ) – containing exactly N_c vectors. It is then useful to present the single–particle energy spectrum in $\vec{k}$–space so

that we obtain an "exploded view" as shown schematically in fig. 1b along certain directions in the BZ: for each **reduced** $\vec{k}$ we depict the energies ε_i for the interacting set of u_i's. The electrons will occupy the lower N states, so that for each $\vec{k}$ the lower N_ν states will occupied. The energy distribution turns out to form quasi–continuous functions of $\vec{k}$, termed energy bands, and the label i may be properly exchanged by $(n,\vec{k})$.

Again, these are properties of any crystalline solid. The characteristic feature of semiconductors is that the number of electrons per cell is even $N_v = 2N_b$, and the band structure results in such a way that N_b bands of each spin are completely occupied or filled, and all higher–lying states are empty at $T = 0$ in the ground state. Furthermore, there is an energy gap between occupied (valence) and empty states which is however not too large, so that interaction with, e.g., a photon in the visible range may excite an electron from the highest valence band to an empty state (where it may respond to an external applied electric field).

These properties, or more properly the properties of impurities in such materials, changed human daily life in the last decades. Semiconductors were and are extensively studied, both from the experimental and theoretical points of view, and to present an unbiased survey of the theory of electron states in semiconductors is today a formidable task. What I will present is a brief and biased survey of the theory: I will discuss some chosen band–structure calculation methods, based on the Hartree–Fock and Density Functional formalisms; and some electronic structure techniques used in the study of defect states. Relativistic and spin–orbit effects will not be treated. Some problems of current interest will be mentioned, but the discussions will be restricted to phenomena in bulk crystalline materials: I will leave out entirely amorphous and alloy semiconductors, surface phenomena, and the new superlattice materials (which are discussed already in the other lectures in this school).

I will further restrict the discussion to the tetrahedral crystals Si and GaAs, the two most important semiconducting materials today. These two materials crystallize in the

diamond and zinc–blende structures, respectively. There are two atoms and $2N_b = 8$ electrons per unit cell, coming from the s and p valence orbitals of the isolated atoms. It is an open structure, shown schematically in Fig. 2, each atom surrounded tetrahedrally by four neighbors. The lattice is conveniently described in terms of two intertwined fcc sublattices, one displaced from the other by $\frac{1}{4}$ along the cube diagonal. The unit cell contains two atoms, one at $\vec{\delta} = 0$ and the other at $\vec{\delta} = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right]a$, if a is the lattice constant of one fcc sublattice.
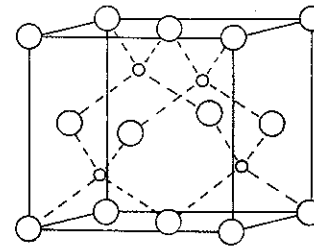

Figure 2– Schematic representation of the structure of tetrahedral semiconductors. For heteropolar compounds such as GaAs, there is a cationic and an anionic sublattice. For homopolar compounds such as Si, both lattices are equivalent.

The space group of the diamond structure is O_h^7, with the 24 operations of the point group T_d (for rotations around axes through $\vec{\delta} = 0$) plus 24 more operations associated to the non–primitive translations $\vec{\delta}$ to the second atom in the cell, reflecting the inversion symmetry about the midpoint between an atom and any of its four neighbors. In the zinc–blende lattice the two atoms in the cell are not equivalent, and the space group is T_d^2 with just the 24 operations of the T_d group.

The special high–symmetry $\vec{k}$–points in the BZ are indicated in Fig. 3. Starting with s and p–type orbitals, wave–functions transforming according to $\vec{k} = \vec{0}$ may be s–type singly degenerate with bonding or antibonding character (Γ_1^i or Γ_2^i, respectively, in O_h^7; Γ_1 in T_d^2); or p–type triply degenerate also with bonding and antibonding character ·

(Γ'_{25} and Γ_{15} in O_h^7 ; Γ_{15} in T_d^2). For $\vec{k}$ along (111) to the L point the degeneracies are similar for diamond and zinc–blende structures: single (L_1 and L'_2 ; L_1) or double (L'_3 ; L_3). For $\vec{k}$ along (100) at the zone edge, X–point, the mentioned inversion symmetry in O_h^7 leads to double degeneracy of all energy levels (X_1 , sp–type, and X_4 , p–type); while for T_d^2 we have both single (X_1,X_3 , sp–type) and double degeneracy (X_5 , p–type).

The zinc–blende structure is characteristic also of the other III–V compounds, AlAs and GaP for instance, and of II–VI semiconductors such as CdTe and ZnSe. Interesting properties are obtained by alloying within one or both sublattices, $Ga_{1-x}Al_xAs$ or $In_{1-x}Ga_xAs_{1-y}P_y$. Very useful heterostructures are tailored from $GaAs/Ga_{1-x}Al_xAs$. Semimagnetic semiconductors $Cd_{1-x}Mn_xTe$, $Zn_{1-x}Mn_xSe$ are other examples of tetrahedral compounds. And also in amorphous semiconductors the shortrange tetrahedral ordering is maintained to a high degree. As such, the discussions included here, even if restricted to Si and GaAs, are relevant to most interesting problems in semiconductor physics.
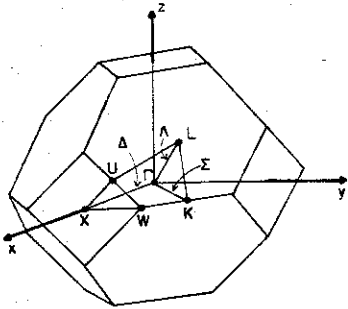

Figure 3– Representation of the Brillouin Zone of tetrahedral semiconductors, showing the usually adopted notations for some high–symmetry points and directions.

II. CRYSTALLINE SEMICONDUCTORS

II.1. The Hartree–Fock Approximation

The independent particle model can best be represented by the Hartree–Fock theory. We write the general hamiltonian for a N–electron system as[1]

$$\mathcal{H} = \sum_i^N h(\vec{r}_i) + \sum_{i \neq j}^N v(r_i,r_j) \qquad (II–1)$$

for a given fixed configuration of nuclei. The one–electron term is

$$h(\vec{r}) = -\frac{\hbar^2}{2m}\nabla^2 - \sum_A Z_A v(\vec{r}_i,\vec{r}_A) \qquad (II–2)$$

with A running through the nuclei (or ion cores); and $v(\vec{r}_i,\vec{r})$ is the Coulomb interaction

$$v(\vec{r}_i,\vec{r}) = \frac{e^2}{|\vec{r}_i - \vec{r}|} . \qquad (II–3)$$

We then incorporate symmetry requirements adopting a product wave function in the form of a Slater determinant

$$\Psi(\{\vec{x}_i\}) = \Psi(\vec{x}_1,\vec{x}_2, ... \vec{x}_N) = (N!)^{-1/2} \det\{u_i(\vec{x}_j)\} , \qquad (II–4)$$

and minimize the total energy with respect to variations in the u_i's, subject to orthogonality constraints. We thus obtain the Hartree–Fock single–particle equations:

$$f u_i(\vec{x}_1) = (h + V_{HF}) u_i(\vec{x}_1) = \varepsilon_i u_i(\vec{x}_1) . \qquad (II–5)$$

The potential is given by a sum of Hartree (V_H) and exchange (V_X) terms $V_{HF} = V_H + V_X$:

$$V_H(\vec{x}_1) = \int d\vec{x}_2 \, v(\vec{r}_1,\vec{r}_2) \, \rho(\vec{x}_2)$$

$$\text{(II–6)}$$

$$V_X(\vec{x}_1,\vec{x}_2) = -\int d\vec{x}_2 \, v(\vec{r}_1,\vec{r}_2) \, \rho(\vec{x}_1,\vec{x}_2) P_{12}$$

where P_{12} is the exchange operator with respect to $\vec{x}_1, \vec{x}_2$ and

$$\rho(\vec{x}_1,\vec{x}_2) = \sum_i^N n_i \, u_i(\vec{x}_1) \, u_i^*(\vec{x}_2)$$

$$\text{(II–7)}$$

where the n_i are either 1 or zero depending whether u_i is occupied or not (occupation numbers).

We note the nonlocality in the exchange term; and the need for selfconsistency in solving (II–5) since the u_i's are involved in the definition of the effective hamiltonian or Fock operator f.

The sums run over all particles $N = 2N_b N_c$ where N_c is the number of cells in the crystal. We now bring in the fact that translation of all electronic coordinates by an allowed vector leaves the system invariant in the crystalline case: if we write the N–electron operator

$$\tau(\vec{t}) = \tau_1^{a_1} \tau_2^{a_2} \tau_3^{a_3}$$

$$\text{(II–8)}$$

with

$$\tau_\nu = \pi_{i=1}^N T_\nu(\vec{x}_i)$$

$$\text{(II–9)}$$

$$\vec{t} = a_1\vec{t}_1 + a_2\vec{t}_2 + a_3\vec{t}_3$$

we have $[\tau(\vec{t}),\mathcal{H}] = 0$. For the Hartree–Fock wave function this results in

$$\tau(\vec{t}) \, \Psi_{\vec{k}}(\vec{x}_1,\vec{x}_2, \dots \vec{x}_N) = \exp\{i\vec{k}\cdot\vec{t}\} \, \Psi_{\vec{k}}(\vec{x}_1,\vec{x}_2, \dots \vec{x}_N)$$

$$\text{(II–10)}$$

where $\vec{k} = (\vec{k}_1 + \vec{k}_2 + \dots + \vec{k}_N)_{BZ}$ is the sum of all individual labels $\vec{k}_i$ reduced to the first Brillouin Zone[3]. Since from (I–2) and (II–7)

$$\rho(\vec{x}_1 + \vec{t}, \vec{x}_2 + \vec{t}) = \rho(\vec{x}_1,\vec{x}_2)$$

$$\text{(II–11)}$$

we obtain that $[\tau(\vec{t}),f] = 0$ also. Periodicity will allow us to solve a similar set of single–particle equations (II–5) for each $\vec{k}$–point in the BZ (as we will emphasize the next section), and for a semiconductor exactly $2N_b$ states per $\vec{k}$–point will be occupied at $T = 0$.

It is usual at this point to substitute the label i for a single–particle state by $(n,\vec{k})$, where n is the "band number", that is, n orders the states at $\vec{k}$ according to energy and symmetry. However, the procedure described in this section is very general and as it will be referred to several times in these lectures, we keep the single label.

We may interpret the eigenvalue ε_i as the negative of the "ionization energy" for an electron in the state u_i, as may be seen if we combine the expressions for the total energies of two systems, with N and $N-1$ electrons. For the total energy we have

$$E(N) = \sum_i n_i \left[\int d\vec{x}_1 \, u_i^*(\vec{x}_1) \left[-\frac{\hbar^2}{2m}\nabla^2 - \sum_A Z_A \, v(\vec{r}_1,\vec{r}_A) \right] u_i(\vec{x}_1) \right]$$

$$+ \frac{1}{2}\sum_{i,j} n_i n_j \left[\int d\vec{x}_1 \, d\vec{x}_2 \, u_i^*(\vec{x}_1) \, u_j^*(\vec{x}_2) \, v(\vec{r}_1,\vec{r}_2) \, u_i(\vec{x}_1) \, u_j(\vec{x}_2) \right]$$

$$-\int d\vec{x}_1 \, d\vec{x}_2 \, u_i^*(\vec{x}_1) \, u_j^*(\vec{x}_2) \, v(\vec{r}_1,\vec{r}_2) \, u_j(\vec{x}_1) \, u_i(\vec{x}_2) \Bigg]$$

$$+\frac{1}{2}\sum_{A\neq B} Z_A \, Z_B \, v \, (\vec{r}_A,\vec{r}_B) \quad . \tag{II-12a}$$

and we introduce a more compact notation

$$E(N) \;=\; \sum_i n_i \, h_{ii} + \frac{1}{2}\sum_{i,j} n_i \, n_j \, [(ij|v|ij) - (ij|v|ji)] + V_{Nucl} \quad . \tag{II.12b}$$

Now for the single–particle energy ε_i we obtain

$$\varepsilon_i \;=\; h_{ii} + \sum_\ell n_\ell \, [(i\ell|v|i\ell) - (i\ell|v|\ell i)] \tag{II-13}$$

which leads to

$$E^0(N) - E\,(N{-}1; n_i = 0) = \varepsilon_i \quad , \tag{II-14}$$

if we assume that the same set of $u_i(\vec{x})$ obtained for the N–electron system describes the system with one electron missing from state i. That is, if we do not take into consideration orbital relaxation effects. In (II-14) $E^0(N)$ is the variational ground state total energy for the N–electron system and, as such, minimum respective to u_i; and $E(N')$ is obtained with the u_i's taken from the N–electron variational set, with different occupation numbers. This interpretation of ε_i is known as the Koopmans[2] approximation and should work best for extended one–particle states (such as Bloch states). As a natural extension, the eigenvalue difference

$$\nu_{ij} \;=\; \varepsilon_i - \varepsilon_j \;\simeq\; E\,(N; n_j = 0, n_i = 1) - E^0(N) \tag{II-15}$$

could be taken as the excitation energy from the ground state to another N–electron configuration with one electron taken from state j and placed at state i (formerly unoccupied). Care must be taken, however, since even if if we calculated $E(N)$ selfconsistently, the variational procedure is only guaranteed to furnish the best u_i's for the ground state of the system in any given symmetry. As such, eigenvalues of virtual (unoccupied) states have no clear interpretation.

The Hatree–Fock approximation is the most important step in the direction of obtaining the electronic structure of a N–electron system. It suffers, however, of serious problems in that it cannot, by construction, treat a correlated N–electron system. If we resort to some product definition of $\Psi(\{\vec{x}_i\})$ as in (I-6), we may say that it is not realistic to describe even the ground state of the system with a single Slater determinant. For finite systems the inclusion of correlation effects may in principle be achieved through a superimposed variational procedure in the space of determinantal functions Ψ_I, known as configuration interaction (CI).

We express the wave function for a given Γ as a combination

$$\Phi_\Gamma \;=\; \sum_I A_{\Gamma I} \, \Psi_I \tag{II-16}$$

where the Ψ_I's are built by excitations from the ground state HF wave function Ψ_{HF} (in the spirit of (II-15) but including multiple excitations). These Ψ_I are then written in terms of the (orthogonal) variational set of u_i's by proper "switching" of occupation numbers. The correlated ground state is obtained through minimization of the energy $<\Phi|\mathcal{H}|\Phi>$ with respect to variation in $A_{\Gamma I}$, which then furnishes, also, the energy and wave functions for excited states.

For an infinite system this scheme is not applicable, and the problem arises of how

to represent the excitation of a single particle in the Bloch approximation. Hence, excited states cannot be properly treated in the HF–limit; ground state properties should, whenever correlation effects are not too important, be reasonably well reproduced.

We will return to these problems later in the chapter. I want to stress, however, that in the semiconductor physics we are interested in we will mostly be concerned with excitations, usually called quasi–particles: holes and electrons (where "electron" is to be taken as an electron in a given conduction or defect state), excitons, phonons. We may therefore expect the HF treatment to fail somewhere along the route: in fact, whenever correlation effects cannot be neglected.

Another point we must discuss is the treatment of spin effects. As derived above from (II–1) to (II–11), the procedure is called unrestricted HF and can become very complicated. The case of closed–shell systems, such as ideal crystalline semiconductors with completely occupied valence bands and empty conduction bands, poses no difficulties if we neglect relativistic effects. For open shells, if the overall symmetry is A_1, the usual procedure is known as "spin polarization": we use a different set of space orbitals for each spin[3], solving as it were two $\frac{1}{2}$ N–electron problems with only parallel spins involved [As a consequence the resulting wave function is not any longer a pure–spin state, due to non–orthogonality of orbitals]. If the symmetry is lower, however, several Slater determinants would have to be considered on equal footing, which may cause severe difficulties as we can have space–symmetries broken.

A very much used procedure, which is sometimes called "mean–field averaging", is to treat open shells (either spin or space) as if they were closed, with fractional occupation numbers: the charge of the open–shell orbitals is equally distributed among up– and down–spin orbitals, or among partners of a space–group representation, thus turning a low–symmetry problem into A_1–symmetric.

II.2. The LCAO or Tight–Biding Formulation of Hartree–Fock and Further Approximations.

The HF equations as described in the preceding section have to be solved in a self–consistent way. We must specify a reasonable set of u_i's as input, solve the determinantal equations to obtain the ε_i's and new u_i's, and so forth. The absence of the central symmetry characteristic of atomic systems causes a serious complication. However, a natural assuption (first introduced in this context by Bloch[4]) is that the constistutent atoms bring their identities with them to the molecule or solid. Thus, a good description of a molecular orbital u_j, extending in principle over the whole system, could be obtained[5] from a suitable linear combination of atomic orbitals ϕ_μ centered over the particular atomic sites (LCAO):

$$u_i(\vec{x}) = \sum_\mu c_{i\mu}\, \phi_\mu(\vec{x}) \;. \qquad (II–17)$$

The set $\left\{ \phi_\mu \right\}$ includes a complete (and usually orthonormal) set of atomic orbitals for each lattice atom. This first approximation leads to a Fock–operator matrix in terms of the atomic orbital basis:

$$f_{\mu\nu} = h_{\mu\nu} + \sum_{\lambda\sigma} P_{\lambda\sigma} \left[(\mu\nu|v|\lambda\sigma) - \frac{1}{2}(\mu\lambda|v|\nu\sigma) \right] \qquad (II–18)$$

where the bond–order or density matrix $P_{\lambda\sigma}$ is given by

$$P_{\lambda\sigma} = \sum_i n_i\, c_{i\lambda}^*\, c_{i\sigma} \qquad (II–19)$$

Variation of u_i's is achieved through variation of the coefficients $c_{i\mu}$, so that equations (II–5) are translated into the Hartree–Fock–Roothaan (HFR) equations[5]

$$\sum_{\mu} \left[f_{\mu\nu} - \varepsilon\, s_{\mu\nu} \right] c_{i\mu} = 0 \qquad \text{(II–20)}$$

which are subject to the normalization conditions

$$\sum_{\mu\lambda} c^*_{i\mu}\, c^*_{i\lambda}\, s_{\mu\lambda} = 1 \qquad \text{(II–21)}$$

where $s_{\mu\lambda}$ is the overlap integral between orbitals μ and λ.

The problem is simplified first by recognizing that the functions $u_i(\vec{x}) = u_{n\vec{k}}(\vec{x})$ must be invariant against decomposition into Bloch projections: if we build the operators

$$P_{\vec{k}} = \frac{1}{N_c} \sum_{\vec{t}} \exp\{i\vec{k}\cdot\vec{t}\}\, T(-\vec{t}) \ , \qquad \text{(II–22)}$$

where N_c is the number of cells in the crystal, we must have $P_{\vec{k}}\, u_{n\vec{k}}(\vec{x}) = u_{n\vec{k}}(\vec{x})$. Operating on (II–17) we note that

$$u_{n\vec{k}}(\vec{x}) = \sum_{\mu} c_{i\mu} \left[P_{\vec{k}}\, \phi_\mu \right] \ , \qquad \text{(II–23)}$$

where the subset $\left\{ P_{\vec{k}}\, \phi_\mu \right\}$ is complete relative to the subspace $\vec{k}$ if $\left\{ \phi_\mu \right\}$ is complete. This procedure corresponds to block–diagonalization of the energy and overlap matrices F and S according to the $\vec{k}$–star representation of the translation symmetry group of the

crystal. Each set $\left\{ P_{\vec{k}}\, \phi_\mu \right\}$ is orthogonal and non–interacting through F to the others, so that we end up as needed, with a set of equations (II–17) to (II–20) for each $\vec{k}$.

Further block–diagonalization of the matrices within each $\vec{k}$–star may follow from the use of symmetry–adapted combination functions according to the crystallographic point group of the crystal.

Another necessary simplification is the truncation of the complete infinite basis set. Formally, this may be accomplished as follows[3]: we separete from the set $\{\phi\}$ an inner, finite set $\{\phi^{in}\}$ so that the quantities are written as

$$\left[f_{\mu\lambda}\right] = \begin{bmatrix} F^{in,in} & F^{in,out} \\ F^{out,in} & F^{out,out} \end{bmatrix}, \quad \left[s_{\mu\lambda}\right] = \begin{bmatrix} S^{in,in} & S^{in,out} \\ S^{out,in} & S^{out,out} \end{bmatrix}, \quad c = \begin{bmatrix} C^{in} \\ C^{out} \end{bmatrix}$$

$$\text{(II–24)}$$

where the blocks (in,in) are of finite dimensions. Now (II–20), in matrix notation, is separated in

$$F^{in,in}\, C^{in} + F^{in,out}\, C^{out} = \varepsilon\, S^{in,in}\, C^{in}$$

$$\text{(II–25)}$$

$$F^{out,in}\, C^{in} + F^{out,out}\, C^{out} = \varepsilon\, S^{out,out}\, C^{out}$$

if we approximate the overlap $S^{in,out} = S^{out,in} = 0$. Solving for C^{out} and reinserting, we obtain the renormalized expression

$$\tilde{F}^{in,in}\, C^{in} = \varepsilon\, S^{in,in}\, C^{in} \qquad \text{(II–26)}$$

with

$$\tilde{F}^{in,in} = F^{in,in} + F^{in,out} \left[\varepsilon\, S^{out,out} - F^{out,out} \right]^{-1} F^{out,in} \qquad \text{(II–27)}$$

which allows us to solve a finite determinantal equation for each group representation, with the second term in (II—27) as a perturbation to be gauged and truncated.

If the procedure were carried out in **ab inition** form, with no further approximations, it would still become unmanageable for rather modest systems. We recall that the resolution of (II—20) implies by (II—18) the computation of multicentered integrals $(\mu\nu|v|\lambda\sigma)$, the number of which grows roughly with the fourth power of the basis size. Some immediate remedies are easily seen, still an ab initio standard, such as the truncation of the basis set at some reasonable convergence—tested size, and the use of analytical functions for the atomic basis, e.g. Slater—type—orbitals or suitably contracted Gaussian—type—orbitals. However, we recall that even if the effort is stretched to some mythical limit, there will still be errors in the final results due to the neglect of correlation: to eliminate these errors we have to move out of HF mean—field theory (through, e.g., inclusion of configuration interaction) which would quite obviously further complicate the problem.

A whole family of techniques — the NDO techniques — were born from the view that the operator in (II—18) could be simplified and results improved at the same time, through convenient semiempirical parametrization of the matrix elemets[6].

Althought these methods were devised for the study of molecules. they have been applied in semiconductor physics (mostly for the study od deep—level defects) hence I will briefly describe the spirit of the approximations.

The first step is to treat the real overlap between different orbitals μ and ν as if it were equal to zero (neglect of differential overlap — NDO):

$$\phi_\mu^*(\mathfrak{x})\,\phi_\nu(\mathfrak{x})\,d\mathfrak{x} \simeq \delta_{\mu\nu} \quad . \tag{II—28}$$

and thus eliminate the computation of three— and four—center integrals, and many of the one— and two—center integrals. Remaining integrals are then parametrized under strict

rules: (i) parameters have to be associated to individual atomic species, and be transferrable from one system to another; and (ii) results should be invariant under rotation of the atomic basis.

Furthermore, atomic basis sizes are usually reduced to the minimum possible to accomodate the existing electrons (plus degeneracies, e.g., two s and six p orbitals for the valence shell of Si,Ga or As, with inner electrons included in the core).

Different techniques aim at describing different properties of the systems. The first scheme to be established, CNDO[7a] (complete NDO), eliminated by II—28 all exchange integrals and is not well suited to treat open—shell systems. INDO[7b] (intermediate NDO) brought back into the calculations exchange integrals for same—center orbitals, and is directed at electronic structure results. MNDO[7c] (modified NDO) alters the way some matrix elements are alculated and aims at obtaining stable geometries and vibrational properties.

All these techniques have in common the use of the NDO approximation to some degree. Matrix elements are parametrized in terms of the atomic species, through properties of free atoms such as electron affinity and ultra—violet spectrum. The quantity that depends on fitting to results for molecular systems is the **bonding parameter** β: all NDO techniques bring back the real calculated overlap integral $s_{\mu\nu}$—ultimately responsible for bonding — to parametrize the one—electron hamiltonian between orbitals μ and λ centered on different atoms A and B respectively:

$$h_{\mu\lambda} = \mathcal{F}(\beta_A,\beta_B)\,s_{\mu\lambda} \quad . \tag{II—29}$$

The fit is usually carried out with reference to several different systems and already within the CI variational formulation (so as to avoid including correlation effects twice).

Results for Si were obtained through MNDO[8] and CNDO[9], this latter in a large unit cell approach. Until very recently[10] no parameters were available for Ga or As, so

there are no calculations of band structure for GaAs.

As said above, the NDO techniques are mostly used for molecular systems. The problem of solving multicenter integrals is however of so severe order that some approximation is needed. The most familiar form in band–structure calculation is know as the **empirical tight–binding** (TB) method. The TB method is mainly used for obtaining the curves ε_{nk} along symmetry lines in the Brillouin zone through interpolation between high–symetry points, where more sophisticated techniques are used[11]. Non–empirical versions of TB are not usually carried out within Hartree–Fock formalism, and will not be discussed in this section.

The **differential overlap** problem is formally circumvented in the TB method by transforming form the $\left\{ \phi_\mu \right\}$ set to a set of orthogonalized atomic orbitals $\left\{ \psi_\mu \right\}$ through, e.g., a Lowdin transformation[3,11]

$$\psi_\nu = \sum_\mu s_{\nu\mu}^{-\frac{1}{2}} \phi_\mu \ . \tag{II–30}$$

Bloch projections are then formed from this set, and a general element of the Fock matrix in the subspace $\vec{k}$ is written (if we recall that the vectors $\vec{t}$ connect lattice sites I):

$$f_{\mu\nu}(\vec{k}) = \sum_J \exp\left[i\vec{k}\cdot\left[\vec{R}_J - \vec{R}_I\right] \right] \int \psi_\mu \left[\vec{x} - \vec{R}_I\right] f \, \psi_\nu \left[\vec{x} - \vec{R}_J\right] d\vec{x} \tag{II–31}$$

if $\vec{R}_L$ are the position vectors of the lattice sites, and $\left[\vec{x} - \vec{R}_L\right] = \left[\vec{t} - \vec{R}_L, \vec{\xi}\right]$; we take $\vec{R}_I$ to be one of the sites in the central unit cell. The usual difficulties are met in a full calculation of **(II–31)**. For instance, each ψ_μ is a combination of ϕ_ν's, the Fock operator

is itself dependent on the ψ's, and we will finally arrive at four center integrals. However, each term in the sum **(II–31)** can be formally assigned to a pair of orbitals centered at given sites, and this convenient feature is the basis of the empirical TB scheme.

The following approximations are made; (i) The form **(II–31)** is normally used, but the integrals are replaced by parameters which are to be adjusted to results from more accurate calculations (or experimental data) for high–symmetry points in the BZ. (ii) To guarantee that the number of parameters is compatible with the amount of data available, integrals between orbitals centered at $\vec{R}_I$ and $\vec{R}_J$ are calculated if $|\vec{R}_I - \vec{R}_J|$ is smaller than a prescribed value, and all other integrals are set to zero (we then work within a nearest–neighbor approximation, or we keep also the second–nearest–neighbor interactions, and so forth).

Apart from (ii), integrals equivalent by symmetry (such as those between s–type orbitals on the central atom and its four nearest neighbors) are identified through the symmetry operations of the point group of the crystal, T_d for tetrahedral semiconductors.

If we apply this scheme to the Si or GaAs case, including up to second–nearest–neighbor integrals and Bloch projections of s and p types, there will remain 8 independent matrix elements **(II–31)**, which can be written in terms of 13 independent integrals. These 13 parameters are then adjusted to known values of ε_{nk} at chosen $\vec{k}$ points, e.g. $\vec{k} = 0$ (Γ), $\vec{k} = \frac{\pi}{a}(100)$ (X), and $\vec{k} = \frac{\pi}{a}(111)$ (L); the whole behavior of ε_{nk} may then be obtained.

However, as said above, the use of the empirical TB method is tied to the existence of more accurate calculations for some high–symmetry points.

II.3. The Local Density Approximation LDA

Most of the difficulties associated with the Hartree–Fock approach are introduced by the non–local exchange term. In 1951 Slater[12] proposed an approximation to the

one–particle equations (II–5) and (II–12) substituting the exchange potential by a local term. The local exchange is calculated as the statistical average of the HF exchange over occupied states $u_i(\vec{x})$; this form is further approximated to the statistical exchange of an uniform electron gas of density $\rho(\vec{r})$:

$$V_{SX} = -3e^2 \left[\frac{3}{8\pi} \rho(\vec{r}) \right]^{1/3} \qquad (II\text{--}32)$$

for a closed–shell system. The local approximation was later discussed by Hohenberg and Kohn[13] who proved that the ground–state energy of a N–electron system is indeed a unique functional of the density $\rho(\vec{r})$, and laid the basis for Density Functional Theory (DFT). In particular, for the uniform electron gas Kohn and Sham[14] found that the exchange term should be only $\frac{2}{3}$ of V_{SX}. This suggested the use[15] of a term $V_{X\alpha} = \alpha V_{SX}$, with the exchange coefficient α bounded $\frac{2}{3} \le \alpha \le 1$. From then up to this date several improved models have been proposed for the exchange–correlation potential V_{XC}, and we will return to this point later. For now we will assume that some form of V_{XC} is given and discuss the implications of LDA.

We write as in Ref. 14 the total energy for the ground state of the interacting electron gas:

$$E^0(N) = T[\rho(\vec{r})] - \sum_J \int \rho(\vec{r}) \, v\left[\vec{r},\vec{R}_J\right] d\vec{r} +$$

$$\qquad (II\text{--}33)$$

$$+ \frac{1}{2} \int \rho(\vec{r}_1) \, \rho(\vec{r}_2) \, v(\vec{r}_1,\vec{r}_2) \, d\vec{r}_1 \, d\vec{r}_2 + E_{XC}[\rho(\vec{r})]$$

in the slowly–varying–density regime where

$$E_{XC}[\rho(\vec{r})] = \int \rho(\vec{r}) \, \varepsilon_{XC}[\rho(\vec{r})] \, d\vec{r} \qquad (II\text{--}34)$$

in terms of the exchange–correlation energy per electron ε_{XC}. The one–particle equation that determines $\rho(\vec{r})$ is in turn given by

$$\left[-\frac{\hbar^2}{2m} \nabla^2 - \sum_J v\left[\vec{r}_1,\vec{R}_J\right] + \int \rho(\vec{r}_2) \, v(\vec{r}_1,\vec{r}_2) \, d\vec{r}_2 + V_{XC}[\rho(\vec{r}_1)] \right] u_i(\vec{r}_1) = \varepsilon_i \, u_i(\vec{r}_1) \, ,$$

$$\qquad (II\text{--}35)$$

with

$$V_{XC}[\rho(\vec{r})] = d\{\rho(\vec{r}_1) \, \varepsilon_{XC}[\rho(\vec{r}_1)]\}/d\rho(\vec{r}_1) \, . \qquad (II\text{--}36)$$

Equation (II–35) is now a local equation for the u_i's that must be solved self–consistently since the effective hamiltonian depends on $\rho(\vec{r})$ through ε_{XC}.

We may now comment on some characteristic aspects of an actual LDA calculation. Each equation (II–35) depends on all ε_i's in a non–linear way through V_{XC}, so the determinantal equation must be solved through numerical iteration every self–consistent cycle, by supplying both a guess set of u_i's and of matching ε_i's ; a search for the zeroes of the secular determinant then follows. When satisfactory sets of u_i's and ε_i's are found, they are used for building the potentials for the next cycle and so forth. Hence, as a general rule, LDA calculations are lengthty.

Another relevant feature of LDA is that since the effective hamiltonian depends on the local charge density, (II–7) must be evaluated each cycle:

$$\rho(\vec{x}) = \sum_n^{occ} \sum_{\vec{k}}^{BZ} |u_{n\vec{k}}(\vec{x})|^2 \qquad (II\text{--}7')$$

(where the sum runs over occupied states only). This implies an extensive sum or integration over the Brillouin Zone. Alternatively, the integration is substituted by a

sampling of the BZ at special points, followed by an averaging procedure. This sampling is thus of fundamental importance for the success of the calculation! The choice of points and weighing procedure may be carried out taking into account the point–symmetry of the crystalline lattice[16] and as a result only a small finite number of $\vec{k}$–points need be explicitly included in (II–7).

As a first example of band structure LDA calculations we show in Table I, compared to experimental results, the results of Ciraci and Batra[17] for Si obtained through the self–consistent first–principles TB approach. The LCAO equations (II–20) are fully solved, after redefining $f_{\mu\nu} \rightarrow f_{\mu\nu}^{LDA}$ in (II–31), with $\alpha = \frac{2}{3}$. The basis set consists of 70 contracted Gaussian–type orbitals of $\ell = 0,1$ and 2 character (s,p and d–type) and up to 5th–neighbor interactions are included. It may be verified that relevant enrgies such as the gap and valence–band–width, and some transitions (detected as reflectance peaks) such as $L_3' - L_3$ are in very good agreement with experiment.

Table I – Energies (in eV) for some relevant electronic states or transitions in crystalline silicon as calculated through different models, compared to experimental values.

	Experiment	Ciraci and Batra Ref.17	Stuckel and Euwema Ref.18		Zunger and Cohen Ref.19
Theoretical Method	——	first–principles tight–binding	first–principles OPW $\alpha = 1$	$\alpha = 2/3$	first–principles pseudo–potentials
Gap	1.14	1.2	1.10	≤ 0.32	0.5
Valence band width	$(12.4 \pm 0.6)^a$	11.7	11.74	12.04	12.20
Direct gap $(\Gamma_{25}'-\Gamma_{15})$	3.45[b]	3.1	2.75[c]	2.33	2.48
$L_3' - L_3^{cond}$	5.48[b]	5.2	5.01	4.38	4.64

a) Ref.20   b) Ref.21   c) in this case, the value corresponds to $\Gamma_{25}' - \Gamma_2'$.

We will discuss in the next section some other LDA results for the band structure of Si (already included in Table I) and GaAs; before doing that, we will take a closer look at LDA itself. It is based on the possibility of writing the total energy for the ground state of a many–particle system in terms of a universal functional of the density, $F[\rho(\vec{r})]$. That is, for an arbitrary external potential $v(\vec{r})$ the correct ground state energy is the minimum value of

$$E[\rho(\vec{r})] \; = \; \int \, v(\vec{r}) \, \rho(\vec{r}) d\vec{r} \, + \, F[\rho(\vec{r})] \; . \qquad (II-37)$$

We may write the energy as a functional of the density since it can be proved[13] that $v(\vec{r})$ is a functional of $\rho(\vec{r})$ for the correct ground–state density. In (II–37), F is written from the hamiltonian $\mathcal{H} = T + V + U$ with

$$T \; = \; \frac{\hbar^2}{2m} \int \vec{\nabla} \, \Psi^+(\vec{r}) \, \vec{\nabla} \, \Psi(\vec{r}) d\vec{r}$$

$$V \; = \; \int \Psi^+(\vec{r}) \, v(\vec{r}) \, \Psi(\vec{r}) d\vec{r} \qquad (II-38)$$

$$U \; = \; \frac{1}{2} \int \Psi^+(\vec{r}_1) \, \Psi^+(\vec{r}_2) \, v(\vec{r}_1,\vec{r}_2) \, \Psi(\vec{r}_1) \, \Psi(\vec{r}_2) \, d\vec{r}_1 d\vec{r}_2$$

where $\Psi^+(\vec{r}), \Psi(\vec{r})$ are field operators for creation or anihilation of a particle at $\vec{r}$. The general density at $\vec{r}$ is defined as

$$\rho(\vec{r}) \; = \; <N| \int d\vec{\xi} \, \Psi^+(\vec{x}) \, \Psi(\vec{x}) \, |N> \qquad (II-39)$$

introducing the general N–particle state $|N>$. The problem now is to find the functional $F[\rho(\vec{r})]$

$$F[\rho(\vec{r})] \; = \; <N|T+U|N> \qquad (II-40)$$

which is still in a non–local formulation. We may separate the problem in a known Coulomb functional of the local density

$$F[\rho] \; = \; \frac{1}{2} \int \, v(\vec{r}_1,\vec{r}_2) \, \rho(\vec{r}_1) \, \rho(\vec{r}_2) \, d\vec{r}_1 \, d\vec{r}_2 + G[\rho] \qquad (II-41)$$

and expand the unknown functional $G[\rho]$ in higher–order density–matrices

$$G[\rho] \; = \; G^0[\rho(\vec{r}_1,\vec{r}_2)] + G'[\rho(\vec{r}_1,\vec{r}_2;\vec{r}_3,\vec{r}_4)] \qquad (II-42)$$

and so forth. The need for approximations is clear, since G is still unknown. LDA derives from the proposal[14] to write

$$G[\rho] \; = \; T_S[\rho] + E_{XC}[\rho] \qquad (II-43)$$

where $T_S[\rho]$ is a simplified kinetic energy functional valid for the gas of non–interacting electrons. From this step on, equation (II–35) is directly recovered, with the main result that V_{XC} is unknown but local $V_{XC}[\rho(\vec{r})]$. Several functionals have been proposed and are in current use[1,22] for V_{XC} , apart from the X_α potentials.

A very important feature inherent to LDA is that the eigenvalues ε_i no longer carry the physically appealing interpretation of negative ionization energies: it is easily seen from II–33 to II–36 that

$$\varepsilon_i \; = \; \frac{\partial}{\partial n_i} [E^0(N)] \; . \qquad (II-44)$$

The problem of how to obtain one–particle ionization and excitation energies has been object of several studies. Slater[2] showed that, to a good approximation, the difference in total LDA energies for two configurations of a N–electron system might be given by:

$$E_{LDA}(N \; ; \; n_j = 0 \; , \; n_i = 1) - E^0_{LDA}(N) = \varepsilon_i(n_i = 1/2) - \varepsilon_j(n_j = 1/2) \quad \textbf{(II-45)}$$

where the difference in eigenvalues is to be extracted from a self–consistent calculation in a transition–state of the system in which the electron is equally shared by the single–particle states defining the excitation. The same argument holds for an ionization transition

$$E^0_{LDA}(N) - E^0_{LDA}(N-1 \; ; \; n_i = 0) = \varepsilon_i(n_i = 1/2) \quad . \qquad \text{(II-46)}$$

The interpretation of the curves $\varepsilon_{n\vec{k}}$ (calculated, of course, with filled bands) is then not so straightforward as in the HF formalism. It is expected, however, that the extended band states will not be much affected by the "absence" of half an electron so that a "LDA analogue of Koopmans result" might hopefully hold. For localized states the transition state construct should be used.

II.4. The OPW and Pseudopotential Methods

In (II–1) we assume that the most natural expansion of the single particle states u_i would be in terms of atomic orbitals. In a periodic system, however, another choice seems also natural, expansion in terms of periodic functions. The nearly–free–electron model envolves from this idea: the movement of the valence electrons in a solid shoud not be much different — apart from scattering events — from the movement of independent electrons in a box or in "periodic free space", provided the screened core potential is weak. It would then seem that the expansion

$$u_i(\vec{r}) = \sum_{\vec{k}} a_{n\vec{k}} \, [\Omega^{-1/2} \exp(i\vec{k}\cdot\vec{r})] \quad , \qquad \text{(II-47)}$$

here $\Omega = N_c \Omega_c$ is the volume of the crystal, could be useful. We write the periodic potential of the solid as a sum of atomic or core potentials, each already including some form of local exchange term (unspecified for the moment)

$$V(\vec{r}) = \sum_J^{N_A N_c} V_c(\vec{r} - \vec{R}_J) \qquad \text{(II-48)}$$

and substitute (II–47) and (II–48) in (II–35) to obtain the secular equations for the coefficients and single–particle energies:

$$\left[-\frac{\hbar^2 k^2}{2m} - \varepsilon_i \right] a_{\vec{k}} \delta_{\vec{k}\vec{k}'} + \frac{1}{\Omega} \sum_{\vec{k}} a_{\vec{k}} \sum_J^{N_A N_c} \exp\{-i(\vec{k}'-\vec{k})\cdot\vec{R}_J\} \times$$

$$\times \int d\vec{r} \; \exp\{-i(\vec{k}'-\vec{k})\cdot(\vec{r}-\vec{R}_J)\} \, V_c(\vec{r}-\vec{R}_J) = 0 \quad . \qquad \text{(II-49)}$$

It may be recognized that the structure factor

$$S_{\vec{k}\vec{k}'} = \frac{1}{N} \sum_J^N \exp\{-i(\vec{k}'-\vec{k})\cdot\vec{R}_J\} \qquad \text{(II-50)}$$

vanishes unless the difference $\vec{k}'-\vec{k}$ equals a lattice wavenumber $\vec{g}$, as is to be expected, and for an actual calculation we should only decide where to truncate (II–49) (in the spirit of (II–27)). It turns out, however, that the convergence of (II–47) is extremely slow for any real crystal, and this is due to the need of describing $u(\vec{r})$ for $\vec{r} \simeq \vec{R}_J$, the close proximity of the atomic cores. The existence of the lower–energy core states, extremely

localized in $\vec{r}$, forces on the valence states very strong short—wavelength oscillations, that is, strong components of large $|\vec{k}'-\vec{k}|$ in (II–47). Estimates[23] of the size of (II–49) for each $\vec{k}$ lie in the range of $\geq [10^6 \times 10^6]$ for Aluminium, and as such the method is not useful at all.

The difficulties of the plane—wave method arise in the neighborhood of the core, while in the other regions (even in the <111> bond direction) the expansion (II–47) should and does work quite well.

The orthogonalized—plane—wave (OPW) method is partly based on the fact that the tight—binding expansion (II–23) works very well for core states, for which there is virtually no differential overlap. We may write for ϕ_μ the usual form (neglecting unspherical deformations)

$$\phi_\mu(\vec{r}) = \phi_{n\ell m}(\vec{r}) = Y_{\ell m}(\Theta,\varphi) R_{n\ell}(r) \qquad (II–51)$$

where $R_{n\ell}(r)$ is the radial part of $\phi_{n\ell m}$, found through adopting some LDA potential, and the crystal core wave functions may be put in the form

$$u^c_{n\ell m,\vec{k}}(\vec{r}) = [N_A N_c]^{1/2} \sum_{J}^{N_A N_c} \exp\{i\vec{k}\cdot\vec{R}_J\} \, \phi_{n\ell m}(\vec{r}-\vec{R}_J) \, . \qquad (II–52)$$

The new basis for expansion is formed from OPWs: periodic functions $\psi_{\vec{k}}(\vec{r})$ which already incorporate the oscillations in the core region, by requiring that they are orthogonal to core states:

$$\psi^{OPW}_{\vec{k}}(\vec{r}) = \Omega^{-1/2} \exp\{i\vec{k}\cdot\vec{r}\} - \sum_{n\ell m} b_{n\ell m,\vec{k}} \, u^c_{n\ell m,\vec{k}} \qquad (II–53)$$

with the coefficients $b_{n\ell m,\vec{k}}$ chosen from

$$\int d\vec{r} \left[\psi^{OPW}_{\vec{k}}(\vec{r}) \right]^* u^c_{n\ell m,\vec{k}} = 0 \qquad (II–54)$$

for every core state (n,ℓ,m) of wave—vector $\vec{k}$ in the crystal.

The specific form $b_{n\ell m,\vec{k}}$ still depends on the choice of $R_{n\ell}(\vec{r})$, but the dependence on the magnetic quantum number m can be made to vanish.

The OPW's are now used in (II–47), so that

$$u_{n\vec{k}}(\vec{r}) = \sum_{\vec{g}} a_{(n\vec{k})\vec{g}} \, \psi^{OPW}_{\vec{k}-\vec{g}}(\vec{r}) \qquad (II–55)$$

and substitution in (II–35) yields the analogue of (II–49).

It is to be expected that a few OPW's in expansion (II–35) will be sufficient and indeed the band structure of Si was obtained using of the order of 3×10^2 OPW's for each $\vec{k}$ point by Stückel and Euwema[18], and $|\vec{k}-\vec{g}|_{MAX} \sim 1.5\pi$ (a.u.) in (II–55).

The authors use the self—consistent scheme in their calculations (as all electronic structure methods, the OPW may be used in a non—self—consistent semiempirical version[24]). It is important to mention that the orthogonalization procedure (II–54) is repeated for each interation in the self—consistency process, since it is allowed that the core states change with the valence—electron rearrangements. As such, the separation in valence and core electrons in the SC OPW method is rather formal, and it is possible to perform "all—electron" OPW calculations[18]. We enter in Table I results for the same calculation using Slater or Kohn—Sham exchange potentials: is is worth noting the large difference in the width of the gap, and the inversion of the order $\Gamma_{15} \leftrightarrow \Gamma_2'$, obtained with

use of one or other exchange potentials.

Summarizing, the idea of the OPW method is to choose an expansion for the single particle states that converges very rapidly, through taking explicity into account the strength of the potential in the regions very close to the atomic cores. As a consequence, the solutions $u_{n\vec{k}}$ all properly exhibit the complicated oscillating behavior at $\vec{r} \sim \vec{R}_J$.

The **pseudopotential method** on the other hand, is directed at finding an expansion which also converges very rapidly, and still consists of smoth, well–behaved functions over all space including the core regions.

The pseudopotential method is usually derived[25] from **(II–53)** (we change notations to

$$|OPW,\vec{k}> \ = \ |\vec{k}> - \sum_{n\ell m, J} |n\ell m, J> <n\ell m, J|\vec{k}> \qquad (II-53')$$

to simplify the writing). The expansion **(II–55)** is written as

$$u_{\vec{k}} \ = \ \sum_{\vec{g}} a_{\vec{k},\vec{g}} \ |OPW, \vec{k}-\vec{g}> \ = \ (1-P_c) \sum_{\vec{g}} a_{\vec{k},\vec{g}} \ |\vec{k}-\vec{g}> \qquad (II-56)$$

where the operator P_c projects the plane waves onto the core functions as in **(II–53')**, and we interchange the sums in $(n\ell m)$ and $\vec{g}$. The pseudofunction

$$\psi_{\vec{k}}^{PS}(\vec{r}) \ = \ \sum_{\vec{g}} a_{\vec{k},\vec{g}} \ [\Omega^{-1/2} \exp\{i(\vec{k}-\vec{g})\cdot\vec{r}\}] \qquad (II-57)$$

is smooth and well–behaved in all space. On the other hand, the operator P_c only "acts" within spheres surrouding the atomic cores: outside these regions the pseudofunction is

equal to the "true" wave function $u_{\vec{k}}(\vec{r})$. Inserting **(II–56)**, in the form $u_{\vec{k}} = (1-P_c)\psi_{\vec{k}}^{PS}$, in **(II–35)** we obtain

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\right]\psi^{PS} - \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\right]P_c\psi^{PS} \ = \ E(1-P_c)\psi^{PS} \qquad (II-58)$$

with $V(\vec{r})$ accounting for all potential terms. If we may assume that

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\right]\phi_{n\ell m}(\vec{r}) \ = \ \varepsilon_{n\ell m} \, \phi_{n\ell m}(\vec{r}) \qquad (II-59)$$

we arrive at the pseudopotential equation:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V^{PS}(\vec{r},\varepsilon)\right]\psi^{PS}(\vec{r}) \ = \ \varepsilon \, \psi^{PS}(\vec{r}) \qquad (II-60)$$

with

$$V^{PS}(\vec{r},\varepsilon) \ = \ V(\vec{r}) + \sum_{n\ell m, J}\left[\varepsilon - \varepsilon_{n\ell m}^{(J)}\right]|n\ell m, J> <n\ell m, J| \ . \qquad (II-61)$$

In this form the pseudopotential method has limited application, because **(II–59)** is not valid in general [$V(\vec{r})$ is not even spherical except in the free atom]. The overall simplicity of **(II–60)** has been interpreted, however, as a freedom in choosing $V^{PS}(\vec{r})$. The empirical pseudopotential method in its several versions is based on this freedom: $V^{PS}(\vec{r})$ is chosen in the simplest form possible which produces: (i)$\varepsilon_{n\vec{k}}$'s adjusted to experimental values for some special points in the BZ, and (ii) smooth transition of the wave function or its logarithimic derivative over the radius of influence of P_c . These empirical versions were very important in the systematic study of several

semiconductors[26], however I will concentrate here on the derivation of the first–principles method.

The basic assumption[19] is that the all electron hamiltonian in (II–35)

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + V(\vec{r}) \right] u_i^{c+v} = \varepsilon_i \, u_i^{c+v} \tag{II–62}$$

where u^{c+v} encompasses core and valence states, can be replaced by

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + V_v(\vec{r}) + V_c^{PS}(\vec{r},\varepsilon) \right] u_j^{PS} = \lambda_i \, u_j^{PS} \tag{II–63}$$

where the pseudofunctions u_j^{PS} only describe valence electrons and may be obtained from the "true" set by a unitary transformation:

$$u_i^{PS} = \sum_j c_{ij} \, u_j^{c+v} \quad . \tag{II–64}$$

The core pseudopotential V_c^{PS} is to be obtained by requiring that $\lambda_i = \varepsilon_i$ for all valence states, maintaining for $V_v(\vec{r})$, the two–electron potential for valence electrons, the same functional form as in $V(\vec{r})$. This is just a rephrasing of II–60 and does not simplify the problem unless we assume that once $V_c^{PS}(\vec{r},\varepsilon)$ is found for some simple situation for a given atom, it may be transfered together with that atom to any other complicated situation. That is, unless we assume that core states are not affected by valence–electron rearranjements from one system to the other, e.g., As 3p states behave the same way in GaAs, AlAs or in the free As neutral atom. In this case, we are entitled to calculate V_c^{PS} in the free atom, and insert it in a band–structure problem: this core

potential is not altered by the self–consistent procedure (it is called a "frozen" core). However in the initial calculations for the atom the full requisites are satisfied.

We thus consider a pseudoatom in such a reference configuration (usually the ground state), with the valence electrons interacting through Coulomb and exchange–correlation terms and moving in the external field of the core

$$V^{ext}(\vec{r}) = -\frac{Z_v e^2}{r} + V_c^{PS}(\vec{r}) \tag{II–65}$$

where Z_v is the effective core charge. The total energy will be given by (II–33) with $\rho(\vec{r})$ replaced by $\rho^{PS}(\vec{r})$, obtained by substituting $u_i^{PS}(\vec{r})$ in (II–7') and summing only over the valence electrons. The single–particle energies are obtained from (II–63), and in the spherical symmetry of the free atom we write (II–64) as

$$u_{n\ell}^{PS}(\vec{r}) = \sum_{n'} C_{nn'\ell} \, u_{n'\ell}(\vec{r}) \quad . \tag{II–66}$$

The searched solutions for equal all–electron and pseudo eigenvalues are

$$V_c^{PS}(\vec{r},\ell) = V_\ell(\vec{r}) + V[\rho(\vec{r})] - V_v[\rho^{PS}(\vec{r})] \tag{II–67}$$

where the difference in potential energies is calculated numerically and

$$V_\ell(\vec{r}) = \frac{\sum_{n'} c_{nn'\ell} \left[\varepsilon_{n\ell} - \varepsilon_{n'\ell} \right] u_{n'\ell}(\vec{r})}{\sum_{n'} c_{nn'\ell} \, u_{n'\ell}(\vec{r})} \quad . \tag{II–68}$$

The transformation (II–66) is not unique and the choice of $V_c^{PS}(\vec{r})$ is subject to "desirable" properties of $u_i^{PS}(\vec{r})$. For instance[19], it is understood that $u^{PS}(\vec{r})$ should be nodeless for the lowest valence state of each symmetry; that it should deviate as little as possible from $u_i(\vec{r})$ in the bonding regions; or other alternative constraints[27].

I insert in Table I results for the band structure of Si obtained by Zunger and Cohen[19] using the first–principles pseudopotential method. The exchange–correlation treatment is somewhat different from statistical exchange. The ℓ–components of V_c^{PS} were kept up to $\ell = 2$, and the plane–wave expansion requires $\sim 3\times10^2$ terms. It is seen that the direct gap transition is $\Gamma_{25}' - \Gamma_{15}$; the indirect gap energy is only 0.5 eV, or less than half the experimental value.

I want to stress that the basic characteristics of valence band structure are very similar, as obtained by any of the methods discussed here. Since LDA calculations must provide $\rho(\vec{r})$, valence–charge contour diagrams can be compared to experimental X–ray data: results compare very well for both empirical and first–principles pseudopotential calculations, for Si[19,26] and GaAs[26]. The main discrepancies between methods appear (see Table I) for virtual states: the $\Gamma_2' - \Gamma_{15}$ ordering, the energy gap values. The work of Stuckel and Euwema emphasizes[18] that the choice of V_{XC} may affect these characteristics. In fact, it was realized that the "band–gap problem" was closely related to correlation effects.

As seen in the last section, the eigenvalues of the single–particle equations in DFT indeed do not correspond to ionization energies, and the transition state construct is only an approximation. However, it was not clear until recently if the failure of LDA calculations with respect to virtual states was to be entirely ascribed to DFT, or if most of the problem came from the local approximation. To gauge the different effects it is necessary to move out of LDA and DFT, which task was undertaken by sevral

workers[28,29]. It now seems that the greater part of the LDA errors are already to be found in DFT[29].

II.5. Structural Properties

We know that the total energy for the electronic ground state should be in principle obtainable within DFT. The main goal is to obtain sufficiently accurate numbers for cohesive energies, lattice constants, or to predict stable structures and structural transitions. We are dealing with huge numbers, and it is reasonable to define the total energy per atom, or per unit cell in a periodic arrangement.

The computation of the total energy is carried out in different ways for different methods. I will only discuss the pseudopotential method, and will follow closely the work of Ihm et al[30] in deriving the equations.

The calculation is feasible mainly because of the frozen–core approximation implicit in pseudopotentials: the core electrons should not have a direct influence on cohesive energies or structural stability; the indirect influence, through core–orthogonalization of valence states, is accounted for reasonably well by the pseudopotential. Apart from that, translational symmetry is again used to simplify the calculations, carried out in reciprocal space.

The authors write the LD total energy (II–33) for the pseudoquantities and include core–core repulsions in:

$$E^0 = T[\rho^{PS}(\vec{r})] + \sum_J V^{PS}(\vec{r},\vec{R}_J) + \frac{1}{2} \int \rho^{PS}(\vec{r}_1)\,\rho^{PS}(\vec{r}_2)\,v(\vec{r}_1,\vec{r}_2)\,d\vec{r}_1 d\vec{r}_2 +$$

$$+ E_{XC}[\rho^{PS}(\vec{r})] + \frac{1}{2}\sum_{I,J} Z_v\,v(\vec{R}_I,\vec{R}_J)$$

(II–69)

where

$$T[\rho^{PS}(\vec{r})] = -\frac{\hbar^2}{2m} \sum_i \int u_i^{PS*}(\vec{r}) \nabla^2 u_i^{PS}(\vec{r}) d\vec{r} \qquad \text{(II--70)}$$

$$V^{PS}(\vec{r},\vec{R}_J) = \sum_{i,\ell} \int u_i^{PS*} U_\ell^{PS}(\vec{r},\vec{R}_J) \hat{P}_\ell u_i^{PS}(\vec{r}) d\vec{r} \quad . \qquad \text{(II--71)}$$

Equation II-69 is simplified, once we know the band structure of the solid, to the following expression (here explicit use is made of the X_α approximation where $\varepsilon_{XC} = \frac{3}{4} V_{XC}$):

$$E^0 = \sum_i \varepsilon_i - \frac{1}{2} \int \rho^{PS}(\vec{r}_1) \rho^{PS}(\vec{r}_2) v(\vec{r}_1,\vec{r}_2) d\vec{r}_1 d\vec{r}_2 +$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(II--72)}$$

$$-\frac{1}{4} V_{XC}[\rho^{PS}(\vec{r})] \rho^{PS}(\vec{r}) d\vec{r} + \frac{1}{2} \sum_{I,J} Z_v v(\vec{R}_I,\vec{R}_J) \quad .$$

Use is then made of the translational invariance by transforming (II--66) to $\vec{k}$--space. For this purpose, a plane--wave expansion for u_i^{PS} is assumed, and we arrive at:

$$E^0 = \sum_i \varepsilon_i - \Omega \left[\frac{1}{2} \sum_{\vec{g}} V_H(\vec{g}) \rho^{PS}(\vec{g}) + \frac{1}{4} \sum_{\vec{g}} V_{XC}(\vec{g}) \rho^{PS}(\vec{g}) \right] +$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(II--73)}$$

$$+ \frac{1}{2} \sum_{I,J} Z_v v(\vec{R}_I,\vec{R}_J) \quad .$$

There are some mathematical problems in computing E_0 from this equation since, for instance, the Coulomb component

$$V_H(\vec{g}) = \frac{4\pi e^2 \rho(\vec{g})}{g^2} \qquad \text{(II--74)}$$

diverges for $\vec{g} = 0$, and the same happens to the corresponding term for the pseudopotential, which is not seen explicitly in II--67 but is implicit in $\rho(\vec{g})$ and in the eigenvalues. The core repulsion energy is also divergent as known. After some lengthy algebra, the final result is put in the form

$$E^0 \text{ (per atom)} = \frac{1}{N_A N_c} \sum_i \varepsilon_i - \frac{1}{2} \frac{\Omega_c}{N_A} \sum_{\vec{g} \neq \vec{0}} V_c(\vec{g}) \rho^{PS}(\vec{g})$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(II--75)}$$

$$-\frac{1}{4} \frac{\Omega_c}{N_A} \sum_{\vec{g}} V_{XC}(\vec{g}) \rho^{PS}(\vec{g}) + \alpha_1 Z_v + \gamma_E$$

where γ_E is the Ewald construction and the constant α_1 is calculated as a functional of the pseudopotential. It is to be understood that the $\vec{g} = 0$ component of the electron--core repulsion is excluded from the calculations in II--69, and the compensating terms $\alpha_1 Z_v + \gamma_E$ are added after selfconsistency is attained for ρ, ε.

Table II — Pseudopotential results for structural properties of solids.

Material		Lattice Constant Å	Bulk Modulus	
Si[a]	Yin & Cohen	5.451	0.98	
				(Mbar)
	Experiment	5.429	0.99 (77K)	
		0.4%	1%	
GaAs[b]	Froyen & Cohen	5.57	72.5	
				(GPa)
	Experiment	5.653	74.8	
		−1.5%	−3.1%	

a) Ref.31   b)Ref.32

Several impresssive results were obtained using this formalism, coupled to one or other first–principles pseudopotential form and local exchange functional. For instance, Yin and Cohen[31] study several lattice structures for Si: the diamond structure is seen to offer the minimum energy over most of the usual pressure range, and the "hexagonal diamond" structure, (equivalent of Wurtzite structure for heteropolar compounds) is never the ground state. The open structures are much higher in energy. In a similar study by Froyen and Cohen[32] for GaAs, the zinc–blende structure was also found to be stable.

In Table II we compile the values found in both references for bulk moduli and lattice constants, compared to experimental values (corrected for zero temperature when needed).

The agreement is really impressive. It coorborates the statement that ground state electronic properties are obtainable, to high accuracy, from the DFT or even the LDA.

III. POINT DEFECTS IN SEMICONDUCTORS

III.1. Shallow–Level Impurities

The possibility of introducing selected defects or impurities in semiconductors is what gives the materials their interesting technological properties. This is so because very small concentrations of certain so–called shallow–level impurities, while introducing virtually no alteration in the band structure of the compound, may drastically increase the conductivity (e.g. from $\sim 4\times10^{-6}$ to $10\Omega^{-1}\,cm^{-1}$ for P in Si). On the other hand, even smaller concentrations of deep–level defects can just as drastically reduce the conductivity. The vast amount of experimental work dedicated to this field is thus justified. For the theoretical physicist the study of defects in semiconductors is very attractive since due to the complexity of the problem progress must rely on models and approximations. It offers as such an excellent opportunity to gain insight on Nature[33]. Here I will focus on a few aspects of the problem only, and very superficially.

A first approach to the problem of impurities may be to consider the scheme of fig.3, and imagine what would happen if we replaced a single atom A out of the compound AB by an impurity atom C. Suppose C carries just one more valence electron than A, $N_v' = N_v + 1$, as would be the case for P in Si ($N_v = 4$, $N_v' = 5$) or Si in the Ga site in GaAs ($N_v = 3, N_v' = 4$). All the rest being the same, it would seem that the spectrum of the compound (fig.1b) would not be altered and the extra electron would be accomodated in the first conduction state. We recall then that the core of the impurity is not the same as that of the replaced host atom, having different effective charge, and we expect the attractive potential to be a little stronger at the impurity site than at any other A site. Translational symmetry is broken, the "Bloch picture" is no longer valid. However, since the atoms are "similar", it is resonable to asssume that the defect potential $\Delta V(\vec{r})$ is weak, i.e. the difference between the external potential felt by a valence (conduction) electron in the system with and without the defect

$$\Delta V(\vec{r}) = V(\vec{r}) - V^o(\vec{r}) \qquad \text{(III--1)}$$

is long--ranged but never too large. In this case, a small binding would lower the impurity level in the gap, very close to the band extremum and associated to a state very similar to the band state it "derives" from. The extra electron would be easily ionizable to the conduction band, hence the denomination "shallow" donor impurities. The same argument would apply to an acceptor impurity with $N_v^! = N_v - 1$ such as Al in Si, only then a "hole" would be "emitted" to the valence band with excitation of a valence--band electron into the impurity gap state.

If, on the other hand, we simply removed an atom from the lattice, thus creating a vacancy, or replaced a lattice atom by an impurity with very different electronic properties, we would impose a severe disturbance on the immediate environment of the defect. There will be broken bonds, weakened or strenghtened bonds, new bonds or alien states: it is to be expected that the defect potential will be much stronger. It usually results for such defects that levels are introduced much closer to midgap: deep levels. These states can act as traps, capturing carriers, or recombination centers "killing" two carriers, an electron and a hole.

Shallow--level impurities and deep--level defects behave indded very differently and have to be studied through different models.

For shallow--level impurities it is convenient to separate the complete hamiltonian with the help of (III--1) as

$$\mathcal{H} = \mathcal{H}^o + \Delta V(\vec{r}) \qquad \text{(III--2)}$$

where $\mathcal{H}^o$ describes the perfect lattice, and to express the impurity states $u_i(\vec{r})$ in terms of the eigenvalues of $\mathcal{H}^o$

$$u_i(\vec{r}) = \sum_{n,\vec{k}} C^i_{n\vec{k}} u_{n\vec{k}} \qquad \text{(III--3)}$$

which leads to

$$(\varepsilon_{n\vec{k}} - \varepsilon_i) C^i_{n\vec{k}} + \sum_{n'\vec{k}'} C^i_{n'k'} <n\vec{k}|\Delta V|n'\vec{k}'> = 0 \qquad \text{(III--4)}$$

where $|n\vec{k}> = u_{n\vec{k}} = \tilde{u}_{n\vec{k}}|\vec{k}>$, $\tilde{u}_{n\vec{k}}$ having full lattice periodicity.

The defect potential is then approxiamted to

$$\Delta V(r) \simeq -\frac{e^2}{\epsilon_s r} \qquad \text{(III--5)}$$

with the impurity at $\vec{r} = 0$, and ϵ_s the static dieletric constant of the host. To justify (III--5) it is assumed that the spatial extent of $u_i(\vec{r})$ is so large that we may neglect the details of $\Delta V(\vec{r})$ over an unit cell, and that $u_i(\vec{r})$ only shows sufficient amplitude far way from the impurity (in the central cell details of ΔV cannot be neglected).

It can be seen from (III--4) that as ΔV tended to zero, the eingenenergies and eigenstates $\varepsilon_i, u_i(\vec{r})$ would tend to band structure results and the extra carrier (e.g. electron in GaAs:Si$_{Ga}$, hole in Si:Al) would occupy the state at the band extremum. A weak but finite ΔV will present a binding strength and a gap state will appear. We expect nonetheless the strongest contributions in (III--3) to come from the states at the extremum $\vec{k} = \vec{e}$. Specializing for the case of a non--degenerate extremum at $\vec{e} = \vec{0}$ we may neglect terms with $n' \neq n_e$ and concentrate on $\vec{k} \sim \vec{0}$ to replace $\varepsilon_{n\vec{k}}$ in (III--4) by the effective mass (EM) functional[25,34] form so that we obtain (with the band effective mass m^*)

$$\left[\frac{\hbar^2 k^2}{2m^*} - \varepsilon_i\right] C^i_{\vec{k}} + \sum_{\vec{k}} C^i_{\vec{k}'} < \vec{k} \left| -\frac{e^2}{\epsilon_s r} \right| \vec{k}' > = 0 \ . \qquad \text{(III–6)}$$

The Fourier transform of (III–6) through $\mathcal{F}^i(r) = \sum_{\vec{k}} C^i_{\vec{k}} |\vec{k}>$ is

$$\left[\frac{\hbar^2 k^2}{2m^*} - \frac{e^2}{\epsilon_s r}\right] \mathcal{F}^i(r) = \varepsilon_i \mathcal{F}^i(r) \qquad \text{(III–7)}$$

which has the hydrogenic solutions

$$\varepsilon_j = \frac{e^4}{2 \ \epsilon_s^2 \ h^2 \ m^*} \frac{1}{j^2} \ , \quad j = 1,2,... \qquad \text{(III–8)}$$

We may then write the impurity wavefunction (III–3) to a first approximation as a fully periodic function modulated by a hydrogenic solution

$$u_i(\vec{r}) \simeq \tilde{u}_{n\vec{0}}(\vec{r}) \, \mathcal{F}^i(\vec{r}) \qquad \text{(III–9)}$$

and the states are accordingly labelled 1s,2s,2p ...

The EM approcah works very well for shallow–level impurities, if we make further central–ccell corrections[35] to account fot the chemical identity of the impurity (i.e., P or As in Si do not present identical impurity states). For stronger, localized potentials $\Delta V(\vec{r})$ it fails completely, and other methods have to be used.

III.2. Deep–Level Defects

A defect which causes a strong perturbation in the crystal will also impose a similar perturbation on the electronic structure, so there will appear deep midgap levels, or strongly localized states within the conduction or valence bands (resonances). Since the impurity states are quite localized in $\vec{r}$, expansion (III–3) is not very useful: we would need $\vec{k}$–vectors from all of the BZ.

The study of such systems is thus more complicated than that of perfect solids, since we do not have translational symmetry, we have to worry about localization effects (e.g., many–electron interactions), and so forth. On the other hand it is more complicated than the study of molecules, because we must relate results to the band structure of the host.

A defect system is characterized if we know most of the following properties: 1) Chemical identity (e.g. a vacancy, or a 3d metal impurity) 2) Structure in the lattice (is the impurity substitutional, or interstitial; does it distort the lattice nearby?), formation energies and reaction kinetics (does such an impurity sit substitutionally or does it "prefer" to form a pair interstitial+vacancy?) 3) Carrier capture and emission properties (is the defect a donor or acceptor? how many levels are there in the gap? how long is the lifetime of an ionized state?) 4) Migration paths, diffusion dynamics (is the impurity a fast diffuser, is it affected by the presence of other defects, or of free carriers?).

The theoretical study usually begins by the choice of a microscopic model (1) for which we obtain the electronic structure (3) in one or more lattice enviroments (2). This procedure is carried out within mean–field theory and then, if it is necessary, many–electron effects are estimated or calculated.

The main assumption for all mean–field approaches is that we can define a region of the crystal surrounding the defect – a "cluster" of atoms – which is relevant. Outside this region the defect potential is supposed to vanish:

$$\Delta V(\vec{r}) \simeq 0 \ , \ |r| > R_c \qquad \text{(III–10)}$$

[for ionized systems the Coulomb tail has to be included, but separately]. It is also understood that the calculations have to be carried through to self—consistency, since charge relaxation in the cluster region is very strong.

Out of this common basis we can distinguish three approaches:

a) The **large unit cell** or **supercell** approach: Once the cluster is chosen it is repeated in space to form a crystal with the defects periodically arranged. This procedure may entail a full band structure calculation[36] for the "defect crystal", or just a molecular calculation for a "cyclic cluster", or we may use expansion (III—3) if we choose carefully the supercell symmetry[37]. The advantages of this approach are that there are no "surfaces" between the cluster and the rest of the crystal, which could introduce spurious states. However, the periodic array of defects may induce spurious defect interactions[37a].

b) The **embedded cluster** or **Green's function** method: In this case[38,39] the formalism is written in terms of the Green's function

$$[\varepsilon - \mathcal{H}] \, G\,(\vec{r},\vec{r}\,',\varepsilon) \;=\; \delta(\vec{r} - \vec{r}\,') \qquad\qquad \text{(III—11)}$$

which can be expressed for the pure crystal as

$$G^o(\varepsilon) \;=\; \sum_{n\vec{k}} \frac{|n\vec{k}> \,<n\vec{k}|}{\varepsilon - \varepsilon_{n\vec{k}}} \quad . \qquad\qquad \text{(III—12)}$$

For the crystal with the defect, adopting (III—3) for the impurity wave functions we may write[38]

$$G - G^o \;=\; G^o \cdot \Delta V \cdot [\Delta V - \Delta V \cdot G^o \cdot \Delta V]^{-1} \cdot \Delta V \cdot G^o \qquad \text{(III—13)}$$

which vanishes outside R_c. Inside R_c the impurity states are expanded in an atomic basis set $\{\phi_\mu\}$ as in (II—17), and we express the matrix G^o as

$$G^o_{\mu\lambda} \;=\; \sum_{n\vec{k}} \frac{<\mu|n\vec{k}> \,<n\vec{k}|\lambda>}{\varepsilon - \varepsilon_{n\vec{k}}} \quad . \qquad\qquad \text{(III—14)}$$

The poles of G yield the energies of the bound states and the effects on the charge density are found through

$$\Delta\rho(\vec{r}) \;=\; -\frac{1}{\pi} \, \text{Im} \left[\sum_{\mu\lambda} \int_{-\infty}^{occ} <\mu|G(\varepsilon + i\eta) - G^o(\varepsilon + i\eta)|\lambda> \right] \quad . \qquad \text{(III—15)}$$

We are then in principle working with the complete system. There might be serious problems for highly localized defect states[37b], since (III—3) has to hold over all space. Lattice distortions are also difficult to include.

c) The **isolated** or **molecular cluster** model: The relevant cluster is studied as a complete unit, disconnected from the crystal, through real—space techniques based on (II—17). The crystalline bonding is simulated either by saturating the broken bonds at the surface of the cluster by terminator atoms or by other suitable boundary conditions[40] depending on the technique used for solving (II—20) or (II—35).

Here again we treat isolated defects, and we may even treat ionized centers. Different lattice configurations can be easily investigated[10,41] and the model has enough flexibility to treat complex defects[42]. The finiteness of the cluster may however induce an artificial localization of defect levels; also there is no longer a clear definition of band edges.

Once a mean—field calculation for some configuration of a given center is completed, we obtain the charge density and eigenvalues for a particular charge state (usually neutral relative to the lattice). If the center is magnetically active the **wave—function amplitudes** may already be compared to **magnetic resonance data** (if many—body effects are not important, and if we are not working with pseudo—charge—densities).

The **depth of levels in the gap** corresponds to the energy needed for the defect to capture or emit a carrier, that is, to **change charge state**, or be ionized. Within the **same charge state** sometimes there may be **excited electronic states** of the defect. We have to remember that within LDA excitation energies or ionization energies to one of the bands have always to be obtained through the transition state construct. The mean–field energies still cannot be compared to experiment, unless many–body effects are of negligible order. After inclusion of many–body corrections[43] we may compare results for the energy levels with non–equilibrium experimental data, e.g. photoconductivity, or other optical absorption techniques.

If we want to study equilibrium phenomena we still have to include[44] interaction with the Fermi reservoir of available carriers: in a real semiconductor there will always be a fair amount of different defects, and the stability of a given charge state for a particular center will be dictated by the relative proportions of these defects. This is shown schematically in Fig.4: in part a are displayed energy differences which give non–equilibrium acceptor (0/−) and donor (+/0) energies relative to the valence band. In part b are total energies for the same centers, but now coupled to the Fermi reservoir. As the Fermi energy μ_F moves to higher values the electron–poor charge states are destabilized and electron–rich situations are favored. The width in energy for which a given charge state is stable is the Mott–Hubbard energy U for that charge state. It may happen, although it seldom occurs, that an intermediate state is not stable (it is only seen in non–equilibrium experiments). These are negative–U systems[44] and the situation is as in parts II or III of Fig.4.
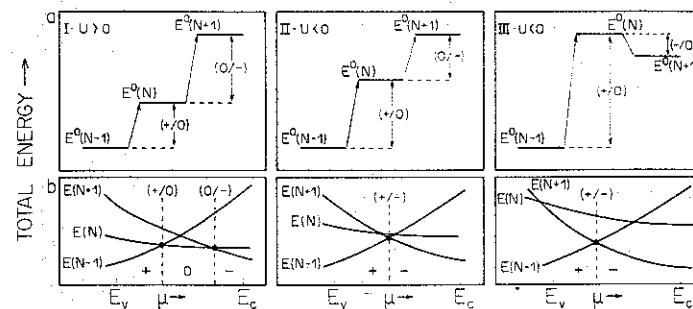
Figure 4— Schematic representation of impurity energy levels in the gap of a semiconductor a) Isolated system [defect + pure crystal] b) System coupled to Fermi reservoir. $E_{v,c}$ mark valence and conduction band edges; (0/−) denotes acceptor energies, (+/0) donor energies; μ_F is the Fermi energy.

Of much interest in recent years are problems of stability of defects[10,36,41,45] which are usually related to lattice distortions. Theoretical models are then constantly being suggested[46], improved and modified, so that, as in experimental physics, we have specific models for different defect centers. We must, though, always take into account the approximations involved in a model when comparing results with experiment. The ideal situation is when several different approaches are used for the same center. We may obtain similar results[47], or results may look controversial for a time: in either case we get closer to understanding the behavior of these complicated systems.

REFERENCES

1. Hedin, L. and Lundqvist, S., *Solid St. Phys.* **23**, 2 (1969).

2. Slater, J.C., *"Quantum Theory of Molecules and Solids"* vol. IV (McGraw–Hill, N.Y., 1974).

3. Lowdin, P.O., *J. Appl. Phys.* **33**, suppl., 251 (1962).

4. Bloch, F., *Z. Physik* **52**, 555 (1929).

5. Roothan, C.C.J., *Rev. Mod. Phys.* **23**, 69 (1951).

6. Klopman, G. and Evans, R.C., in *"Semiempirical Methods of Electronic Structure Calculation. Part A: Techniques"* (G.A. Segal ed., Plenum Press, N.Y., 1977) ch.2.

7. a) Pople, J.A. and Segal, G.A., *J. Chem. Phys.* **44**, 3289 (1966).
   b) Ridley, J. and Zerner, M., *Theor. Chim. Acta* **32**, 111 (1973).
   c) Dewar, M.J.S. and Thiel, W., *J. Am. Chem. Soc.* **99**, 4899 (1977).

8. De Leo, G.G., Fowler, B. and Watkins, G., *Phys. Rev.* **B29**, 3193 (1984).

9. Harker, A.H. and Larkins, F.P., *J. Phys.* **C12**, 2497 (1979).

10. Caldas, M.J. and Fazzio, A., *Mat. Sci. For.* **38–41**, 119 (1989).

11. Slater, J.C. and Koster, G.F., *Phys. Rev.* **94**, 1498 (1954).

12. Slater, J.C., *Phys. Rev.* **81**, 385 (1951); id. ibid. **82**, 538 (1951).

13. Hohenberg, P. and Kohn, W., *Phys. Rev.* **136**, B864 (1964).

14. Kohn, W. and Sham, L., *Phys. Rev.* **140**, A1133 (1965).

15. Schwarz, K., *Phys. Rev.* **B5**, 2466 (1972).

16. Chadi, D.J. and Cohen, M.L., *Phys. Rev.* **B8**, 5747 (1973).

17. Ciraci, S. and Batra, I.P., *Phys. Rev.* **B15**, 4923 (1977).

18. Stückel, D.J. and Euwema, R.N., *Phys. Rev.* **B1**, 1635 (1970).

19. Zunger, A. and Cohen, M.L., *Phys. Rev.* **B18**, 5449 (1978); id. ibid. **20**, 4082 (1979).

20. Ley, L., Kowalcysk, S., Pollack, R. and Shirley, D.A., *Phys. Rev. Lett.* **29**, 1088 (1972).

21. Zucca, R., Walter, J.P., Shen. Y.R. and Cohen, M.L., *Solid St. Commun.* **8**, 627 (1970).

22. Ceperley, D.M. and Alder, B.J., *Phys. Rev. Lett.* **45**, 566 (1980).

23. Heine, V., *Solid St. Phys.* **24**, 1 (1970).

24. Herman, F., Kortum, R.L. and Kuglin, C.D., *Int. J. Quantum Chem.* **15**, 533 (1967).

25. Harrison, G., *"Solid State Theory"* (Dover, N.Y., 1980).

26. Chelikowsky, J. and Cohen, M.L., *Phys. Rev.* **B14**, 556 (1976).

27. Hamann, D.R., Schlüter, M. and Chiang, C., *Phys. Rev. Lett.* **43**, 1494 (1979).

28. Hyberstein, M.S. and Louie, S.G., *Phys. Rev.* **B37**, 2733 (1988).

29. Godby, R.W., Schlüter, M. and Sham, L.J., *Phys. Rev.* **B37**, 10159 (1988).

30. Ihm, J., Zunger, A. and Cohen, M.L., *J. Phys.* **C12**, 4409 (1979).

31. Yin, M.T. and Cohen, M.L., *Phys. Rev. Lett.* **45**, 1004 (1980).

32. Froyen, S. and Cohen, M.L.., *Physica* **117&118B**, 561 (1983).

33. Eddington, A.S., *"The Internal Constitution of Stars"* (University Press, Cambridge, 1926).

34. Kohn, W., *Solid St. Phys.* **5**, 258 (1957).

35. Pantelides, S.T., *Rev. Mod. Phys.* **50**, 797 (1978).

36. a) Louie, S.G., Schlüter, M., Chelikowsky, R.J., and Cohen, M.L., *Phys. Rev.* **B13**, 1654 (1976).
    b) Biernacki, S., Scherz, U., Gillert, R. and Scheffler, M., *Mat. Sci. For.* **38–41**, 625 (1989).

37. a) Lindefelt, U., *J. Phys.* **C11**, 3651 (1978).
    b) Lindefelt, U. and Zunger, A., *Phys. Rev.* **B26**, 846 (1982).

38. Baraff, G.A. and Schlüter, M., *Phys. Rev.* **B19**, 4965 (1979).

39. Bernholc, J., Lipari, N.O. and Pantelides, S.T., *Phys. Rev.* **B21**, 3545 (1980).

40. Fazzio, A., Leite, J.R. and De Siqueira, M.L., *J. Phys.* **C12**, 3469 (1979).

41. Oshyiama, A., and Ohnishi, S., *Phys. Rev.* **B33**, 4320 (1986).

42. Assali, L.V.C. and Leite, J.R., *Mat. Sci. For.* **38–41**, 409 (1989).

43. Fazzio, A., Caldas, M.J. and Zunger, A., *Phys. Rev.* **B30**, 3430 (1984).

44. Baraff, G.A., Kane, E.O. and Schlüter, M., *Phys. Rev.* **B21**, 5662 (1980).

45. Chadi, D.J. and Chang, K.J., *Phys. Rev. Lett.* **61**, 873 (1988).

46. Car, R. and Parrinello, M., *Phys. Rev. Lett.* **55**. 2471 (1989).

47. a) Katayama–Yoshida, H., and Zunger, A., *Phys. Rev.* **B33**, 2961 (1986).

    b) Caldas, M.J., Figueiredo, S.K. and Fazzio, A., *Phys. Rev.* **B33**, 7102 (1986).