The Ab Initio Cluster Method and the Dynamics of Defects in Semiconductors

R. Jones,

DEPARTMENT OF PHYSICS, UNIVERSITY OF EXETER,
EXETER, EX4 4QL, UK.

P. R. Briddon,

DEPARTMENT OF PHYSICS, UNIVERSITY OF NEWCASTLE UPON TYNE,
NEWCASTLE UPON TYNE, NE1 7RU, UK.

(To appear in *Identification of Defects in Semiconductors*, ed. M. Stavola, *Semiconductors and Semimetals*, treatise editors, R. K. Willardson, A. C. Beer, and E. R. Weber, Academic Press.)

I. Introduction

First-principles methods of determining the structure and electronic properties of materials have become very popular in the fifteen years or so since the pioneering studies of Yin and Cohen (1980). Usually, these calculations are carried out on a supercell employing a basis of plane waves. For many applications such an approach is not the most efficient. For example, molecules, topological defects like dislocations, kinks or adsorbates on surfaces are cases where a cluster approach has definite advantages. Even for point defects in a crystalline environment, there are advantages arising from a cluster method using a localized basis set of orbitals. Such methods can give a direct interpretation of defect wavefunctions in terms of hybridized orbitals on atoms local to the defect and can treat the strict symmetry of a point defect which is lost through interaction between defects in different unit cells. Unlike the supercell approach, the cluster method can easily treat the induced dipole moments of dynamic defects which is important for determining the integrated absorption intensity of infra-red radiation.

In this review, we shall discuss the basis of the two main *ab initio* methods: Hartree-Fock and density functional theories. We then discuss in detail the application to atomic clusters using a localized basis set. We shall almost exclusively deal with the implementation developed by us and finally we discuss some applications that have been made of the formalism to defects in diamond, silicon, and other semiconductors, although the theory has been used to treat many other systems. However, it seems relevant to begin by discussing our motivation for introducing the technique in the first place.

In 1985, we became interested in the problem of water reacting with the cores of dislocations in quartz. It was believed that water molecules could react with the strained Si-O bonds within the core, or at kinks, breaking them and creating two Si-O-H bonds (Griggs and Blacic, 1965). This mechanism might then explain the very dramatic effect of hydrolytic weakening where the yield stress of dry quartz is more than an order of magnitude higher than wet quartz (Doukhan and Trepied, 1985; Heggie and Jones, 1987). To investigate such a process requires a theoretical method that is able to account not only for the strength of chemical bonds, but is able to deal with their unusual environment within dislocation cores. The most satisfactory technique would be one that did not rely on empirical information whose applicability would be uncertain in this case. There are two *ab initio* schemes that do not rely on empirical methods: the traditional Hartree-Fock method and density functional methods. Hartree-Fock methods have been championed by chemists but are unable to account for the quasi-particle spectrum of metals, and perhaps for this reason density functional methods have been favored by physicists. Moreover, an important consideration is that the latter have a well-developed pseudopotential scheme which makes applications to materials containing for example germanium no more difficult than those composed of carbon. In addition, the evaluation of the exchange-correlation energy is simpler but at a cost of the lack of systematic improvements which progressively reduce the errors in the energies of bonds or multiplets. Another problem area for density functional theory is in the description of Mott insulators such as NiO. Spin density functional theory predicts these materials to be either metallic or narrow gap semiconductors, whereas they are observed to be highly insulating. It is usually the case that density functional theory finds band gaps smaller than the experimental values. Hartree-Fock methods on the other hand usually predict band gaps much larger than experimental values. Both methods could be used to treat clusters or supercells but the problem of dislocations and of kinks is so demanding and requires so many atoms, that it seemed desirable to use a cluster containing a single dislocation – not possible in a supercell — whose surface was terminated by hydrogen. So it seemed sensible at that time to invest a considerable amount of effort in developing an *ab initio* local density functional cluster method which incorporated pseudopotentials.

The particular code developed and used by the Exeter, Newcastle, Sussex and Luleå groups is called AIMPRO which is an acronym for *Ab Initio Modeling Program.* The code has undergone a great many modifications and improvements since it was first written. These developments have extended the range of applications and most importantly have led to a considerable speed up so that nowadays very large clusters of atoms can be considered. At the time of writing, the largest cluster considered is an 840 atom bucky-onion which was run, without using any symmetry acceleration options, on a T3D using 256 processors (Heggie *et al.,* 1996a). Typically, about two hours were required to carry out one conjugate gradient iteration which generates the relaxed structure. This extreme application illustrates the power of the method but of course most applications to solid state problems use much smaller 70-150 atomic clusters. Such clusters can be run on simple RISC workstations taking several days.

We begin by giving an overview of the problem of determining the equilibrium structure of a multi-atom system, then we shall discuss the cluster method in some detail before describing some of the applications that have been made.

II. The many-body problem

It is desirable to choose a system of units where the fundamental constants are removed from the equations. We shall use atomic units throughout except in dealing with applications. In terms of these units, $\hbar, e, m$, and $4\pi\epsilon_0$ are taken to be unity. The Schrödinger equation for the electron in the hydrogen atom for example, then becomes:

$$\{-\frac{1}{2}\nabla^2 - \frac{1}{r} - E\}\psi(\mathbf{r}) = 0.$$

The 1s solution is then, $\psi = \frac{1}{\sqrt{\pi}}e^{-r}$, and has energy, $E = -\frac{1}{2}$. This establishes the unit of energy to be 1 a.u. = 27.212 eV, and as the radius of the atom is 1 a.u., the unit of length is the Bohr radius of 0.529 Å.

The non-relativistic many-body Schrödinger equation for the electrons in a fixed field due to ions of charges Z_a at sites $\mathbf{R}_a$ is:

$$\{-\frac{1}{2}\sum_{\mu}\nabla_{\mu}^2 + \frac{1}{2}\sum_{\nu\neq\mu}\frac{1}{|\mathbf{r}_\mu - \mathbf{r}_\nu|} - \sum_{\mu,a}\frac{Z_a}{|\mathbf{r}_\mu - \mathbf{R}_a|}$$
$$+\frac{1}{2}\sum_{a\neq b}\frac{Z_a Z_b}{|\mathbf{R}_a - \mathbf{R}_b|} - E\}\Psi(r) = 0,$$

or, in an obvious notation,

$$(H - E)\Psi(r) = \{T + V_{e-e} + V_{e-i} + V_{i-i} - E\}\Psi(r) = 0. \tag{1}$$

Here r denotes the positions and spins of the electrons, ie, $(\mathbf{r}_1, s_1, \mathbf{r}_2, s_2, ...)$. We shall be mainly concerned with the ground state solution to this equation, and the greater part of this article is devoted to a discussion of the techniques we employ to obtain this.

It should be noted that the Hamiltonian of Eq. (1) does not include the kinetic energy of the ions themselves, if they are also regarded as a quantum–mechanical system. The full Hamiltonian, H_T, includes this kinetic energy :

$$H_T = H - \sum_a \frac{m}{2M_a}\nabla_a^2.$$

We have written the term m/M_a to remind the reader this ratio involves the electron and atomic masses. The structure and properties of all atomic clusters are contained in the solution of the Schrödinger equation of the full Hamiltonian. However, to develop a practical method of solving this equation it is necessary to first decouple the motions of the electrons and ions; then to construct an effective potential acting on each electron due to the other electrons as well as the surrounding nuclei; and finally to calculate the forces acting on each ion so that both the equilibrium structure of the cluster as well as its vibrational modes can be found.

1. THE BORN-OPPENHEIMER APPROXIMATION

The first step in which the dynamical equations of the ions are separated from those of the electrons is made using this approximation. We assume that the ions are so much more massive than the electrons that their movement simply modulates the wavefunction of the electrons. The total wavefunction can then be written

$$\Psi_T(r,R) = \chi(R)\Psi(r,R),$$

where $\chi(R)$ is an amplitude dependent on the nuclear coordinates alone, and $\Psi(r,R)$ is a solution of Eq. (1). We have denoted the nuclear coordinates by $R = (\mathbf{R}_1, \mathbf{R}_2, ...)$. If this is substituted into the Schrödinger equation for the full Hamiltonian, then we find, after multiplying through by $\Psi^*(r,R)$,

$$\{-\sum_a \frac{m}{2M_a}\nabla_a^2 + E(R) + W(R) - E_T\}\chi(R) = \sum_a \int \Psi^*(r,R)\frac{m}{M_a}\nabla_a\Psi(r,R)\nabla_a\chi(R)dr. \qquad (2)$$

The term dr represents the integration over all the electron coordinates, $\mathbf{r}_\mu$ and the summation over all their spins s_μ. The left hand side represents the Schrödinger equation for the ions moving in a potential $E+W$, where W is a small correction, invariably neglected, due to the electrons moving along with the nuclei:

$$W(R) = -\sum_a \frac{m}{2M_a}\int \Psi^*(r,R)\nabla_a^2\Psi(r,R)dr.$$

The term on the right hand side of (2) vanishes if $\Psi(r,R)$ is real corresponding to a non-degenerate ground state. Otherwise, it usually represents a small perturbation but can be particularly important for degenerate ground states — as perturbations usually are — for then it can lead to symmetry breaking as in the Jahn-Teller effect (Stoneham, 1975). If we neglect this term, then the ionic and electron motions are decoupled.

$E(R)$ is an effective potential energy of the ions averaged over the state $\Psi(r,R)$. The minimum value of $E(R)$ is then the ground state energy of the cluster and one of the principal objectives of the theory is to deduce this energy and the corresponding ionic positions. If $E(R)$ is expanded about its minimum value, we find:

$$E(R) = E(R_o) + \frac{1}{2}\sum_{la,mb}\Big(\frac{\partial^2 E}{\partial R_{la}\partial R_{mb}}\Big)\Delta R_{la}\Delta R_{mb} + .. \qquad (3)$$

Here ΔR_{la} represents the displacement of the ion a in direction l from the equilibrium configuration R_o. The harmonic frequencies of vibration ω_i and their normal coordinates, u^i_{la}, are related to the eigenvalues and eigenvectors of the dynamical matrix calculated from the energy derivatives (Born and Huang, 1954) *i.e.*,

$$\sum_{mb} E_{la,mb} u^i_{mb} = \omega_i^2 u^i_{la},$$

$$E_{la,mb} = \frac{1}{\sqrt{M_a M_b}} \Big(\frac{\partial^2 E}{\partial R_{la} \partial R_{mb}} \Big).$$

We shall discuss this further in section VII. Now that we have separated the motions of the ions and electrons we are confronted by the problem of the interaction between the electrons implicit in Eq. (1). To deal with this, further approximations are required.

2. HARTREE-FOCK THEORY

The assumption behind this method is that there exists a set of M orthonormal one-electron spin-orbitals $\psi_\lambda(r)$ from which the many-body wavefunction can be constructed as a single Slater determinant:

$$\Psi(r) = \frac{1}{\sqrt{M!}} det|\psi_\lambda(r_\mu)|, \quad \psi_\lambda(r) = \psi_i(\mathbf{r})\chi_\alpha(s),$$

and $\chi_\alpha(s)$ is a spin-function satisfying:

$$\sum_s \chi^*_\alpha(s)\chi_\beta(s) = \delta_{\alpha\beta}.$$

The sum being over 2-values of s and α being 'up' or 'down'. The orbitals $\psi_i(\mathbf{r})$ satisfy:

$$\int \psi^*_i(\mathbf{r})\psi_j(\mathbf{r}) d\mathbf{r} = \delta_{ij}.$$

The many-body wavefunction is clearly antisymmetric with respect to the interchange of two particles as is required by the Pauli exclusion principle. As a simple example, we may write down the wavefunction for a two electron problem such as H_2 by expanding the determinant. We obtain the well-known result for a two-particle fermion system:

$$\Psi(r_1, r_2) = \frac{1}{\sqrt{2}} \{\psi_1(r_1)\psi_2(r_2) - \psi_1(r_2)\psi_2(r_1)\}.$$

The average energy of the single normalized determinental wavefunction is $\langle\Psi|H|\Psi\rangle$ and can be shown to be (Slater, 1960) :

$$E = \sum_\lambda \langle\lambda|T + V_{e-i} + V_{i-i}|\lambda\rangle + \frac{1}{2} \sum_{\lambda,\mu} \{\langle\lambda\mu|V_{e-e}|\lambda\mu\rangle - \langle\lambda\mu|V_{e-e}|\mu\lambda\rangle\}. \quad (4)$$

Here, the first term involves the matrix elements of one-particle operators: the kinetic, electron-ion and ion-ion interactions respectively, and the sum is over the occupied spin-orbitals λ. The second and third terms involve four-center integrals of the electron-electron interaction:

$$\langle\lambda\mu|V_{e-e}|\nu\kappa\rangle = \sum_{s_1 s_2} \int \psi^*_\lambda(r_1)\psi^*_\mu(r_2) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \psi_\nu(r_1)\psi_\kappa(r_2) d\mathbf{r}_1 d\mathbf{r}_2.$$

We can rewrite E in the form :

$$E = -\frac{1}{2}\sum_{\lambda s}\int \psi^*_\lambda(\mathbf{r},s)\nabla^2\psi_\lambda(\mathbf{r},s)d\mathbf{r} + \int n(\mathbf{r})V_{e-i}d\mathbf{r} + E_H + E_x + E_{i-i}, \tag{5}$$

$$E_H = \frac{1}{2}\int \frac{n(\mathbf{r}_1)n(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}d\mathbf{r}_1 d\mathbf{r}_2, \tag{6}$$

$$E_x = -\frac{1}{2}\sum_{\lambda\mu}\sum_{s_1 s_2}\int \psi^*_\lambda(r_1)\psi^*_\mu(r_2)\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\psi_\mu(r_1)\psi_\lambda(r_2)d\mathbf{r}_1 d\mathbf{r}_2, \tag{7}$$

$$E_{i-i} = \frac{1}{2}\sum_{a\neq b}\frac{Z_a Z_b}{|\mathbf{R}_a - \mathbf{R}_b|}. \tag{8}$$

Here we have introduced the electron density

$$n(\mathbf{r}) = \sum_{\lambda s}|\psi_\lambda(\mathbf{r},s)|^2, \tag{9}$$

and the Hartree energy E_H, the exchange energies E_x, and the ion-ion energy E_{i-i}.

The ground state orbitals ψ_λ are determined by the requirement that E is minimized subject to orthonormal ψ_λ. This constrained minimization problem can be solved by introducing Lagrange multipliers, $E_{\lambda\mu}$, such that the function

$$E - \sum_{s,\lambda\neq\mu}E_{\lambda\mu}\int \psi^*_\lambda\psi_\mu d\mathbf{r} - \sum_\lambda E_\lambda\{\sum_s\int|\psi_\lambda|^2 d\mathbf{r} - 1\}$$

is minimized with respect to $\psi^*_\lambda, E_\lambda$ and $E_{\lambda\mu}$ without constraint. From Eq. (5), we then get the Hartree-Fock equations for each orbital λ:

$$\{-\frac{1}{2}\nabla^2 + V_{e-i}(\mathbf{r}) + V^H(\mathbf{r}) + V^x_\lambda(\mathbf{r}) - E_\lambda\}\psi_\lambda(r) = \sum_{\mu\neq\lambda}E_{\lambda\mu}\psi_\mu(r), \tag{10}$$

$$\sum_s\int\psi^*_\mu\psi_\lambda d\mathbf{r} = \delta_{\lambda\mu},$$

$$V^H(\mathbf{r})\psi_\lambda(r) = \frac{\delta E_H}{\delta\psi^*_\lambda} = \int\frac{n(\mathbf{r}_1)\psi_\lambda(r)}{|\mathbf{r} - \mathbf{r}_1|}d\mathbf{r}_1,$$

$$V^x_\lambda(\mathbf{r})\psi_\lambda(r) = \frac{\delta E_x}{\delta\psi^*_\lambda} = -\sum_{\mu s_1}\int\psi^*_\mu(r_1)\psi_\lambda(r_1)\frac{1}{|\mathbf{r}_1 - \mathbf{r}|}\psi_\mu(r)d\mathbf{r}_1.$$

V^H, V^x_λ are the Hartree and exchange potentials. The last involves a sum over occupied orbitals μ whose spin is the same as that of λ.

Now we can carry out a unitary transformation on the Slater determinant which diagonalizes $E_{\lambda\mu}$ and then the right hand side of the differential Eq. (10) vanishes. The exchange potential can be written in terms of the exchange density, n^x_λ, as:

$$V^x_\lambda(\mathbf{r}) = \int\frac{n^x_\lambda(\mathbf{r},\mathbf{r}_1)}{|\mathbf{r}_1 - \mathbf{r}|}d\mathbf{r}_1$$

$$n^x_\lambda(\mathbf{r},\mathbf{r}_1) = -\frac{\sum_{\mu s_1}\delta(s_\lambda,s_\mu)\psi^*_\mu(r_1)\psi_\mu(r)\psi^*_\lambda(r)\psi_\lambda(r_1)}{\psi_\lambda(r)\psi^*_\lambda(r)}.$$

The exchange density satisfies:

$$\int n^x_\lambda(\mathbf{r},\mathbf{r}_1)d\mathbf{r}_1 = -1,$$
$$n^x_\lambda(\mathbf{r},\mathbf{r}) < 0.$$

These relations show that an exchange hole of total charge unity is introduced around each electron. The exchange integral is a very difficult term to evaluate numerically as it involves the product of orbitals, each of which oscillates in a complicated way. Further, as it depends on λ, it has to be evaluated many times. This makes practical versions of the theory rather slow even on the fastest computers.

The total energy E can be found by multiplying the Hartree-Fock equations (10) by $\psi^*_\lambda(r)$ and integrating over $\mathbf{r}$ and summing over s and λ. This gives:

$$E = \sum_\lambda E_\lambda - E_H - E_x + E_{i-i}.$$

The sum is over occupied orbitals only. Notice that the interaction terms must be subtracted from the sum of energy eigenvalues.

The Hartree-Fock equations in (10) are solved by a self-consistent method. An initial set of orbitals $\psi_\lambda(r)$ are selected, which are usually related to atomic orbitals, and the Hartree and exchange potentials found. Then the Hartree-Fock equations are solved for an output set of orbitals. These are used to construct a new set of input orbitals and the process repeated until the output and input sets are equal. This process is called the self-consistent cycle.

The energy-eigenvalues, E_λ, can be given an interpretation through Koopman's theorem which states that the difference in energy between two configurations differing by the occupation of an orbital λ, while all the other orbitals ψ_μ are unchanged, is E_λ. Hence $-E_\lambda$ as the ionization energy for the λ electron. The correspondence is not exact as all the orbitals will alter, in general, whenever the configuration is modified.

Hartree-Fock theory usually predicts structures and vibratory modes of small molecules quite accurately. However, bond lengths are usually underestimated leading to an overestimate of mode frequencies. Excitation energies are also overestimated.

3. THE HOMOGENEOUS ELECTRON GAS

As an example, we apply the theory to a homogeneous electron gas, sometimes called jellium, where the ions form a uniform background of density n. The total spin S is then a good quantum number and we begin by looking at non-polarized states where $S = 0$. A solution of the Hartree-Fock equations consists of orbitals corresponding to plane-waves. Then λ refers to the wave-vector $\mathbf{k}$ and spin state α:

$$\psi_\lambda(\mathbf{r},s) = \frac{1}{\sqrt{\Omega}}e^{i\mathbf{k}.\mathbf{r}}\chi_\alpha(s).$$

Here Ω is the volume of the system. The charge density, n, is uniform and hence the Hartree term, E_H, and the ion-ion term E_{i-i}, which in this case is just the electrostatic energy of the uniform positive background charge, exactly cancels the electron-ion term, E_{el-i}, in the total energy E.

The energy levels are therefore:

$$E_\lambda = E_{\mathbf{k},\alpha} = \frac{1}{2}k^2 + V^x_{\mathbf{k}},$$

where the exchange potential, $V^x_\lambda(\mathbf{r})$, is given by

$$V^x_{\mathbf{k}}(\mathbf{r}) = -\sum_{k_1<k_f} \int \frac{e^{i(\mathbf{k}-\mathbf{k}_1).(\mathbf{r}_1-\mathbf{r})}}{\Omega|\mathbf{r}_1-\mathbf{r}|} d\mathbf{r}_1.$$

This is in fact independent of $\mathbf{r}$ and spin state α and depends on the magnitude k of $\mathbf{k}$ alone.

$$V^x_{\mathbf{k}} = -\sum_{k_1<k_f} \frac{4\pi}{\Omega|\mathbf{k}-\mathbf{k}_1|^2} = -\frac{1}{8\pi^3}\int_{k_1<k_f} \frac{4\pi}{|\mathbf{k}-\mathbf{k}_1|^2} d\mathbf{k}_1.$$

Here k_f is the Fermi wave-vector related to the electron density by $n = \frac{1}{3\pi^2}k_f^3$. To carry out the integral, we write $\eta = k/k_f$, and a simple calculation shows:

$$V^x_{\mathbf{k}} = -4\Big(\frac{3n}{8\pi}\Big)^{\frac{1}{3}} F(\eta),$$
$$F(\eta) = \frac{1}{2} + \frac{1-\eta^2}{4\eta} ln(\frac{1+\eta}{1-\eta}).$$

The function $F(\eta)$ tends to 1 as $\eta \to 0$, and to 1/2 as $\eta \to 1$. Its derivative has a weak singularity as η tends to 1 which has catastrophic implications for the applicability of the theory to simple metals. This follows as the density of states, per unit energy range and for each spin, is

$$N(E) = \frac{4\pi k^2}{8\pi^3}\frac{1}{|\nabla E_{\mathbf{k}}|},$$

and tends to zero as $k \to k_f$, thus showing that the density of states is zero at the Fermi level. This is incorrect and is due to absence of correlation in the theory. The form of the Hartree–Fock wavefunction does not include correlated movement of the electrons. This can be included by constructing wavefunctions built out of combinations of determinants. These are known as configuration interaction (CI) calculations, but the computational demands are so high and the scaling with the number of electrons so poor that such calculations can only be done for a very small number of atoms.

The total energy is

$$E = \frac{1}{2}\sum_{\mathbf{k}\alpha} k^2 + E_x,$$

which gives the energy density:

$$\frac{3n}{10}\Big(3\pi^2 n\Big)^{\frac{2}{3}} + n\epsilon_x(n), \quad \epsilon_x(n) = -\frac{3}{2}\Big(\frac{3n}{8\pi}\Big)^{\frac{1}{3}}.$$

The quantity ϵ_x is the exchange energy per electron.

4. THE SPIN POLARIZED ELECTRON GAS

We now consider solutions of the Hartree-Fock equations for non-zero spin values S. This means we have more 'up' spins say than 'down' spins and each state is specified by the densities of 'up' and 'down' spins, $n_\uparrow$ and $n_\downarrow$ respectively. The orbitals remain plane waves defined by a wavevector $\mathbf{k}$ but each spin population has its own Fermi wavevector and the total energy is then:

$$E = \Omega \sum_s \{ \frac{3n_s}{10} \left(6\pi^2 n_s\right)^{\frac{2}{3}} - \frac{3}{2}\left(\frac{3}{4\pi}\right)^{\frac{1}{3}} n_s^{\frac{4}{3}} \}. \tag{11}$$

Improved estimates of the ground state energy can be found by going beyond Hartree-Fock theory. Ceperley and Alder (1980) used a quantum Monte-Carlo method to find the correlation energy E_c, — the difference between the ground state and the Hartree-Fock energy — for polarized and non-polarized electron gases for a low density homogeneous electron gas. This can be combined with results of perturbation theory for the high density case to produce an energy for a wide range range of densities. If the correlation energy per electron, ϵ_c, polarization ξ and the Wigner-Seitz radius of each electron r_s are defined by:

$$E_c = \Omega n \epsilon_c(n, \xi), \quad \xi = \frac{(n_\uparrow - n_\downarrow)}{n}, \quad r_s = (4\pi n/3)^{-1/3},$$

then ϵ_c for the non-polarized and fully polarized electron gases are given by Perdew and Zunger (1981) as:

$$\epsilon_c = \begin{cases} \gamma\{1 + \beta_1\sqrt{r_s} + \beta_2 r_s\}^{-1}, & \text{for } r_s \geq 1 \\ B + (A + Cr_s)ln(r_s) + Dr_s, & \text{for } r_s < 1 \end{cases}$$

The values of the coefficients are given for both cases in Table 1.

In the case of a partially polarized gas, where $1 > \xi > 0$, the correlation energy is averaged over the polarized and non-polarized cases using the procedure due to von Barth and Hedin (1972):

$$\epsilon_c(n, \xi) = \epsilon_c^{np}(n) + f(\xi)(\epsilon_c^p - \epsilon_c^{np})$$
$$f(\xi) = \frac{(1+\xi)^{4/3} + (1-\xi)^{4/3} - 2}{2^{4/3} - 2}.$$

We show in Fig. 1 the exchange-correlation energy per unit volume for the non-polarized and fully polarized gases for the same density. It is clear that, to a good approximation, these energies are power series in the densities n and n_s respectively. In developing the theory of clusters in IV, it is necessary to simplify the expressions for E_{xc}. For the non-polarized case, we can write

$$E_{xc} = \Omega A n^p, \tag{12}$$

where p is 1.30917. This fit is accurate to within 0.002 a.u. for $n < 1.2$. For larger values the error increases with E_{xc} but the percentage error is less than 3% for n up to 15.

For polarized gases, we use

$$E_{xc} = \Omega \sum_{i,s} A_i n_s^{p_i+1} n_{1-s}^{q_i}, \tag{13}$$

where A_i, p_i and q_i are given in Table 1. The error in this expression is less than 0.001 a.u. for $n_\uparrow, n_\downarrow < 1$. For larger density values it is desirable to use the values of A_i', p_i' and q_i' also given in Table 1. The error then is less than 3% for large n but is 4% for n around 0.1.

It is necessary when dealing with the core electrons of heavy elements to multiply ϵ_x by a small factor arising from relativistic corrections.

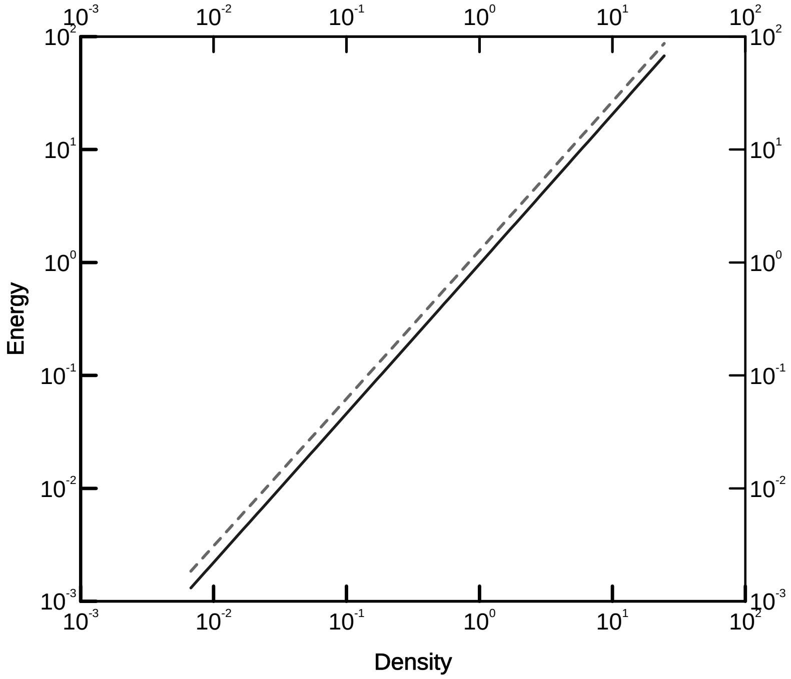

Figure 1: Variation of polarized (full) and non-polarized (dashed) exchange-correlation energy, × (-1) a.u., per unit volume with density.

Table 1: Parametrization of the exchange-correlation energy

	γ	β_1	β_1	
Non-polarized	-.1423	1.0529	0.3334	
Polarized	-.0843	1.3981	0.2611	
	A	B	C	D
Non-polarized	0.0311	-0.0480	0.0020	-0.0116
Polarized	0.0155	-0.0269	0.0007	-0.0048
	i	A_i	p_i	q_i
	1	-0.9305	0.3333	0
	2	-0.0361	0	0
	3	0.2327	0.4830	1
	4	-0.2324	0	1
	i	A'_i	p'_i	q'_i
	1	-0.9305	0.3333	0.
	2	-0.0375	0.1286	0.
	3	-0.0796	0.	0.1286

5. DENSITY FUNCTIONAL THEORY

The difficulty of evaluating the exchange energy and the need to include correlation has prompted the development of alternative methods. Density functional theory is one such development which has proved to be very successful. There are several ways of deducing the relevant equations. The simplest approach is to argue that the non-local exchange energy in Hartree-Fock theory is a very complicated integral which is considerably simplified if we treat the inhomogeneous problem locally as jellium and replace the exchange energy in Eq. (7) with its known electron gas value:

$$\int n(\mathbf{r})\epsilon_{xc}(n_\uparrow, n_\downarrow)d\mathbf{r}.$$

The 'up' and 'down' spin densities are defined in terms of the orbitals $\psi_\lambda(\mathbf{r}, s)$ through:

$$n_s(\mathbf{r}) = \sum_\lambda \delta(s, s_\lambda)|\psi_\lambda(\mathbf{r}, s)|^2.$$

The theory then proceeds by minimizing the total energy with respect to the orbitals.

This approach, called local spin density functional (LSDF) theory, has several advantages. Many of the problems with Hartree Fock theory are solved and it is far more efficient computationally. Nevertheless, there are problems caused by the replacement of the Hartree-Fock exchange energy by its electron gas value. In particular we have introduced a self-interaction term. In Eq. (4) the diagonal term with $\lambda = \mu$ in the expression for the Hartree energy cancels out the diagonal term in the exchange energy. This is no longer the case in LSDF theory and there is then a potential due to an electron acting on itself. This has the consequence that when the theory is applied to the H atom, for example, the $1s$ energy level is found to be -0.269 a.u. and the energy of the neutral atom is -0.479 a.u. instead of each being -.5 a.u. For this reason, the ionization energies of atoms are not in very good agreement with experimental values. This deficiency can be corrected by incorporating an extra term which removes the self-interaction (Perdew and Zunger, 1981). This approach has been successful in treating transition metal oxides (Svane and Gunnarsson, 1990; Szotek *et al.,* 1993).

The more usual approach to LDF theory is based on the work of Hohenberg and Kohn (1964), and Kohn and Sham (1965). These authors showed there is a 1:1 correspondence between a non-degenerate non-polarized ground state wavefunction $\Psi(r)$ and the electron density $n(\mathbf{r}_1)$ defined by

$$n(\mathbf{r}_1) = \sum_\mu \int \delta(\mathbf{r}_1 - \mathbf{r}_\mu)|\Psi(r)|^2 dr.$$

The proof rests on the preliminary result that in the Hamiltonian

$$H = T + V_{e-e} + V_{e-i},$$

the ground state electron density is in 1-1 correspondence with the external potential V_{e-i}. Suppose this is false, *i.e.* there exist two external potentials V_1 and V_2 having the same n. Then from the variational principle, if Ψ_1 and Ψ_2 are the corresponding normalized wavefunctions, and if H_i is the Hamiltonian with potential V_i and energy E_i, then

$$\begin{aligned} E_1 = \langle\Psi_1|H_1|\Psi_1\rangle &< \langle\Psi_2|H_1|\Psi_2\rangle \\ &= E_2 + \langle\Psi_2|V_1 - V_2|\Psi_2\rangle \\ &= E_2 + \int (V_1 - V_2)n(\mathbf{r})d\mathbf{r}. \end{aligned}$$

But in a similar way we can show

$$E_2 < E_1 + \int (V_2 - V_1) n(\mathbf{r}) d\mathbf{r}.$$

Adding these equations gives us

$$E_2 + E_1 < E_1 + E_2,$$

which is a contradiction unless the states are degenerate.

This shows that V_{e-i} is determined uniquely by n and hence the Schrödinger equation for Ψ can be solved in terms of n. Thus Ψ is a unique functional of n. Since E is the expectation of H with respect to Ψ, it also follows that E is determined uniquely by n. This is a remarkable result as it shows that the many-body wavefunction dependent on the set of $\mathbf{r}_\mu$ and s_μ, *i.e.* $4M$ variables where M is the number of electrons, is determined uniquely by a function of three variables, *i.e.* $n(\mathbf{r})$. Clearly instead of trying to guess the wavefunction as in Hartree-Fock theory and minimize this function of $4M$ variables, it is advantageous to try to find how the energy depends on $n(\mathbf{r})$ and minimize this as a functional of $n(\mathbf{r})$. In fact, subsequent to the work of Hohenberg and Kohn, it has been shown that this result also holds if the ground state is degenerate.

Hohenberg and Kohn went on to show that all the terms in the expression for the total energy may be evaluated as functionals of the charge density:

$$E[n] = T[n] + E_{e-i}[n] + E_H[n] + E_{xc}[n] + E_{i-i}. \tag{14}$$

Clearly, $E_{e-i}[n]$ and $E_H[n]$ are manifestly functionals of the charge density:

$$\begin{aligned} E_{e-i}[n] &= -\int n(\mathbf{r}) \sum_a \frac{Z_a}{|\mathbf{r} - \mathbf{R}_a|} d\mathbf{r}, \\ E_H[n] &= \frac{1}{2} \int \frac{n(\mathbf{r}_1) n(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2, \end{aligned}$$

and, as before, the ion-ion term is given by

$$E_{i-i} = \frac{1}{2} \sum_{a \neq b} \frac{Z_a Z_b}{|\mathbf{R}_a - \mathbf{R}_b|}.$$

The main difficulty is to write down expressions for the exchange-correlation and kinetic energies as functionals of the charge density. The exchange correlation is apparently the more challenging of the two, as this is describing the complicated many-body interactions that take place. An exact expression for this is not available, and in practice a number of approximate forms are used, the most common being the local density approximation (LDA) and the local spin density approximation (LSDA). LDA is for systems that are not spin polarized, and the exchange correlation energy is written as

$$E_{xc}[n] = \int n(\mathbf{r}) \epsilon_{xc}(n) d\mathbf{r}.$$

where the exchange-correlation energy density, $\epsilon_{xc}(n)$ is the function obtained for the homogeneous electron gas and detailed in the previous section. It is readily seen that this is a *local* approximation – it is assumed that for any small region in the system the contribution made to the exchange-correlation energy is just the same as in a uniform electron gas with the same density.

In the LSDA, we write

$$E_{xc}[n] = \int n(\mathbf{r})\epsilon_{xc}(n_\uparrow, n_\downarrow)d\mathbf{r}.$$

which is similar in principle to the LDA, but uses the expression for the energy density of a polarized homogeneous electron gas.

The use of this approach in a real system (*i.e.* with a varying charge density) is an approximation, but it is one that has been shown to be successful and to have acceptable accuracy for modeling materials containing atoms from many parts of the periodic table.

Recently, there has been interest in going beyond local expressions for exchange and correlation and to include some terms dependent on the gradient of the density (see Kutzler and Painter, 1992, for applications to first row diatomic molecules). The general experience is that the energy is improved but at the cost of an inferior structure.

It is possible to write down a functional for the kinetic energy $T[n]$ using the same approach as used for the exchange-correlation, i.e. using the result for the homogeneous electron gas:

$$T[n] = \int \frac{3n(\mathbf{r})}{10}\left(3\pi^2 n\right)^{\frac{2}{3}} d\mathbf{r}.$$

We then arrive at a similar result to the Thomas–Fermi theory. Unfortunately, this is not accurate enough to describe the small changes in total energies that take place on chemical bonding. This problem was solved by Kohn and Sham (1965). They introduced a set of orthonormal orbitals as a basis for the charge density. In the spin polarized theory the spin densities would be written in terms of these as

$$n_s(\mathbf{r}) = \sum_\lambda \delta(s_\lambda, s)|\psi_\lambda(r)|^2.$$

This is effectively claiming that the charge density can always be written as that derived from a wavefunction consisting of a single Slater determinant. It should be emphasised that this is wholly different to Hartree Fock theory where the many–particle wavefunction was written as a single determinant and used as a variational function. In the Kohn–Sham procedure, this wavefunction is used only as a means of expanding the charge density and at no stage is the energy considered to be obtainable from an expression of the form $\langle\Psi|H|\Psi\rangle$. Strictly, therefore, we cannot interpret the Kohn-Sham orbitals as single particle states.

In terms of these, the kinetic energy can be written down as

$$T = -\frac{1}{2}\sum_{\lambda,s}\int \psi_\lambda^* \nabla^2 \psi_\lambda d\mathbf{r}$$

which completes the terms that make up the total energy.

The Kohn–Sham orbitals will be determined by the requirement that E is minimized with respect to n_s subject to the total number of electrons M and spin S being fixed where,

$$M = \sum_s \int n_s(\mathbf{r})d\mathbf{r},$$

$$S = \int (n_\uparrow - n_\downarrow)d\mathbf{r}.$$

The Kohn-Sham equations are now derived by a variational principle remembering that the orbitals ψ_λ are orthonormal. Thus the quantity,

$$E - \sum_\lambda E_\lambda\{\sum_s \int |\psi_\lambda(r)|^2 d\mathbf{r} - 1\},$$

is minimized with respect to $\psi_\lambda, \psi_\lambda^*$ and E_λ. This yields:

$$\{-\frac{1}{2}\nabla^2 - \sum_a \frac{Z_a}{|\mathbf{r}-\mathbf{R}_a|} + V^H(\mathbf{r}) + V^{xc}_{s_\lambda}(n_\uparrow, n_\downarrow) - E_\lambda\}\psi_\lambda(r) = 0$$

$$\sum_s \int |\psi_\lambda(r)|^2 d\mathbf{r} = 1.$$

Here

$$V^{xc}_s = \frac{d(n\epsilon_{xc})}{dn_s}.$$

The main differences with the Hartree-Fock equations are the exchange-correlation term and the interpretation of the energy levels E_λ. In the case of jellium, the density of states $N(E)$ is no longer zero at E_f and hence identifying the Kohn-Sham energies E_λ with quasi-particle energies is natural. This, however, is not strictly correct. Janak's Theorem (Janak, 1978) asserts that if we change the occupancy of level λ by δf_λ, then the change of energy is $E_\lambda \delta f_\lambda$. This is not the same as the energy change that results from the addition or subtraction of a single electron. Such quasi-particle energies should be derived from extensions of the theory: the GW approximation of Hedin (1969) being the most successful. Implementation of this theory has given fundamental energy gaps to within a few tenths of an eV (Delsole *et al.,* 1994).

The exchange-correlation potential V^{xc}_s depends on the densities of both 'up' and 'down' spins. Consequently the spin 'up' solutions are not the same as the spin 'down' ones if S is non-zero. This means that there are two sets of equations each corresponding to a particular spin-state.

An alternative formula for E is found by multiplying by $\psi^*_\lambda(r)$ and integrating, followed by a sum over the occupied orbitals λ.

$$E = \sum_\lambda E_\lambda + E_{i-i} - E_H + \sum_s \int n_s(\mathbf{r})\{\epsilon_{xc} - V^{xc}_s)\} d\mathbf{r}. \qquad (15)$$

This formula is the starting point of approximate methods such as tight binding schemes (Sutton *et al.,* 1988). It is argued that the terms E_{i-i} and E_H largely cancel; as they must for jellium. For simple systems, the sum of energy eigenvalues acts as an attractive potential — it gets more positive as the separation between atoms increases — but the other three terms combine to act as a repulsive potential. It is often assumed that the repulsive one is short ranged and falls off quickly to zero. The matrix elements of a tight binding Hamiltonian are constructed by either fitting to a band structure derived by a combination of *ab initio* theory and experiment, or evaluated from a localized basis set (Porezag *et al.,* 1995; Seifert and Eschrig, 1985; Seifert *et al.,* 1986) The short ranged repulsive potential is then fitted so that the structures of representative systems are reproduced. We have not followed this approach ourselves, and will confine ourselves to this brief comment here.

The derivation of density functional theory presented here shows it to be essentially a ground state theory. Properties of the ground state can be obtained with quantitative accuracy. To consider the ground state, the lowest–lying Kohn-Sham orbitals are filled. However, a problem arises when we have several degenerate orbitals and not all should be occupied. Examples of this occur in atoms (for example in carbon, two electrons need to be placed into the three degenerate $2p$ states) and defects (for example the vacancy in diamond where two electrons need to be placed into three degenerate states of t_2 symmetry. This is essentially a multiplet problem and should lie beyond density functional theory. However, a method for obtaining approximate multiplet energies has been given by von Barth (1979). In this approach, the density functional energy is taken to be the energy of a single Slater determinant, the contents

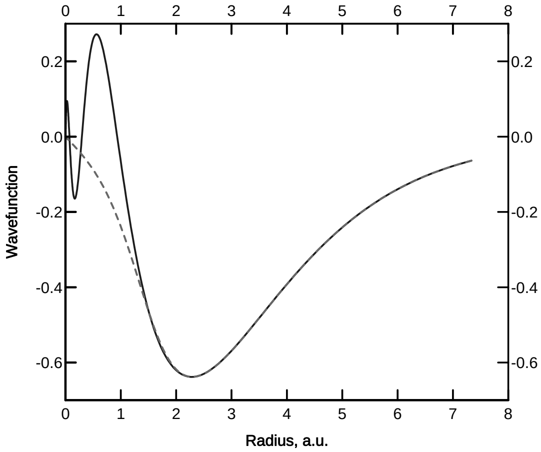

Figure 2: The $4s$ (full) and pseudo- (dashed) radial wavefunction (a.u.) for the Ni atom.

of which are governed by the choice of Kohn–Sham states to be filled. The energies of different multiplets which can be built out of determinants corresponding to different configurations may be found by writing each determinant, $|D\rangle$, as a linear combination of multiplets, $|M_a\rangle$, as

$$|D\rangle = \sum_a c_a |M_a\rangle.$$

If we operate on this with the Hamiltonian and identify the expectation value of the left hand side as the energy E found using density functional theory, then

$$E = \sum_a |c_a|^2 \langle M_a|H|M_a\rangle.$$

We can now choose different configurations – each corresponding to single Slater determinants — and then deduce several equations relating the multiplet energies which can, in favorable cases, be solved.

III. Pseudopotential theory

Pseudopotentials are a very important component of first principles calculations as they remove the need to consider core electrons – only the valence electrons need be considered. This is extremely important if one wants to be able to treat all the elements of the periodic table. Using the full Coulomb potential can cause considerable problems. The total energy then becomes extremely large and since one is interested in differences in energies between similar sized clusters, there can be a significant loss of accuracy. Secondly, the fitting of core wavefunctions with Gaussian orbitals, and even more so with plane waves, is extremely difficult and small errors can make large differences in the core eigenvalues. Fig. 2 shows the $4s$ wavefunction and pseudo-wavefunctions in Ni. It is clear that the pseudo-wavefunction is a much simpler and smoother function to approximate than the all-electron wavefunction. Thirdly, for the heavier atoms, relativistic effects are important and the Dirac equation is required. However, the valence electrons can continue to be treated non-relativistically and a spin-orbit potential can be introduced which describes polarized valence electrons.

The justification for the use of a pseudopotential lies in the fact that the highly localized core wavefunction cannot take part in the bonding of atoms. Nevertheless, the valence electrons undergo exchange interactions with core ones and this makes the problem of constructing pseudopotentials non-trivial.

The precise description of generating a pseudopotential is a complicated procedure and there are many different prescriptions given in the literature. These prescriptions differ because the resulting potentials have different uses. Methods using plane waves require pseudopotentials whose momentum matrix elements decay as quickly as possible. This is not an important requirement for the real-space basis used here although it is of great advantage to deal with smooth wavefunctions which have as few nodes as possible, for such functions are easier to represent in terms of Gaussian basis sets. Here, we follow closely the prescription given by Bachelet *et al.,* (1982) who have produced a comprehensive table of pseudopotentials for all the elements between H and Pu. These potentials are called norm-conserving since they yield the exact atomic charge density outside the core. This is an important property for self–consistent calculations.

The first step is the solution of the Kohn-Sham equations for all the electrons in an atom. This is done by choosing a configuration leading to a spherically symmetric charge density and hence the atomic Kohn-Sham levels are labeled by angular momentum numbers for light elements, and $j = l \pm 1/2$ symbols for heavy elements when the Dirac equation must be used. The final pseudopotentials will possess the same valence eigenvalues and give pseudo-wavefunctions which agree exactly with the true ones outside a core radius. Now, if one filled up the Kohn-Sham energy levels in ascending order, choosing for carbon, for example, the $1s^2 2s^2 2p^2$ configuration, then the $3d$ level would be empty. It might seem then that these unoccupied d-levels do not have to be considered in describing the binding of C atoms with other elements. This, however, is not strictly correct. The wavefunctions for states in solids will be made up of combinations of all atomic states including d-states and although the atomic d-states may not play a major role, it is not clear their influence can be neglected. The pseudopotential then should possess the same s, p and d valence energy levels as the all-electron atom for the valence states. To accommodate this, the atom is solved in the ground state configuration $1s^2 2s^2 2p^2$ for $l = 0$ and 1, but for $l = 2$ one chooses an ionized excited state configuration such as $1s^2 2s^{0.75} 2p^1 d^{0.25}$ in which these d-levels are occupied with a small amount of charge. The fractional occupancy of the s-shell is chosen to eliminate 'bumps' in the potential. Different configurations are used for other elements. In Si for example, the d-pseudopotential is derived from the configuration $3s^1 3p^{0.75} 3d^{0.25}$.

The Kohn-Sham equations for the atom using these configurations, ν, yield the all-electron wavefunctions and energy levels. The spin densities can then be found from the wavefunctions as well as the all-electron potential $V^\nu(r)$. This is the sum of the nuclear, Hartree and exchange-correlation potentials and possesses a Coulomb singularity at $r = 0$. A first-step pseudopotential for each configuration and angular momentum index l (or $j = l \pm 1/2$ for the Dirac equation) is then constructed, eliminating the Coulomb singularity by defining

$$V_l(r) = V^\nu(r)(1 - f(r/r_{c,l})) + c_l^\nu f(r/r_{c,l}).$$

Here $f(x)$ is a function which is unity at the origin and vanishes rapidly for x much bigger than 1, *e.g.* $e^{-x^{3.5}}$. Hence, for r close to the origin, V_l is a constant, c_l^ν, which is chosen so that the lowest energy level for each l is exactly the same as the solution of the all-electron atom for the same l.

The corresponding normalized pseudo-wavefunction, w_{1l}^ν, is clearly equal (up to a normalization factor) to the all-electron wavefunction for large r as the potential $V_l(r)$ is exactly the same as $V^\nu(r)$ there. The value of $r_{c,l}$ is called the core radius and it determines when the all-electron wavefunction approaches the pseudo-wavefunction. Clearly it must not be too big, but if chosen too small, then it lies in a region where the wavefunction is rapidly varying and difficult to represent by any basis. It is usually chosen to be about half-way to the outermost node and the outermost extrema of the all-electron wavefunction.

The next step is to modify w_{1l}^{ν} by introducing a second wavefunction w_{2l}^{ν} by

$$w_{2l}^{\nu} = \gamma_l^{\nu}\{w_{1l}^{\nu}(r) + \delta_l^{\nu} r^{l+1} f(r/r_{cl})\}.$$

The constants γ_l^{ν} and δ_l^{ν} are selected so that the normalized function w_{2l}^{ν} agrees exactly with the all-electron wavefunction outside the core and is not just proportional to it. The potential giving rise to the w_{2l}^{ν} wavefunctions is then found by inverting the Schrödinger equation using the energy levels which agree with the all-electron values. This potential then has the correct eigenvalues and a wavefunction which agrees exactly with the all-electron one outside the core. Finally, the contribution of the Hartree and exchange-correlation potentials arising from the *pseudo-wavefunctions* w_{2l}^{ν} are subtracted leaving a bare ion potential $V_l(r)$. This last step is exact for the Hartree potential as this is linear in the core and valence charge densities. However, it is an approximation for the non-linear exchange-correlation potential. The approximation can be improved, along with the transferability of the pseudopotential, by subtracting instead the exchange-correlation potential corresponding to the all-electron charge density and spin-polarization (Louie *et al.,* 1982).

For relativistic atoms, where the states are labeled by $j \pm \frac{1}{2}$, an average pseudopotential is defined:

$$V_l(r) = \frac{1}{2l+1}\{lV_{l-1/2}(r) + (l+1)V_{l+1/2}(r)\}.$$

This is called the scalar relativistic potential. The spin-orbit potential is:

$$V_l^{so}(r) = \frac{2}{2l+1}\{V_{l+1/2}(r) - V_{l-1/2}(r)\},$$

and the full pseudopotential is then:

$$V^{ps}(\mathbf{r}) = \sum_l |l\rangle\{V_l(r) + V_l^{so}(r)\mathbf{L}.\mathbf{S}\}\langle l|. \tag{16}$$

The potentials have been parametrized by fitting them to simple functions in the following way:

$$V_l(r) = -\frac{Z_\nu}{r}\{\sum_{i=1}^{2} c_i^c \; erf(\sqrt{\alpha_i^c} r)\} + \sum_{i=1}^{3}\{A_{i,l} + r^2 A_{i+3,l}\}e^{-\alpha_{i,l}r^2}$$

$$V_l^{so}(r) = \sum_{i=1}^{3}\{B_{i,l} + r^2 B_{i+3,l}\}e^{-\alpha_{i,l}r^2}.$$

Here Z_ν is the valence charge, α_i^c is the inverse of the extent of the core charge density and erf is the error function. The coefficients c_i^c are independent of l and hence this first term is a simple function called the local pseudopotential. The sum of the coefficients c_i^c is unity so the local potential gives rise to a potential $-Z_\nu/r$ for $\alpha_i^c r^2 >> 1$. The second term does depend on l and is called the non-local pseudopotential. Tables of values of $Z_\nu, c_i^c, \alpha_i^c, A_{i,l}, B_{i,l}$ and $\alpha_{i,l}$ are given in Bachelet *et al.,* (1982).

A crucial property of the pseudo-wavefunction is that it can accurately describe different bonding configurations. One test is to compare the energy differences between configurations corresponding to the promotion of valence electrons. For example, the energy difference between C in the s^2p^2 and sp^3 configurations is 8.23 eV when the all-electron theory is used and 8.25 eV using the pseudopotential. The agreement is not quite as good for Ni, as the corresponding $d^8s^2 \rightarrow d^9s^1$ energies are -1.66 and -1.36 eV respectively.

IV. The real space cluster method

We now discuss applying LSDF theory described above to a cluster of atoms. The wavefunctions of the cluster are expanded in a basis of localized orbitals $\phi_i(\mathbf{r}-\mathbf{R}_i)$ as:

$$\psi_\lambda(\mathbf{r},s)=\chi_\alpha(s)\sum_i c_i^\lambda\phi_i(\mathbf{r}-\mathbf{R}_i). \tag{17}$$

In this way the Kohn-Sham differential equations are converted to matrix equations for c_i^λ. The localized orbitals are often taken to be Gaussian ones of the form:

$$(x-R_{ix})^{n_1}(y-R_{iy})^{n_2}(z-R_{iz})^{n_3}e^{-a_i(\mathbf{r}-\mathbf{R}_i)^2},$$

where n_1, n_2 and n_3 are integers. If these are all zero they correspond to s-orbitals of spherical symmetry. Orbitals of p-symmetry correspond to one of these integers being unity and the others zero, whereas five d-like and one s-like orbital can be generated if $\sum_i n_i=2$.

The advantage with Gaussian orbitals is that the many integrals required can be evaluated analytically but their disadvantages are that they quickly become 'over-complete', and unlike Slater orbitals they do not individually approximate solutions to the Kohn-Sham equations. The over-completeness is exemplified by the singular nature of the overlap matrix when two orbitals with similar exponents are sited too close together. The basis functions are real and hence all matrix elements are real as well as the coefficients c_i^λ. We can therefore drop complex conjugates from the equations.

The density for each spin is then given in terms of the density matrix, $b_{ij,s}$,

$$n_s(\mathbf{r})=\sum_{ij}b_{ij,s}\phi_i(\mathbf{r}-\mathbf{R}_i)\phi_j(\mathbf{r}-\mathbf{R}_j),$$

$$b_{ij,s}=\sum_\lambda\delta(s,s_\lambda)c_i^\lambda c_j^\lambda. \tag{18}$$

The sum is over occupied orbitals λ with spin s. The charge density $n(\mathbf{r})$ can be written:

$$n(\mathbf{r})=\sum_s n_s(\mathbf{r})=\sum_{ij}b_{ij}\phi_i(\mathbf{r}-\mathbf{R}_i)\phi_j(\mathbf{r}-\mathbf{R}_j),$$

$$b_{ij}=\sum_s b_{ij,s}.$$

Let us now consider the various terms in the LSDF expression for the energy E in Eq. (14) when this basis of localized orbitals is used. The kinetic energy and pseudopotential terms involve integrals of the form:

$$T_{ij}=-\frac{1}{2}\int\phi_i(\mathbf{r}-\mathbf{R}_i)\nabla^2\phi_j(\mathbf{r}-\mathbf{R}_j)d\mathbf{r}$$

$$V_{ij}^{ps}=\int\phi_i(\mathbf{r}-\mathbf{R}_i)\sum_a V_a^{ps}(\mathbf{r}-\mathbf{R}_a)\phi_j(\mathbf{r}-\mathbf{R}_j)d\mathbf{r},$$

which are easily found. The evaluation of the Hartree energy requires $O(N^4)$ integrals, where N is the number of basis functions, which is prohibitively large for a cluster where N might be 1000 or more. Many of these integrals are negligible, particularly those associated with basis functions with fast decay rates. However, many remain. For example, if the smallest exponent a_i is about 0.1 a.u., then the 'overlap' of two such orbitals will be non-negligible for separations of centers less than about 15 a.u. Thus for Si and diamond, there would be between 100 and 500 atoms in a sphere of this size. This means for most of the clusters considered, *all* these integrals need to be evaluated. For this reason it is essential to approximate the Hartree energy in some way. One way is to carry out a numerical integration over a finely meshed grid (Pederson *et al.,* 1991; Chen *et al.,* 1995). However, unless the mesh is chosen very carefully the resulting matrix elements of the Hartree and exchange-correlation potentials will not transform correctly under the point group operations of the cluster. This will mean that the eigenfunctions of the Hamiltonian, and the normal-coordinates of vibrational modes, will not have the correct symmetries, nor the eigenvalues the correct degeneracy. There is then a great advantage in being able to compute these integrals using an analytic formula which preserves any point group symmetry.

Usually this is done by introducing an approximate but analytic expression for the Hartree and exchange-correlation energies from which the corresponding potentials can be easily found. In addition, it is essential to be able to differentiate them with respect to the positions of the nuclei so that the forces acting on each atom can also be found analytically. These approximate expressions are written in terms of an approximate density for each spin $\tilde{n}_s$ which is expanded in a set of basis functions (Dunlap *et al.,* 1979; Jones and Sayyash, 1986):

$$\tilde{n}_s(\mathbf{r}) = \sum_k d_{k,s} g_k(\mathbf{r}), \quad \tilde{n}(\mathbf{r}) = \sum_s \tilde{n}_s(\mathbf{r}).$$

We now consider these approximate expressions for the Hartree and exchange-correlation energies.

1. THE HARTREE ENERGY

The exact Hartree energy is, from Eq. (30),

$$E_H = \frac{1}{2}\int \frac{n(\mathbf{r}_1)n(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2,$$

and is replaced by an approximate value, $\tilde{E}_H$, involving an approximate charge density $\tilde{n}$

$$\tilde{E}_H = \int \frac{n(\mathbf{r}_1)\tilde{n}(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2 - \frac{1}{2}\int \frac{\tilde{n}(\mathbf{r}_1)\tilde{n}(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2.$$

The replacement is exact when $\tilde{n} = n$. Now, we expand the density in terms of a basis set $g_k(\mathbf{r})$ so that

$$\tilde{n}(\mathbf{r}) = \sum_k c_k g_k(\mathbf{r}),$$

and c_k is chosen to minimize the error in estimating the Hartree energy:

$$E_H - \tilde{E}_H = \frac{1}{2}\int \frac{\{n(\mathbf{r}_1) - \tilde{n}(\mathbf{r}_1)\}\{n(\mathbf{r}_2) - \tilde{n}(\mathbf{r}_2)\}}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2.$$

Differentiating this with respect to c_k to determine the minimum gives:

$$\sum_l G_{kl} c_l = \sum_{ij} t_{ijk} b_{ij}. \tag{19}$$

Here,

$$t_{ijk} = \int \phi_i(\mathbf{r}_1 - \mathbf{R}_i)\phi_j(\mathbf{r}_1 - \mathbf{R}_j) g_k(\mathbf{r}_2) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2,$$

$$G_{kl} = \int g_k(\mathbf{r}_1) g_l(\mathbf{r}_2) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2.$$

We notice that $\tilde{E}_H$ is always bounded above by E_H and the quality of the fit can be assessed by the increase in $\tilde{E}_H$ when the number of basis functions is increased.

We now consider the choice of g_k. The simplest consists of Gaussian functions $e^{-b_k(\mathbf{r}-\mathbf{R}_k)^2}$ defined by a site $\mathbf{R}_k$ and an exponent b_k. The sites need not correspond to the location of atoms but can include, for example, bond centers. All the integrals can be computed analytically which leads to considerable time saving. However, for clusters of less than about 100 atoms, it is the evaluation of the t_{ijk} which is often the most time consuming procedure. The number of basis functions g_k is usually proportional to the number of basis functions ϕ_i, *i.e.* N, and hence there are $O(N^3)$ integrals of the type t_{ijk}. These cannot be stored in main memory and must either be evaluated once and stored on disk or, for very fast processors, be repeatedly evaluated during each of the self-consistent cycles. There is then an advantage in choosing a set of g_k which leads to a simple analytical form for t_{ijk} which can be evaluated very quickly.

This can be done when g_k is divided into two sets. The first set has g_k defined by:

$$g_k = \{1 - \frac{2b_k}{3}(\mathbf{r} - \mathbf{R}_k)^2\} e^{-b_k(\mathbf{r}-\mathbf{R}_k)^2}. \tag{20}$$

These functions give a potential of Gaussian form:

$$\int \frac{g_k(\mathbf{r}_1)}{|\mathbf{r} - \mathbf{r}_1|} d\mathbf{r}_1 = \frac{3b_k}{2\pi} e^{-b_k(\mathbf{r}-\mathbf{R}_k)^2},$$

and thus integrals in t_{ijk} involve a product of three Gaussian functions and can be evaluated very quickly. Also, many of these are now zero, as all three Gaussians must overlap to give a non-zero value. However, in order to get the short–ranged Gaussian potential and to avoid the long-ranged coulomb potential, the integral of g_k must vanish. It is readily verified that this indeed happens for the functions in Eq. (20). It is therefore necessary to select additional functions g_k which are purely Gaussian and whose integrals do contribute to the total number of electrons. It is, however, sometimes possible to choose the coefficients of these as fixed quantities related to the anticipated total charge on the atom or ion. Thus their contribution to the Hamiltonian does not change during the self-consistent cycle and they behave in the same way as the external potential of the nuclei. This leads to a considerable speed up in the code.

2. THE EXCHANGE-CORRELATION ENERGY

In the same way the exact expression for this energy,

$$E_{xc} = \int \epsilon_{xc}(n_\uparrow, n_\downarrow) n d\mathbf{r},$$

is replaced by an approximate one $\tilde{E}_{xc}$ involving an approximate density, $\tilde{n}_s$.

$$\tilde{E}_{xc} = \int \epsilon_{xc}(\tilde{n}_\uparrow, \tilde{n}_\downarrow) \tilde{n} d\mathbf{r}. \tag{21}$$

Clearly the error we make is negligible if $\tilde{n}_s$ is close to n_s. The first step then is to fit n_s to a set of functions. It is possible to choose the same $g_k(\mathbf{r})$ as was used in the construction of $\tilde{E}_H$. However, the least squares procedure used there minimizes the electrostatic energy of the error in the charge density, *i.e.* $n - \tilde{n}$, and it does not mean that at each value of $\mathbf{r}$, $n(\mathbf{r})$ and $\tilde{n}(\mathbf{r})$

are almost equal. Moreover, the choice of g_k was selected to reflect the difficulty of working out the integrals t_{ijk}. Hence, in dealing with $\tilde{E}_{xc}$ it is better to use a sum of simple Gaussian functions h_k so that:

$$\tilde{n}_s(\mathbf{r}) = \sum_k d_{k,s} h_k(\mathbf{r}), \tag{22}$$

where $d_{k,s}$ is found from minimizing

$$\int \{n_s(\mathbf{r}) - \tilde{n}_s(\mathbf{r})\}^2 d\mathbf{r}.$$

Differentiating this with respect to the coefficients $d_{k,s}$ leads to the equations:

$$\sum_l H_{kl} d_{l,s} = \sum_{ij} u_{ijk} b_{ij,s}, \tag{23}$$

where,

$$H_{kl} = \int h_k(\mathbf{r}) h_l(\mathbf{r}) d\mathbf{r}, \tag{24}$$

$$u_{ijk} = \int \phi_i(\mathbf{r} - \mathbf{R}_i)\phi_j(\mathbf{r} - \mathbf{R}_j) h_k(\mathbf{r}) d\mathbf{r}. \tag{25}$$

We note that the integrals are the same for each spin-index s and that u_{ijk} are simply proportional to t_{ijk} if g_k is chosen as in Eq. (20) above. This saves a considerable amount of computer time.

We consider first the non-polarized or spin-averaged case where we can dispense with the spin label and write:

$$\tilde{E}_{xc} = \sum_k d_k \int h_k(\mathbf{r}) \epsilon_{xc}(\tilde{n}) d\mathbf{r}.$$

If h_k is chosen to be a positive definite localized function such as a Gaussian, then each integral is proportional to the average value of the exchange-correlation density under h_k,

$$\langle \epsilon_{xc}(\tilde{n}) \rangle_k.$$

We next note from Fig. 1 that $\epsilon_{xc}(n)$ varies slowly with n and hence we expect

$$\langle \epsilon_{xc}(\tilde{n}) \rangle_k \approx \epsilon_{xc}(\langle \tilde{n} \rangle_k),$$

$$\langle \tilde{n} \rangle_k = \frac{\sum_l d_l \int h_k h_l d\mathbf{r}}{I_k}, \tag{26}$$

where I_k is simply the integral of h_k. This approximation is tantamount to replacing the exact exchange-correlation density at $\mathbf{r}$ by its homogeneous electron gas value for the average density $\langle \tilde{n} \rangle_k$. We can improve on this approximation as follows. Now, as discussed in II.4, the exchange-correlation density behaves with high accuracy as a power series in n,

$$\epsilon_{xc}(n) = A n^s$$

with $s = 0.30917$. Let us now consider the function $f(s)$ where

$$f(s) = \ln\left(\frac{\langle \tilde{n}^s \rangle_k}{\langle \tilde{n} \rangle_k^s}\right).$$

Clearly for $s = 0$ or 1, $f(s)$ is 0, while for $s = 2$, $f(s)$ must be a positive quantity. Since we are interested in values of s around 0.3, we can approximate $f(s)$ by

$$f(s) = s(s-1)f(2)/2.$$

The right hand side of this equation is found from the second moment of $\tilde{n}$, ie,

$$\langle \tilde{n}^2 \rangle_k = \frac{\sum_{lm} d_l d_m \int h_k h_l h_m d\mathbf{r}}{I_k}. \tag{27}$$

These integrals can all be evaluated analytically. We have finally:

$$\tilde{E}_{xc} = \sum_k d_k \epsilon_k, \tag{28}$$

where,

$$\epsilon_k = I_k \epsilon_{xc}(\langle \tilde{n} \rangle_k) e^{f_k},$$
$$f_k = \frac{1}{2} s(s-1) ln\Big(\frac{\langle \tilde{n}^2 \rangle_k}{\langle \tilde{n} \rangle_k{}^2}\Big).$$

This theory has been extended to the spin-polarized case (Lister and Jones, 1988). The spin-polarized exchange-correlation energy is written as in Eq. (13),

$$E^{xc}(n_\uparrow, n_\downarrow) = \sum_{i,s} A_i \int n_s^{p_i+1} n_{1-s}^{q_i} d\mathbf{r},$$

and we replace n_s on right hand side by $\tilde{n}_s$ obtaining:

$$\tilde{E}_{xc} = \sum_{ks} d_{k,s} \epsilon_{k,s}, \tag{29}$$

where,

$$\epsilon_{k,s} = \sum_i A_i I_k \langle \tilde{n}_s^{p_i} \tilde{n}_{1-s}^{q_i} \rangle_k.$$

Now, we define the quantity f by:

$$\langle \tilde{n}_s^p \tilde{n}_{1-s}^q \rangle_k = \langle \tilde{n}_s \rangle_k^p \langle \tilde{n}_{1-s} \rangle_k^q e^{f(p,q)}$$
$$f(p,q) = ln\Big(\frac{\langle \tilde{n}_s^p \tilde{n}_{1-s}^q \rangle_k}{\langle \tilde{n}_s \rangle_k^p \langle \tilde{n}_{1-s} \rangle_k^q}\Big).$$

We now approximate f by the formula:

$$f(p,q) = \frac{1}{2} p(p-1) f(2,0) + \frac{1}{2} q(q-1) f(0,2) + pq f(1,1),$$

which interpolates f between the known integer values. In this way the spin-polarized exchange-correlation energy is evaluated.

3. MATRIX FORMULATION

In terms of the approximate Hartree and exchange-correlation energies, the total energy can now be written:

$$E = \sum_{ij} \{T_{ij} + V_{ij}^{ps}\} b_{ij} + \tilde{E}_H + \tilde{E}_{xc} + E_{i-i}, \tag{30}$$

where,

$$b_{ij} = \sum_{\lambda} c_i^{\lambda} c_j^{\lambda}$$
$$\tilde{E}_H = \frac{1}{2} \sum_{kl} c_k c_l G_{kl}$$
$$\tilde{E}_{xc} = \sum_{ks} d_{k,s} \epsilon_{k,s},$$

and E_{i-i} is given by (8). The fitting coefficients c_k and $d_{k,s}$ are defined in terms of $b_{ij,s}$ by Eqs. (19) and (23).

E is minimized subject to an orthonormal set of wavefunctions, *i.e.*

$$\sum c_i^{\lambda} c_j^{\mu} S_{ij} = \delta_{\lambda\mu},$$

where the overlap matrix S, is defined by:

$$S_{ij} = \int \phi_i(\mathbf{r} - \mathbf{R}_i) \phi_j(\mathbf{r} - \mathbf{R}_j) d\mathbf{r}. \tag{31}$$

This can be done by introducing Lagrange undetermined multipliers, E_{λ}, so that we minimize without constraint,

$$\sum_{ij\lambda} c_i^{\lambda} \{T_{ij} + V_{ij}^{ps} - E_{\lambda} S_{ij}\} c_j^{\lambda} + \tilde{E}_H + \tilde{E}_{xc} + E_{i-i}, \tag{32}$$

with respect to c_i^{λ}. Now, this introduces the matrix elements of the Hartree and exchange-correlation potentials through:

$$\frac{\partial \tilde{E}_H}{\partial c_i^{\lambda}} = \sum_j V_{ij}^H c_j^{\lambda}, \quad V_{ij}^H = \sum_{kl} G_{kl} c_l \frac{\partial c_k}{\partial b_{ij}}$$
$$\frac{\partial \tilde{E}_{xc}}{\partial c_i^{\lambda}} = \sum_j V_{ij,s_{\lambda}}^{xc} c_j^{\lambda}, \quad V_{ij,s}^{xc} = \sum_k \{\epsilon_{k,s} + \sum_l d_{l,s} \frac{\partial \epsilon_{l,s}}{\partial d_{k,s}}\} \frac{\partial d_{k,s}}{\partial b_{ij,s}}$$

From Eqs. (19) and (23), we find:

$$\sum_l G_{kl} \frac{\partial c_l}{\partial b_{ij}} = t_{ijk}$$
$$\sum_l H_{kl} \frac{\partial d_{l,s}}{\partial b_{ij,s}} = u_{ijk}.$$

Differentiating Eq. (32) with respect to c_i^{λ} we get the Kohn-Sham equations:

$$\sum_j \{T_{ij} + V_{ij}^{ps} + V_{ij}^H + V_{ij,s_{\lambda}}^{xc} - E_{\lambda} S_{ij}\} c_j^{\lambda} = 0. \tag{33}$$

We note that the total number of electrons M and the spin S are given by:

$$M = \sum_{ijs} S_{ij} b_{ij,s}, \tag{34}$$

$$S = \sum_{ij} S_{ij}(b_{ij,\uparrow} - b_{ij,\downarrow}).$$

We can now write Eq. (33) more compactly in matrix form. Eq. (33) is written in terms of two generalized eigenvalue problems, one for each spin, as:

$$\sum_j (H_{ij} - ES_{ij})c_j = 0,$$

or in matrix notation as:

$$(H - ES)c = 0.$$

For the cluster sizes and values of the exponents of the basis sets typically used, the matrices H and S are not sufficiently sparse to warrant special numerical techniques and consequently the overlap matrix S is written in terms of an upper triangular matrix using a Choleski decomposition:

$$S = U^t U.$$

U and its inverse can be evaluated in $O(N^3)$ operations. We then define a vector d by $Uc = d$, and the generalized eigenvalue problem is converted into the usual one:

$$\{(U^{-1})^t H U^{-1} - E\}d = 0$$

The eigenvalues of this can be found by a standard Householder scheme which first reduces the matrix to tridiagonal form from which the eigenvalues can be found. As the number of occupied states is much smaller than the dimension of the Hamiltonian, the eigenvectors are best found by inverse iteration.

All these matrix operations can be carried out on a parallel computer (Briddon, 1996) using the PBLAS and SCALAPACK libraries which provide routines to carry out all the required operations, provided the matrices are correctly distributed between nodes. In fact the matrices are divided up into square blocks, and these are allocated to different nodes in a specified manner. In this way efficient parallel code is easily written. The time dominant step for large clusters is the computation of the eigenvalues and eigenvectors which scales as N^3 for the dense matrices found in practice.

V. Self-consistency and atomic forces

1. SELF-CONSISTENCY

Self-consistency is the situation obtained after a successful solution of the Kohn–Sham equations when the charge density that would be produced by the Kohn–Sham orbitals gives rise to the same potential as was used in the equation that determined them. In short, it is the process by which charge is distributed around the cluster minimizing its energy for a fixed structure. The self-consistency cycle is initiated by choosing sets of charge density coefficients c_k and $d_{k,s}$

taken from either neutral atoms, or a previous run. The Kohn-Sham equations given in Eq. (33) are then solved and the density matrix $b_{ij,s}$ found. Eqs. (19) and (23) are then used to determine the output charge density coefficients, $c^o_k, d^o_{k,s}$. The next step consists of selecting a new input charge density c'_k, defined in terms of c^o_k and c_k. This is done by using a weighted combination as in:

$$c'_k = c_k + w(c^o_k - c_k).$$

The same weighting is used to define the new spin density coefficients $d'_{k,s}$. If the process of generating the output charge density is denoted by the (non-linear) operation

$$c^o_k = L_k(c),$$

where we have written c to stand for the vector c_k, then

$$c^{o\prime}_k = L_k(c') = L_k(c + w(c^o - c)).$$

Now, provided w is small enough, we can linearize this equation to get

$$\begin{aligned} c^{o\prime}_k &= L_k(c) + w\sum_l D_{kl}(c^o_l - c_l) \\ &= c^o_k + w\sum_l D_{kl}(c^o_l - c_l). \end{aligned}$$

The condition for self-consistency is that the input and output charge densities are equal, *i.e.*

$$c^{o\prime}_k = c'_k,$$

or

$$c_k + w(c^o_k - c_k) = c^o_k + w\sum_l D_{kl}(c^o_l - c_l).$$

Hence

$$(w^{-1} - 1)(c_k - c^o_k) = \sum_l D_{kl}(c^o_l - c_l). \tag{35}$$

The right hand side can be determined by choosing a small value of w, say w_1, and the output charge density c^o_{1k} then found. This gives

$$c^o_{1k} = c^o_k + w_1\sum_l D_{kl}(c^o_l - c_l).$$

Hence,

$$\sum_l D_{kl}(c^o_l - c_l) = (c^o_{1k} - c^o_k)/w_1.$$

Inserting this into Eq. (35) gives us an equation for w which is solved by a least squares procedure. We denote the difference between the sides of this equation as:

$$e_k = (1 - w)(c_k - c^o_k)/w - (c^o_{1k} - c^o_k)/w_1,$$

and choose w by minimizing the electrostatic energy of the 'charge density' e_k defined by $\sum_k e_k g_k(\mathbf{r})$. Thus the energy defined by:

$$\frac{1}{2}\sum_{kl} e_k G_{kl} e_l,$$

is made as small as possible. It is possible to generalize this procedure so that the predicted charge density is built up from several previous iterates c_{lk}, *i.e.*

$$c'_k = \sum_l \{c_{lk} + w_l(c^o_{lk} - c_{lk})\}.$$

In practice, the self-consistency cycle converges exponentially quickly, taking between four to ten iterations with the difference in the input and output Hartree energies typically becoming less than 10^{-5} a.u. Convergence is particularly rapid when there is a gap between the highest filled and lowest empty level but problems can arise when this gap is very small or vanishes. These are often related to an attempted crossing of an occupied and unoccupied energy level whereupon the charge density changes discontinuously. This can be avoided by 'smearing out' the occupation of levels by using Fermi statistics. Thus we suppose that the level E_λ is occupied by f_λ electrons. This means that the energy to be be minimized now includes an entropy term as well as a term constraining the total number of electrons to M:

$$F = E + k_B T \sum_\lambda \{f_\lambda ln f_\lambda + (1 - f_\lambda) ln(1 - f_\lambda)\} - \mu\{\sum_\lambda f_\lambda - M\}. \tag{36}$$

Here the sum is now over all orbitals λ. Minimizing the free energy F with respect to f_λ and μ gives

$$f_\lambda = \frac{1}{e^{(E_\lambda - \mu)/k_B T} + 1},$$

and

$$\sum_\lambda f_\lambda = M.$$

Eq. (18) must also be generalized to

$$b_{ij,s} = \sum_\lambda \delta(s, s_\lambda) f_\lambda c^\lambda_i c^\lambda_j.$$

In practice, $k_B T$ is taken to be about 0.04 eV. Often, where we have two energy levels separated by 0.1eV that 'cross' in the approach to self-consistency, this will remove the discontinuous change, but when self-consistency is achieved, provided the final splitting is more than 0.04 eV one state is found to be fully occupied and the other empty. In this sense, we are using variable filling purely as a computational tool and are not attempting to simulate materials at finite temperatures.

It is worth pointing out here that incorrect use of Fermi statistics can lead to incorrect structures. This occurs when a Jahn-Teller effect operates, as for example for a substitutional Ni^- impurity in Si (Jones *et al.,* 1995b). Here, the gap contains $t_2^{\uparrow\downarrow}$ levels with the upper one containing two electrons. The system distorts leading to a lowering of symmetry and a splitting of the t_2 levels into a_1, b_1 and b_2. The occupied levels will be displaced downwards and the unoccupied level upwards, leading to a lowering in the energy, provided the strain energy arising from the distortion is less than the lowering in the occupied level. Fermi statistics, however, result in an equal occupation of the levels and the driving force for the distortion vanishes. This can be overcome by, for example, occupying the b_1 and b_2 levels throughout the self-consistency cycle, even though the unoccupied a_1 level might lie below one or both of the occupied levels during part of the cycle. Such an approach is found to successfully model the defect.

There are other troublesome cases where even the use of Fermi statistics is unable to give a self-consistent solution. Often this means that the starting structure is physically unreasonable but the problem disappears once partial structural relaxation has occurred.

2. EVALUATION OF FORCES

Once the self-consistent charge density has been found, then the force acting on each atom can be evaluated. It is essential to determine the forces accurately in order to relax the cluster and calculate its vibrational modes. The force on an atom a, in direction l, is given by:

$$f_{la} = -\frac{\partial F}{\partial R_{la}} = -\frac{\partial E}{\partial R_{la}}.$$

This can be evaluated by considering the change to each term in the energy in Eq. (32) when R_{la} is displaced by ΔR_{la}. Thus

$$\Delta E = \sum_{ij} b_{ij} \Delta\{T_{ij} + V_{ij}^{ps}\} + \sum_{ij}\{T_{ij} + V_{ij}^{ps}\}\Delta b_{ij} + \Delta\tilde{E}_H + \Delta\tilde{E}_{xc} + \Delta E_{i-i}$$

$$\Delta\tilde{E}_H = \sum_{kl} c_k G_{kl} \Delta c_l + \frac{1}{2}\sum_{kl} c_k c_l \Delta G_{kl}$$

$$\Delta\tilde{E}_{xc} = \sum_{k,s} \epsilon_{k,s} \Delta d_{k,s} + \sum_{k,s} d_{k,s} \Delta\epsilon_{k,s}.$$

Δc_k can be evaluated from Eq. (19):

$$\sum_l G_{kl}\Delta c_l = \sum_{ij}\{t_{ijk}\Delta b_{ij} + b_{ij}\Delta t_{ijk}\} - \sum_l c_l \Delta G_{kl}.$$

In the same way $\Delta d_{k,s}$ can be evaluated from Eq. (23):

$$\sum_l H_{kl}\Delta d_{l,s} = \sum_{ij}\{u_{ijk}\Delta b_{ij,s} + b_{ij,s}\Delta u_{ijk}\} - \sum_l d_{l,s}\Delta H_{kl}.$$

Now, if we gather together the terms in Δb_{ij} and $\Delta b_{ij,s}$ we get,

$$\sum_{ij}\{T_{ij} + V_{ij}^{ps} + V_{ij}^{H}\}\Delta b_{ij} + \sum_{ijs} V_{ij,s}^{xc}\Delta b_{ij,s}.$$

From the Kohn-Sham Eq. (33) this equals $\sum_{ij\lambda} E_\lambda S_{ij} \Delta c_i^\lambda c_j^\lambda$, which can be written as:

$$\sum_\lambda E_\lambda \Delta\{\sum_{ij} c_i^\lambda c_j^\lambda S_{ij}\} - \sum_{ij\lambda} E_\lambda c_i^\lambda c_j^\lambda \Delta S_{ij}.$$

The first term on the right hand side vanishes as Eqs. (18) and (34) show that the expression in braces is the total number of electrons M which is constant. Thus the force does not contain any derivatives in the wavefunction coefficients – as required by the Hellmann-Feynman theorem (Slater, 1960).

The term $\Delta\epsilon_{k,s}$ contains $\Delta\langle\tilde{n}_s\rangle_k$ and $\Delta\langle\tilde{n}_s^2\rangle_k$. These can be found from Eqs. (26) and (27):

$$\Delta\langle\tilde{n}_s\rangle_k = \frac{1}{I_k}\sum_l\{H_{kl}\Delta d_{l,s} + d_{l,s}\Delta H_{kl}\}$$

$$\Delta\langle\tilde{n}_s^2\rangle_k = \frac{1}{I_k}\sum_{lm}\{2u_{klm}d_{l,s}\Delta d_{m,s} + d_{l,s}d_{m,s}\Delta u_{klm}\}.$$

Terms involving the matrix elements T_{ij} and S_{ij} depend on R_{la} only through the basis functions $\phi_i(\mathbf{r} - \mathbf{R}_a)$, but the pseudopotential term has an additional dependence arising from $V_a^{ps}(\mathbf{r} - \mathbf{R}_a)$. This can be evaluated by integrating by parts:

$$\int \phi_i(\mathbf{r} - \mathbf{R}_i)\Delta V_a^{ps}(\mathbf{r} - \mathbf{R}_a)\phi_j(\mathbf{r} - \mathbf{R}_j)d\mathbf{r} = -\int\{\phi_j(\mathbf{r} - \mathbf{R}_j)\Delta\phi_i(\mathbf{r} - \mathbf{R}_i) + \phi_i(\mathbf{r} - \mathbf{R}_i)\Delta\phi_j(\mathbf{r} - \mathbf{R}_j)\}V_a^{ps}(\mathbf{r} - \mathbf{R}_a)d\mathbf{r}.$$

Despite the complexity of the equations, the time taken to evaluate the forces is small in comparison with that taken to determine the self-consistent energy.

VI. Structural Optimization

1. UNCONSTRAINED RELAXATION

Usually the positions of atoms in the starting cluster do not correspond to the equilibrium ones. The first step then is to determine the energy and forces acting on them but once this has been achieved, it is necessary to consider efficient algorithms that allow the equilibrium sites to be found. The optimization strategy that is often used is a conjugate gradient one (Press *et al.,* 1987). This requires only the forces to be known at any stage. The atoms are moved to a new set of positions whose energy is lower than that of the previous set. Suppose in some configuration the forces acting on atom a are f'_{la} in direction l. Then the atoms are moved along a conjugate direction d'_{la} so that the new atomic position is

$$R'_{la} = R_{la} + wd'_{la}.$$

Here w is chosen so that the free energy F in Eq. (36), is least. This is usually accomplished by quadratic or cubic interpolation. The directions d'_{la} are related to the forces f_{la} through

$$d'_{la} = f'_{la} - xd_{la},$$

where d_{la} is the previous search direction where the force was f_{la}.

The value of x is just:

$$x = \frac{\sum_{la} f'_{la}(f'_{la} - f_{la})}{\sum_{la} f^2_{la}},$$

and is set to zero initially.

The efficiency of the optimization strategy depends on the number of constraints. For an atom strongly bonded to at least three others in a non-planar configuration, relaxation is very fast as the atom is over-constrained. Thus we find about 10 iterations are required to relax the inner atoms of a tetrahedrally bonded cluster reducing the forces on each atom to less than 0.001 a.u. On the other hand, if the atoms have low coordination, then the structure is more floppy and the number of relaxations required is much greater. This can happen if the surface H atoms are allowed to move. In such cases, the movement of the H atoms sets up an elastic wave in the bulk whose reflections repeatedly affect the surface. This can be overcome by attaching springs to the H atoms simulating the outer crystal. The choice of spring constants is, however, somewhat arbitrary.

Other problems occur with clusters of water molecules where weak hydrogen bonds co-exist with strong covalent intramolecular O-H bonds. In this case, the choice of a single quantity w may not be the best strategy and it would be desirable to include information on the derivatives of the forces.

The most serious problem with the optimization strategy is that it finds a local minimum in the energy. There may be – and often are – other lower minima separated by barriers from the one found. The only way to reach these with the static relaxation method described here is to start the calculation from different structures but even then there can be no guarantee that the global minimum has been found. The global minimum, of course, may not really be of any physical interest. For example, the global minimum for a vacancy or dislocation in silicon corresponds to the defects lying on the surface of the cluster.

2. CONSTRAINED RELAXATION

For some purposes it is important to relax the cluster with constraints, as for example, in determining the saddle point for defect migration or reorientation. At the saddle point, there is at least one direction along which the energy falls when the atoms are displaced along it, while along other directions the energy rises. A commonly used procedure for finding the saddle point would be to average the atomic coordinates corresponding to the beginning and end points of the migration path and then to calculate the Hessian or energy second derivatives at this point. If the structure is close to the saddle point, the Hessian matrix has at least one negative eigenvalue. The eigenvector corresponding to this eigenvalue gives a direction, d_{la}, in which the energy decreases. The cluster is then relaxed so that the displaced coordinates, ΔR_{la}, lie orthogonal to this direction, *i.e.,*

$$\sum_{la} d_{la}\Delta R_{la} = 0.$$

The saddle point is then located by moving along the direction d_{la} so that the energy *increases* to a maximum. However, this procedure is not a practical one because of the time taken to evaluate the Hessian. Clearly, some constraints must be imposed on the coordinates otherwise the conjugate gradient algorithm would push the coordinates away from the saddle point. To deal with this problem, it is important to select a few variables for which the energy varies rapidly. These include the bond lengths nearest to the defect core. For example, in oxygen migration they would include the Si-O and Si-Si lengths nearest the defect. The relevant variables are held fixed while the remainder are allowed to vary, minimizing the energy. If variables that are not relevant are selected, for example the position of an atom outside the defect core or some angle, then the structure could slide from one configuration into the other, rapidly passing the barrier, which often manifests itself as a cusp. Again, it is important to be able to calculate the forces on all the atoms allowing for the constraint. The procedure that we have found successful is as follows.

Suppose that an atom a is hopping from one site to another during which one bond $a-b$ is broken and the bond $a-c$ is created. Then the constraint used involves the relative bond lengths, $|\mathbf{R}_a - \mathbf{R}_b|$ and $|\mathbf{R}_a - \mathbf{R}_c|$, and the cluster is relaxed maintaining this constraint. For technical reasons, the actual constraint used is:

$$x = (\mathbf{R}_a - \mathbf{R}_b)^2 - (\mathbf{R}_a - \mathbf{R}_c)^2 = constant. \tag{37}$$

This provides a linear equation for one of the Cartesian components, say l, of the 'central' atom a which can be solved for any value of x.

$$R_{la} = \frac{2\sum_{k\neq l}(R_{kb} - R_{kc})R_{ka} + \sum_k(R_{kb}^2 - R_{kc}^2) - x}{2(R_{lb} - R_{lc)}}. \tag{38}$$

Now, if the atoms are moved in any way, the change to the energy is

$$\Delta E = -\sum_{m,d} f_{md}\Delta R_{md}. \tag{39}$$

Now,

$$(\mathbf{R}_c - \mathbf{R}_b).\Delta\mathbf{R}_a = (\mathbf{R}_a - \mathbf{R}_b).\Delta\mathbf{R}_b - (\mathbf{R}_a - \mathbf{R}_c).\Delta\mathbf{R}_c,$$

and hence the term in ΔR_{la} in Eq. (39) can be written in terms of $\Delta R_{ka}, k \neq l$, $\Delta\mathbf{R}_b$ and $\Delta\mathbf{R}_c$ and hence be eliminated from the expression for ΔE. This modifies the forces on the atoms a, for $k \neq l$ as well as the atoms b and c, but these changes can now be easily found. The value of l is selected to give the greatest value of the denominator in Eq. (38). The analysis can be generalized to deal with several constraints simultaneously and the method has proved successful in dealing with a number of problems where bonds are switched between different configurations.

VII. Determination of vibrational modes

1. ENERGY SECOND DERIVATIVES AND MUSGRAVE POPLE POTENTIALS

The vibrational modes of clusters can be found from the dynamical matrix $E_{la,mb}$ as described in II 1. This is related to the second derivative of the energy with respect to displacements of the atom a at $\mathbf{R}_a$, along the Cartesian direction l and atom b at $\mathbf{R}_b$ along direction m. The cluster is first relaxed so that the forces on all the atoms, or those around the defect, are essentially zero. Then the atom a is displaced by ϵ (≈ 0.025 a.u.) along the l axis. The electrons will attempt to 'follow' this displacement so that that the self-consistent charge density will be different from that in the equilibrium configuration. This new self-consistent charge density must then be found. When this has been done, the forces on all the atoms of the cluster are evaluated and these will no longer be zero because of the change in charge density and structure. Suppose the force $f^+_{mb}(l,a)$ acts on the atom b in direction m. This adiabatic force includes the effect of the screening charge density seeking to oppose the change caused by moving the atom a. The whole process is now repeated by moving the atom a by -ϵ along the same direction l producing forces $f^-_{mb}(l,a)$. The energy second derivative is then, up to second order in ϵ,

$$D_{la,mb} = (f^+_{mb}(l,a) - f^-_{mb}(l,a))/2\epsilon.$$

It is important to realize that these are not infinitesimal derivatives. They include contributions from all even powers of ϵ and the frequencies that they give rise to contain anharmonic contributions. For this reason, the latter are sometimes called *quasi-harmonic* frequencies (Jones *et al.,* 1994b).

Only some of the entries of the dynamical matrix of a large cluster can be found in this way. This is because it is a very time consuming procedure to evaluate the second derivatives and those for atoms near the surface are irrelevant for the frequencies of vibration for bulk solids. The next step is to fit the calculated derivatives to those arising from a valence force potential. The potential can then be used to generate the dynamical matrix for any type of cluster or unit cell composed of the same elements and bonding configuration.

One could choose many types of potential but one that is particularly useful is due to Musgrave and Pople (1962). This includes all possible bond length and bond angle distortions up to second order. The potential for atom a is:

$$\begin{aligned} V_a = \quad & 1/4\sum_b k_r^{(a)}(\Delta r_{ab})^2 + r_0^2/2\sum_{b>c} k_\theta^{(a)}(\Delta\theta_{bac})^2 + r_0\sum_{c>b} k_{r\theta}^{(a)}(\Delta r_{ab}+\Delta r_{ac})\Delta\theta_{bac} \\ & + \sum_{c>b} k_{rr}^{(a)}\Delta r_{ab}\Delta r_{ac} + r_0^2\sum_{d>c>b} k_{\theta\theta}^{(a)}\Delta\theta_{bac}\Delta\theta_{cad}. \end{aligned}$$

Here Δr_{ab} and $\Delta\theta_{bac}$ are the changes in the length of the $a-b$ bond and angle between the $a-b$ and $a-c$ bonds, respectively. The sums are over the nearest neighbor atoms of atom a. The Musgrave-Pople potential is superior to a Keating potential, for example, since it includes correlations between bond stretch and bend.

This potential can be used to derive phonon dispersion curves for the bulk solid. This has been done in a number of cases such as diamond (Jones, 1988), Si (Jones, 1987), Ge (Berg Rasmussen *et al.,* 1994), GaAs (Jones and Öberg, 1991a), AlAs (Jones and Öberg, 1994a), InP (Ewels *et al.,* 1996a) and quartz (Purton *et al.,* 1992). The potential usually gives the highest frequencies accurate to a few wave-numbers, although it cannot account for the splitting of the longitudinal and transverse optic modes due to a long range Coulomb field. The worst errors occur at low frequencies where the assumption that forces beyond second shell atoms are zero is inadequate. Moreover, the elastic constants derived from the potential are in poor agreement with experiment. Nevertheless, their main use is in describing the local modes of defects and the disagreement arising at low frequencies is not important.

Let us now consider how the local and resonant modes of a defect are calculated. The second derivatives of the cluster containing the defect are found in the same way as described above. These derivatives are usually evaluated between the defect atoms and their nearest neighbors. Other entries in the dynamical matrix are then found from the Musgrave-Pople potential. The normal modes and their frequencies are then found by direct diagonalization of the dynamical matrix with the masses of the terminating H atoms set to infinity.

This procedure works well for frequencies well away from the one phonon spectrum. Usually infra-red absorption spectroscopy is only able to detect local modes and for such problems this method is satisfactory. However, there are cases where it is the resonant modes that are of interest as, for example, nitrogen related defects in diamond. We shall discuss their evaluation below. Usually, the number of modes calculated for a defect exceeds the number observed. The remaining ones are not detected for a number of reasons. They might fall into a part of the spectrum where there is strong absorption arising from an overtone or combination band from the bulk or substrate; their life-times might be very short due to anharmonic interactions and this implies a very broad spectrum as, for example, the highest modes of the interstitial carbon dimer, C_i-C_s, in Si (Jones *et al.,* 1995a; Leary *et al.,* 1996) or finally the defect may possess a very small transition dipole moment. It is this last issue that can be addressed with the method.

2. THE EFFECTIVE CHARGES

Leigh and Szigeti (1967) give the integrated intensity of absorption due to a defect as

$$\frac{2\pi^2\rho}{ncM'}\eta^2.$$

Here η is the effective charge, c is the velocity of light, n the refractive index of the material, M' and ρ are the mass and concentration of the impurity respectively. η^2 is given by the sum over any degenerate modes of

$$M'\left(\frac{\partial M_x}{\partial Q_i}\right)^2, \tag{40}$$

where M_x is the dipole moment in the direction of the polarization of the electromagnetic field. Q_i is the normal coordinate of the mode i. That is, the displacement of each atom is $Q_i u^i_{la}/\sqrt{M_a}$. The induced dipole can be evaluated from the changes in the dipole moment of the cluster,

$$\mathbf{M} = \sum_a Z_a \mathbf{R}_a - \int \mathbf{r} n(\mathbf{r}) d\mathbf{r},$$

when the atoms are subjected to this displacement.

There are several points to note here. The effective charge depends on the mode and its displacement pattern. It can be very different for different modes of the same defect. This is illustrated by the H wag mode of the passivated C acceptor in GaAs (Jones and Öberg, 1991a). Although the effective charge for the stretch mode is almost unity (Kozuch *et al.,* 1990), that of the wag mode is almost zero as it was originally undetected by infra-red absorption measurements but had been observed by Raman scattering (Wagner *et al.,* 1995).

The effective charge in general depends on the polarization of the electrical field and the orientation of the defect. If thermal equilibrium prevails leading to defects assuming all possible orientations which have degenerate energies, then η^2 must be averaged over these orientations. The important case of a defect with C_{3v} symmetry in cubic crystals has been considered by Clerjaud and Côte (1992) who showed that the average effective charges are given by:

$$\eta^2_{A_1} = \frac{M'}{3}\sum_l \Big(\frac{\partial M_l}{\partial Q_{A_1}}\Big)^2,$$

for the A_1 non-degenerate mode and

$$\eta^2_E = \frac{2M'}{3}\sum_l \Big(\frac{\partial M_l}{\partial Q_E}\Big)^2,$$

for the two-fold degenerate E-mode. These effective charges are independent of the direction of the polarization of light. The assumption of equal numbers of point defects in different orientations related by symmetry is not always valid, even in the absence of stress, as for example the case of single passivated substitutional C dimer in GaAs. Here, the defects form during growth on the surface and are frozen in during cooling. The $\mathrm{C_{As}}$-Ga-$\mathrm{C_{As}}$ unit is oriented along only one of the two possible $\langle 110\rangle$ orientations for a (001) growth plane (Cheng *et al.,* 1994; Davidson *et al.,* 1994).

The effective charge is independent of the mass M' in only simple cases such as a H stretch and wag mode. For in these cases, essentially only the H atom is undergoing a displacement and the derivative of $M_x(R_{la} + Q_i u^i_{la}/\sqrt{M_a})$ scales as $1/\sqrt{M'}$. In this case η is the same for H as for D and the integrated intensity is then a factor two smaller for D than for H. Of course, this argument neglects anharmonic effects which are more important for H.

3. RESONANT MODES

A Green function method has been developed to analyze these modes. The Green function for the bulk crystal can be evaluated directly from the dynamical matrix constructed from the Musgrave-Pople potential. If the atoms in the unit cell are located at $\mathbf{R}_{\tau_a}$ then the Green function for a vector $\mathbf{k}$ in the Brillouin zone is given by:

$$G^0_{l\tau_a,m\tau_b}(\mathbf{k}) = \sum_i \frac{u^i_{l\tau_a}(\mathbf{k})u^i_{m\tau_b}(\mathbf{k})}{\omega^2 - \omega_i^2(\mathbf{k})}.$$

We can find the crystal Green function in real space between atoms a and b related to basis atoms by

$$\mathbf{R}_a = \mathbf{R}_{\tau_a} + \mathbf{R}_L, \quad \mathbf{R}_b = \mathbf{R}_{\tau_b} + \mathbf{R}_M$$
$$G^0_{la,mb} = \frac{1}{\Omega}\sum_{\mathbf{k}} e^{i\mathbf{k}.(\mathbf{R}_L - \mathbf{R}_M)} G^0_{l\tau_a,m\tau_b}(\mathbf{k}).$$

The Green function in the presence of a defect, whose contribution to the dynamical matrix differs by $V_{la,mb}$, is:

$$G_{la,mb} = G^0_{la,mb} + \sum_{nc,pd} G^0_{la,nc} V_{nc,pd} G_{pd,mb}.$$

This equation can be readily solved for a point defect since the elements of $V_{la,mb}$ are taken to be zero for either a or b outside the second shell of neighbors surrounding the defect. The density of phonon states projected onto an atom a is then found from the trace of the imaginary part of this Green function *i.e.*,

$$-\frac{2\omega}{\pi}\Im.\sum_l G_{la,la}.$$

In this way the contribution of the resonance to the local density of states is found. However, the infra-red absorption is controlled by the induced dipole moment, and it is better to evaluate this assuming reasonable values of the charge distributed over atoms of the defect. If the charge q_a is located on atom a, then the induced dipole is

$$M_l = -\frac{2\omega}{\pi}\Im.\sum_a q_a G_{la,la}.$$

This procedure has been used to discuss the absorption of resonant modes due to complex N defects in diamond (Jones *et al.,* 1992b). In this case, 27000 points in the Brillouin zone were used to construct G^0.

VIII. Practical considerations

1. CHOICE OF BASIS SETS

There are two different basis sets used in the method. The first is a basis used to describe the wavefunctions. This is invariably a set of Gaussian functions defined by an exponent a_i and sited on an atom or at the center of a bond or some other location, R_i. As described in IV.1, this Gaussian is multiplied by a polynomial in $x - R_{xi}, y - R_{yi}, z - R_{zi}$. For spherically symmetric s-functions, the polynomial is trivially unity. For p-orbitals, the three possible polynomials are $x - R_{xi}, y - R_{yi}$ or $z - R_{zi}$. For d orbitals, all 6 polynomials of degree 2 are used, which generates a linear combination of five d- and one s-orbitals. The code also includes f-orbitals generated by polynomials of degree 3. The complete basis is then a linear superposition of these orbitals for different exponents a_i and centers $\mathbf{R}_i$.

The use of bond centered orbitals seems unique to the AIMPRO code. They serve a useful function of representing the pseudo-wavefunction in the region where it is often large. They make it unnecessary to use atom centered d-orbitals for Si, Ge and GaAs. They are particularly important in dealing with some impurities like oxygen in silicon where the strong strain effects on atoms distant from the impurity make it necessary to use a basis which gives as accurate as possible elastic constants for the host. For other materials, as for example diamond, relaxation effects are often very small and bond centered orbitals have very little effect. Although the location of the bond centers is often kept fixed at a specified point (e.g. the mid–point) of a bond, it is possible to allow them to relax or float with the atoms of the cluster until the minimum energy is found. This has not been commonly used since they often move close to atoms leading to instabilities.

The optimum exponents $a_i, i = 1, 2...m$, for a particular atom can be found by minimizing the energy of the pseudoatom as a function of a_i. This procedure also generates the coefficients of the wavefunction: $c_i^\lambda, i = 1, 2, ..., m$ for each valence state λ. For example, it generates a set of coefficients for an s-orbital and a set for the three p-orbitals. When an application is made to a large cluster, the same fixed linear combination of the Gaussian orbitals with different exponents can then be used. This gives a basis of 4 orbitals for each Si atom for example and 10 for a transition element like Ni. Such a basis is called a minimum one. In many applications, the minimum atomic basis is used for atoms far away from the core. For other atoms, the coefficients which multiply the Gaussian orbitals are treated as variational parameters as described in IV.

A second basis is used to expand the charge density. This again is a set of Gaussian functions, or modified Gaussian functions as described in IV, defined by an exponent b_k and center $\mathbf{R}_{fk}$. Again, the centers $\mathbf{R}_{fk}$ can be chosen to lie at nuclei, bond-centers or other locations. The optimum basis consists of exponents and sites which *maximize* the estimated Hartree energy $\tilde{E}_H$ as described in IV.1.

It is always desirable to locate the Gaussian orbitals at a symmetrical site or the set of sites generated by symmetry, since otherwise the energy levels, vibrational modes *etc.*, will not possess the required degeneracy. It is expedient to define the basis in terms of $N - M\ X$ which means that a basis of N Gaussian s, p or d orbitals are placed at the location of each atom of type X to describe the wavefunctions, while a basis of M Gaussian s-functions are used to describe the charge density. In addition the sites treated in terms of a minimal basis set need to be defined as well as any orbitals and fitting functions placed at bond centers. A minimal basis is often placed on the surface H atoms.

One basis set which has been of occasional use is an icosahedral set of 'bond centers' sited close to an atom. This has an advantage that the d and f degeneracy of an atom is not compromised but in general the point group symmetry of a defect will be lost.

The basis size has a significant effect on calculated properties: with structures being least sensitive and energies and wavefunctions being most sensitive. It is not possible to converge total energies with the same degree of rigor as is occasionally obtained in plane-wave treatments. This is because simply increasing the number of exponents used to describe the basis eventually results in a numerical instability for the Choleski decomposition of the overlap matrix. However, in practice it is energy *differences* that are important as, for example, between a H atom at a bond centered and tetrahedral interstitial site. In this case the dependence of the total energy difference can be easily checked.

2. THE CONSTRUCTION OF A SUITABLE CLUSTER

In dealing with defects within semiconductors, H terminated clusters have invariably been used. These saturate the dangling bonds at the surface of the cluster leading to widely separated filled and empty surface states for 'bulk' clusters, *i.e.* clusters comprised with the same stoichiometry and atomic arrangement as the bulk semiconductor. If the surface H bond lengths are close to their equilibrium values, the band gaps are much greater than those of the bulk solids, with the exception of diamond. Values for representative clusters are given in Table 2. These were calculated for tetrahedral clusters with an 8-8 basis on the inner 5 atoms and a minimal basis on all the others. Two bond centered Gaussian basis functions with different exponents were sited on all the bond centers between host atoms.

These gaps are much larger than those found using density functional theory in supercells which are in turn smaller than experimental gaps. The cluster band gaps vary only slightly with the basis size but become smaller if longer H bonds are allowed. It is not advisable to use long H bonds as this imposes a strain on inner bonds around defects and this can certainly modify their structure, seriously perturbing the local vibrational modes. The band gap also decreases slowly with cluster size.

Despite the large band gaps, some information on the position of energy levels can be obtained. There are two common ways of 'correcting' the band gap to make allowance for the difference with experiment. The first is to simply scale defect levels by the band gap. Clearly this is simply pushing both valence and conduction band states closer together. The second is to use a 'scissors' operator. This is added to the Hamiltonian and displaces the unoccupied states of the 'perfect' cluster upwards by V. The scissor operator is

$$\Delta(r, r') = V \sum_{\lambda'} \psi_{\lambda'}(r)\psi_{\lambda'}(r'),$$

where the sum is over unoccupied levels. It can also be expressed in terms of the occupied

Table 2: Lowest Kohn-Sham level, E_1, highest occupied level E_v, calculated band gap E_g and experimental gap for various clusters in eV.

	Cluster Size		E_1	E_v	E_g	Exptal. gap
Diamond	71	$C_{35}H_{36}$	-23.42	-6.09	5.01	5.5
	131	$C_{71}H_{60}$	-23.69	-5.36	5.92	
	297	$C_{181}H_{116}$	-24.17	-4.70	5.37	
Silicon	71	$Si_{35}H_{36}$	-16.31	-6.48	3.82	1.17
	131	$Si_{71}H_{60}$	-16.77	-6.41	3.13	
	297	$Si_{181}H_{116}$	-16.91	-5.96	2.51	
Germanium	71	$Ge_{35}H_{36}$	-16.70	-6.12	3.53	0.75
	131	$Ge_{71}H_{60}$	-16.92	-5.93	2.70	
	297	$Ge_{181}H_{116}$	-17.25	-5.58	2.14	
Gallium Arsenide	71	$(Ga_{19}As_{16}H_{36})^{3-}$	-10.91	0.31	2.95	1.42
	131	$(Ga_{31}As_{40}H_{60})^{9+}$	-34.92	-22.59	2.27	
	297	$(Ga_{89}As_{92}H_{116})^{3+}$	-22.76	-10.10	1.92	
	71	$(As_{19}Ga_{16}H_{36})^{3+}$	-24.92	-13.13	2.83	
	131	$(As_{31}Ga_{40}H_{60})^{9-}$	-1.47	10.16	2.64	
	297	$(As_{89}Ga_{92}H_{116})^{3-}$	-14.24	-1.752	2.09	

states and, for the spin-averaged case,

$$\Delta(\mathbf{r}, \mathbf{r}') = V \sum_{ij} (\delta_{i,j} - b_{ij}) \phi_i(\mathbf{r} - \mathbf{R}_i) \phi_j(\mathbf{r}' - \mathbf{R}_j).$$

V is chosen to give the correct band gap. This is then applied to a cluster containing a defect. Few calculations have been carried out using this method.

Energy differences between different defect states can be found using the Slater transition method (Slater, 1960). Here, the difference in total energies of the configurations where an electron is promoted from the λ to the μ orbital is found from the eigenvalues of the configuration corresponding to half of the promoted electron being placed in each orbital. Then the total energy difference is accurately given by $E_\mu - E_\lambda$. This differs from the zero-phonon line of the optical transition by the relaxation energy of the defect. This method has been used to treat optical transitions for vacancy impurity complexes in diamond (Goss *et al.,* 1996a).

The size of the cluster used varies with the application and for point defects at least one, and usually two shells of host atoms surround the defect. In the compound semiconductors such as GaAs, one can choose between stoichiometric clusters which contain as many Ga as As atoms and are chemically neutral, or others possessing a greater number of As than Ga atoms for example. In both cases, for clusters representing bulk material it is important to occupy them with M electrons where M is twice the number of covalent bonds. This results in all the bonding states being filled and the anti-bonding ones being empty. Again there is a large gap between the two. Such a procedure results in bond lengths close to experimental values. If the cluster is not stoichiometric then this procedure necessarily leads to charged clusters. This arises as the number of protons is determined by the numbers of As, Ga and terminating H atoms but it is the topology that determines the number of electrons and clearly for four-fold coordinated atoms, the number of electrons is exactly the same as if the cluster was made from Si atoms and terminated by H. The charged clusters do not appear to cause serious problems. It is especially desirable to use charged non-stoichiometric clusters when the defect has a high symmetry and consequently has degenerate energy levels or vibrational modes, *e.g.* a C atom substituting for As in AlAs. A disadvantage with charged clusters is that the energy levels are shifted up for negative and down for positively charged clusters and the position of a defect level with respect to a band edge becomes more difficult to assess. For this reason a stoichiometric cluster is to be preferred. However, this is not free from difficulty as then the cluster has more As atoms in its upper half than Al ones, for example, and leads to a dipole moment. This has an effect on the bond lengths parallel and perpendicular to this direction. For example, an 86 atom stoichiometric cluster, $Al_{22}As_{22}H_{42}$, has C_{3v} symmetry and is centered on the middle of an Al-As bond. It has a dipole moment along the C_3 axis leading to Al-As bond length of 2.476 Å compared with 2.428 Å for the other six bonds (Jones and Öberg, 1994a). These are all within 2% of the experimental value of 2.43 Å.

3. MULLIKEN POPULATIONS

In many cases it is important to understand the nature of gap states and the hybridization state they refer to. For example, in the C_i defect in Si, the gap states are localized on *p*-orbitals on the C and Si atom sharing a lattice site (Leary *et al.,* 1996). The simplest way of determining the character of a state is to plot its wavefunction. However, the use of pseudopotentials implies that the amplitude is invariably small near a nucleus. It is not then easy to deduce which atoms gap states arise from. One way which gives some information is to evaluate the Mulliken bond populations m_i^λ. These are defined from the integral of the square of the wavefunction. Eqs. (17) and (31) show:

$$\sum_s \int \psi_\lambda^2(\mathbf{r}, s) d\mathbf{r} = \sum_{ij} c_i^\lambda c_j^\lambda S_{ij} = \sum_i m_i^\lambda$$

where

$$m_i^\lambda = c_i^\lambda \sum_j S_{ij} c_j^\lambda.$$

If the state is localized on an atom a, then c_i^λ and hence m_i^λ will be large for basis functions i localized there. One problem is that m_i^λ can be large and negative because the phases of Gaussian orbitals with different exponents, but centered on atom a, are rarely the same.

4. RADIATIVE LIFETIMES

For complicated defects, as for example Ni in diamond, there are many gap levels whereas only one or two optical transitions related to the defect have been observed. There is then a problem in assigning the transition. The calculated radiative lifetimes of the various transitions can be very different and the most intense transition will be associated with the smallest lifetime. It is then necessary to calculate this quantity.

The rate of electrical dipole transitions between two states λ and μ can be found using the expression (Svelto, 1976):

$$\frac{1}{\tau_{\lambda\mu}} = \frac{4n\omega^3}{3c^3\hbar} \frac{e^2}{4\pi\epsilon_o} |\mathbf{r}_{\lambda\mu}|^2, \tag{41}$$

where

$$\mathbf{r}_{\lambda\mu} = \sum_s \int \psi_\lambda(\mathbf{r}, s) \mathbf{r} \psi_\mu(\mathbf{r}, s) d\mathbf{r},$$

is the dipole matrix element, n the refractive index, e the electron charge, c the speed of light and ω the transition frequency. Estimates of the radiative lifetime are sensitive to the transition energy and spatial extent of the wavefunction. We shall describe an application to the Si-V center in diamond in the next section.

IX. Applications

1. GENERAL

Many applications of the formalism have been made to molecules as well as point, line and surface defects in large clusters simulating bulk material. Investigations into molecular entities such as fullerenes (Estreicher *et al.,* 1992; Eggen *et al.,* 1996), water dimers and octamers (Heggie *et al.,* 1996b) and surface problems associated with the growth of CVD diamond (Latham *et al.,* 1994) will not be reviewed here. We shall instead emphasize some of the applications to point and line defects in bulk solids that have brought about a deeper insight into the properties of the defect.

2. POINT DEFECTS IN BULK SOLIDS

a) Diamond

Nitrogen is one of the most important impurities in diamond occurring in concentrations as large as 10^{20} cm^{-3} (Evans *et al.,* 1981). It readily complexes with itself and with other impurities and intrinsic defects and the resulting complexes are often important optical centers. The high solubility of N is attributed to the low misfit energy of inserting an N atom with N-C bond lengths of 1.47 Å into the diamond lattice where the C-C bonds are 1.54 Å. The neutral substitutional defect exhibits trigonal symmetry as convincingly demonstrated by electron paramagnetic resonance, EPR, (Smith *et al.,* 1959; Cook *et al.,* 1966). In type Ib or synthetic diamonds, N_s is present as an isolated defect, but in annealed synthetic diamonds, the nitrogen aggregates to give complexes with more than one N atom. These complexes are also found in the great majority of natural (type Ia) diamonds. It is a long standing problem to elucidate the final fate of N aggregation in diamond when it is annealed for long periods. The *ab initio* calculations have helped to clarify the properties of many of these nitrogen complexes.

The substitutional defect was investigated by Briddon *et al.,* (1991) and led to an explanation of the 'anomalous' vibrational mode associated with the defect. This local mode at 1344 cm^{-1} was observed (Collins *et al.,* 1982) in an infra-red absorption study on type Ib diamonds and its intensity correlated with the EPR signal due to N_s suggesting that it is associated with the vibrations of N_s. Surprisingly, however, the mode does not shift with ^{15}N doping. The cluster calculation revealed that for the neutral substitutional defect, not only N was displaced from a lattice site along [111] by 0.2 Å, but also the neighboring C atom was displaced along $[\bar{1}\bar{1}\bar{1}]$ by the same amount thus leading to back C-C bonds about 5% shorter than the normal C-C bonds. This was independently found by plane wave pseudopotential calculations (Kajihara *et al.,* 1991). There are two gap-levels of a_1 symmetry: a bonding state between N and the unique C atom and an anti-bonding state containing one electron. The presence of two a_1 levels is consistent with stress studies on the 4.059 eV optical center associated with N_s (Koppitz *et al.,* 1986; Vaz *et al.,* 1987). The absence of degenerate levels for the metastable T_d defect suggests that the mechanism for the off-site distortion is a chemical rebonding one rather than a Jahn-Teller distortion (Bachelet *et al.,* 1981). The vibrational modes of the defect were found from the Green function method discussed in VII 3. This gave three bands centered at 1320, 1122 and 1032 cm^{-1} in good agreement with observed ones at 1344, 1130 and 1080 cm^{-1} (Collins *et al.,* 1982). The highest mode was localized on the unique C atom and its C neighbors and does not shift with a change in the N isotope. This explains the anomalous mode. The N related modes fell below the Raman frequency at 1332 cm^{-1}. It is not surprising with hindsight to understand the character of the C related mode arising from the sp^2 bonding of the unique C atom. Recently, the reorientation energy of the defect has also been calculated

(Breuer and Briddon, 1996) and found to be 0.7 eV — the same as that observed experimentally by Ammerlaan and Burgemeister (1981).

The N_s defect is not stable during prolonged annealing at high temperatures and aggregates firstly into A centers, which are N_s dimers (Davies, 1976) and secondly into B centers which are believed to be vacancies surrounded by four N atoms (Loubser and Van Wyk, 1981). The calculated (Jones *et al.,* 1992b) vibrational modes of the defects are in reasonable agreement with observation giving further support to the assignments. Furthermore, they give a clue as to why N atoms should aggregate. The highest filled level in the A center is around mid-gap which is considerably lower than that of the N_s donor. Thus the driving force for aggregation is the lowering of the one-electron energy. It is an insight such as this which makes the theory so useful. The B defect also has deep mid-gap states which should make it optically active and this has prompted investigations into vacancies and vacancy-impurity defects which are also known to be very important optical centers in diamond.

Vacancies and interstitials were investigated by Breuer and Briddon (1995) who confirmed the importance of many-body effects and the need to determine the energies of different multiplets. The theory found that V^- was a spin 3/2 defect, in agreement with experiment, and the calculated optical transition energy agreed well with the observed value. However, the Von Barth procedure, discussed in section 5, yields too few equations to find the multiplets for the neutral vacancy and hence the theory is unable to describe this important case. In many cases, vacancies will complex with impurities and a recent study of N-V and Si-V centers (Goss *et al.,* 1996a) concluded that they possess very different structures. Whereas the N-V defect has C_{3v} symmetry, Si-V possesses D_{3d} symmetry where the Si atom sits mid-way between two adjacent vacancies. This finding explains the surprising optical properties of the defect. The dangling bonds on each of the two sets of three C atoms nearest Si form a_1- and e-states. The two sets of e-states combine to form bonding and anti-bonding e-levels around mid-gap, the highest of which is occupied by two electrons. Now, in synthetic or type Ib diamonds, the higher e-level traps an additional electron from N donors leading to a ^2E ground state. An internal optical transition can then take place with an electron promoted from the lower e-level. This $^2E \rightarrow {}^2E$ transition leads to four close-by luminescence lines if the e-levels are split by a Jahn-Teller effect, or possibly a spin-orbit interaction. The presence of three isotopes of Si causes a slight shift in the zero-point energy and leads to the appearance of a remarkable twelve line spectrum (Clark *et al.,* 1995). The computed radiative lifetime of 3 ns is in very good agreement with the experimental values around 1-4 ns (Sternschulte *et al.,* 1994). This example shows the method is able to explain very simply a complicated optical spectrum which would otherwise be difficult to understand.

b) Silicon

A great deal of effort has been devoted to understanding the properties of the light impurities: H, B, C, O and N in silicon, and especially their vibrational frequencies as local mode spectroscopy has been such a valuable experimental tool. Calculations have been made for the various substitutional defects: VH_n (Bech Nielsen *et al.,* 1995), B_s, C_s (Jones and Öberg, 1992d), VO_n (Ewels *et al.,* 1995), N_s (Jones *et al.,* 1994c) with substantial agreement obtained with experimental results for the local vibrational modes in each case. In addition, interstitial defects such as H (Jones, 1991b), C_i (Jones *et al.,* 1995a; Leary *et al.,* 1996), O_i (Jones *et al.,* 1992c), and N_i (Jones *et al.,* 1994c), and complexes of these have also been investigated.

One example is the C_i-O_i defect which turned out to have very surprising structure. The C interstitial on its own takes the form of a [100] oriented split-interstitial as shown in Fig. 3 (Watkins, 1964; Zheng *et al.,* 1994). Two of the C-Si bonds along [011] are 1.8 Å long and pull

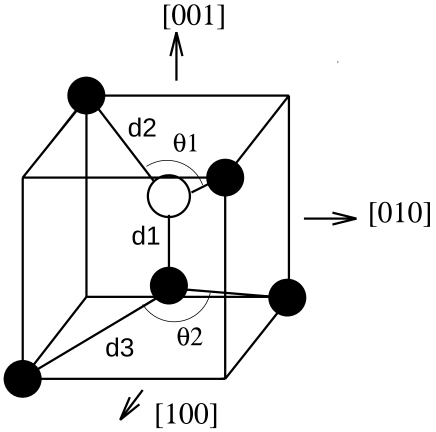

Figure 3: The C_i defect.

in the Si neighbors lying there. This leaves the Si-Si bonds along this direction extended and hence are favorable sites for attack by oxygen (Trombetta and Watkins, 1987). The calculation (Jones and Öberg, 1992a) shows that O does not lie at a bond center site within these dilated bonds but rather moves towards the Si radical shown in Fig. 3. The reason for this is the electronegativity of C exceeds Si rendering the Si radical positively charged. This in turn attracts the O atom so that its becomes over-coordinated leading to rather long Si-O bonds. The three Si-O bonds are by no means equal in strength. A consequence is that the O-related vibrational mode lies well below that of interstitial and even substitutional oxygen. The same process occurs for interstitial N. But now the state arising from the dangling bond on the Si_3 atom is occupied. This has led to a remarkable finding (Ewels *et al.,* 1996b): the O atom, being negatively charged, squeezes itself into dilated Si-Si bonds adjacent to N and pushes up the donor level due to Si_3. For N_i-O_2, the level is displaced almost to the conduction band. This defect might explain the occurrence of shallow thermal donors which arise when Czochralski Si, containing N, is annealed to 650°C (Suezawa *et al.,* 1986).

This 'wonderbra' mechanism of deep to shallow level conversion is not unique to N. A shallow donor level also arises when a C-H unit replaces N. Of course, this begs the question as to whether an interstitial C-H defect with this structure is stable at these temperatures. But this unit is known to be a constituent of the T-center, which is stable to 600°C (Minaev and Mudryi, 1981). This photoluminescent center has an emission band at 0.9351 eV and a rich spectrum of local mode satellites. The isotope shifts of these local modes with ^{13}C and D have recently been calculated and agree very well with the observed ones (Safonov *et al.,* 1996). The defect has a gap level, occupied by a single electron and it would be interesting to know if this level is displaced upwards becoming a shallow donor if O atoms cluster around this defect.

The N interstitial referred to above is not the dominant N defect in Si. This consists of a close-by pair of [100] split-interstitials. The evidence comes partly from channeling experiments showing that each N atom is displaced about 1 Å from lattice sites; partly from infra-red spectroscopy showing that the N atoms are equivalent and the high frequencies are due to an interstitial complex; and partly from the theoretical modeling (Jones *et al.,* 1994c). The model refutes earlier suggestions that nitrogen forms molecules within silicon. A complex of the N pair with oxygen yielding an electrically inactive NNO defect has also been investigated both experimentally and theoretically (Jones *et al.,* 1994d; Berg Rasmussen *et al.,* 1995).

c) Compound semiconductors

The cluster calculations were the first to describe and detail the structure and modes of H passivated Si donors and Be acceptors in GaAs (Briddon and Jones, 1990). Subsequent studies of trigonal C-H complexes in GaAs have been particularly fruitful. The local modes of the defect (Jones and Öberg, 1991a) exhibit several unusual properties. The E^--mode, which involves a movement of H perpendicular to the C_3 axis, and out of phase with C, was placed around 715 cm^{-1}. The C related A_1 and E^+ modes, which involve motion of H in phase with C in respective directions parallel and perpendicular to the C-H bond, were calculated to lie at 413 and 380 cm^{-1} respectively. Infra-red spectroscopy on GaAs containing high concentrations of C and H grown by molecular beam epitaxy and chemical vapor deposition methods located modes at 453 (X) and 563 cm^{-1} (Y) (Woodhouse *et al.,* 1991). Both were subsequently shown to be due to the C-H defect as they exhibited shifts with C and H isotopes. A Raman scattering experiment (Wagner *et al.,* 1991) assigned the 453 cm^{-1} mode to C-A_1. Y is now believed to be the E^+ mode (Davidson *et al.,* 1993). The E^- mode was not observed in these early experiments. However, in deuterated samples, the E^- mode was detected at 637 cm^{-1}. This must imply that the unobserved E^- mode in the H samples lies above 637 cm^{-1} and a simple force constant model (Davidson *et al.,* 1993) predicted it to lie at 745.2 cm^{-1}. The failure of the early infra-red experiments to locate the H-E^- mode was explained by the *ab initio* cluster theory as the consequence of a small transition dipole moment. Very recently this mode has been detected at 739 cm^{-1} by Raman scattering experiments (Wagner *et al.,* 1995). Similar calculations have been carried out for C in AlAs (Jones and Öberg, 1994a). The effect of anharmonicity on the stretch mode has also been investigated (Jones, *et al.,* 1994b) .

It is possible to grow heavily C doped GaAs by chemical beam epitaxy using CBr_4 as a doping source so that the films are free of hydrogen. C is a very electronegative element and naturally favors an As site. The calculations (Jones and Öberg, 1994a) show a large build up of charge around C and there is no evidence that C can occupy Ga or Al sites and behave as a donor. However, when C-rich samples are annealed at 850°C, there is a loss of C from As sites together with a reduction in the hole density. It was first suggested by Jones and Öberg (1994e), and independently by Cheong and Chang (1994), that rather than C_{Ga} defects being formed, a [100] oriented C-C dimer located at an As site, and which acts as a single donor, could be created. Thus for every pair of C atoms lost from As sites, there would be a loss of three holes. The computed stretch frequency of the C-C dimer in GaAs is 1799 cm^{-1} and is Raman but not infra-red active. Recently, *two* dimers have been detected by Raman scattering (Wagner *et al.,* 1996) with modes at 1742 and 1858 cm^{-1} in the annealed material. The hole density is about 10% of its pre-annealed value, 2.5 $\times 10^{20}$ cm^{-3}, but the concentration of C_{As} dropped to 30% of its pre-annealed value, also about 2.5 $\times 10^{20}$ cm^{-3}. Hence some donors or hole traps must have been introduced by the annealing. If only $(\text{C-C})_{\text{As}}$ dimers were introduced by the annealing, then we would require their concentration to be 5×10^{19} cm^{-3} to account for the carrier density, but that would result in about 8×10^{19} cm^{-3} of carbon unaccounted for. Presumably this forms the second type of dimer which possibly lies at an interstitial T_d site and acts as a single acceptor.

Other types of dimers are present in epitaxially deposited material containing H. A pair of substitutional C atoms at neighboring As sites traps H or H_2 with modes slightly shifted from those of C_{As}-H. An unexpected finding is that the dimers are preferentially oriented along one of the two [011] directions perpendicular to the (100) growth surface probably because of kinetic reasons (Cheng *et al.,* 1993; Davidson *et al.,* 1994). Recent calculations (Goss *et al.,* 1996b) confirm that H lies at a site accounting for the observed polarization in the C_2-H defect but in the case of C_2-H_2 the theory predicts one mode polarized along [011] and another – the

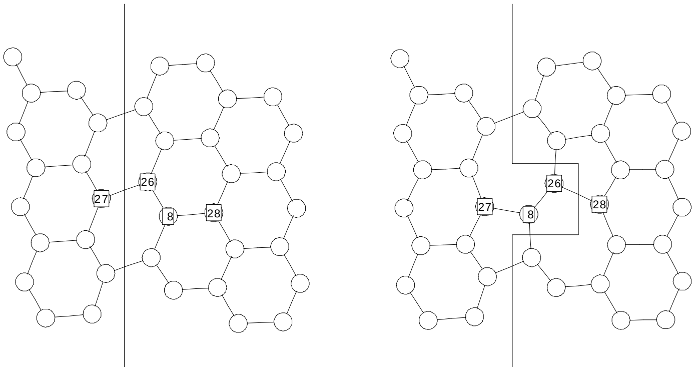

Figure 4: The reconstructed 90° dislocation and double kink in silicon. Vertical axis is $[01\bar{1}]$, horizontal axis is $[\bar{2}11]$.

higher mode – along $[0\bar{1}1]$. This has not yet been observed.

The C-H defect is unusual in possessing a resonant electron trap which has profound consequences for the dissociation of the defect. The calculated activation energy for dissociation is about 1 eV lower in the presence of minority carriers which can be trapped in the resonant level (Breuer *et al.,* 1996). This calculation anticipated experimental results (Fushimi and Wada, 1996) confirming this reduction in the activation energy.

3. LINE DEFECTS

In addition to the work carried out on point defects there has been a considerable attempt to understand the structure and kinetics of dislocations in group IV and III-V semiconductors. In these materials dislocations are dissociated into partials separated by a stacking fault. Commonly occurring partials are 90° and 30° ones. The cluster theory was the first *ab initio* one to reveal that 90° degree partial dislocations in Si (Heggie *et al.,* 1991) and GaAs (Öberg *et al.,* 1995) are reconstructed as shown in Fig. 4. The reconstruction leads to electrical inactivity of the line and is to be contrasted with earlier models of deep states arising from a line of dangling bonds. Intriguingly, impurities like P and N have a pronounced effect on the reconstruction in Si and actually break it (Heggie *et al.,*, 1993; Sitch *et al.,*, 1994). This effect might explain the very strong locking effect of these impurities – especially N – which has important technological implications.

An important question concerns the mobility of dislocations as this controls their rate of growth and ultimately their density in the crystal. This is especially important as dislocations bind point defects like vacancies and interstitials as well as impurities, all of which possess deep gap levels which can greatly affect the electronic and optical properties of the material. Now, it is believed that dislocations propagate by creating double kinks as shown in Fig. 4 which then expand under the influence of stress leading to motion of the dislocation. The energetics of this process can be followed by embedding the dislocation in a cluster. The kink formation energy was found (Öberg *et al.,* 1995) to be a very small value in these materials: about 0.1 eV, whereas the activation energy necessary to break the reconstructed bonds was considerable. The total activation energy for dislocation motion was found to be 1.9 eV in Si and 1.4 and 0.8 eV for β- and α-partials respectively in GaAs. These energies are in fair agreement with observations: 2.1 eV in Si, (Imai and Sumino, 1983) and 1.24 -1.57 eV for β-, and 0.89-1.3 eV for α-partials in GaAs (Matsui and Yokoyama, 1986; Yonenaga and Sumino, 1989). Moreover, these activation energies are sensitive to the Fermi-level. This arises because during the transition to the saddle point structure, a level moves from close to the band edge to become deep in the gap. Clearly, then the activation energy will depend on whether this level is occupied or not. In this way the pronounced reduction in the activation energy for β partials in p-type material, and α partials in n-type material, can be explained. A similar effect occurs for SiC and has led to predictions of the effect of doping on dislocations in that material (Sitch *et al.,* 1995).

X. Summary

The cluster theory that we have described has led to significant advances in understanding defects in bulk solids, atomic processes in molecules and interaction effects of hydrogen on diamond surfaces. The method is a straightforward application of local density functional theory, with a localized basis, to large clusters. It is remarkably stable with bonding patterns quite insensitive to cluster size and has been remarkable for the accuracy of the predicted local and resonant vibrational modes. It is perhaps this aspect that has caught the greatest attention of experimental groups, several of whom have sought help in the understanding of defects of interest to themselves. In many cases, this collaboration has been very successful and the theory has built upon experimental findings to elucidate the detailed geometry of a defect or a key ingredient in an atomic process.

The future advances in computing power — especially the development of cheaper parallel processor machines — will enable clusters as large as 1000 atoms to be *routinely* relaxed and investigated. This will pave the way for an exploration of the structure of larger clusters and extended defects, such as interstitial aggregates. However, this will take the theory into areas where few experiments can probe the microstructure and the results described in outline here must provide the underlying confidence in any predictions that emerge.

Acknowledgements

The many developments and applications of the formalism presented here have been carried out by a great number of collaborators. Among these, Sven Öberg, Malcolm Heggie, Paul Sitch, Chris Latham, Steve Breuer, Chris Ewels, Jon Goss, Paul Leary, Bernd Eggen, Deepak Jain, Grenville Lister, Andrey Umerski and Vitor Torres are especially to be thanked.

References

Ammerlaan, C. A. J., and Burgemeister, E. A. (1981). *Phys. Rev. Lett.* **47**, 954.

Bachelet, G. B., Baraff, G. A., and Schlüter, M. (1981). *Phys. Rev. B* **24**, 4736.

Bachelet, G. B., Hamann, D. R., and Schlüter, M. (1982). *Phys. Rev. B* **26**, 4199.

Bech Nielsen, B., Hoffmann, L., Budde, M., Jones, R., Goss, J., and Öberg, S. (1995). *Mat. Sci. Forum* **196-201**, 933.

Berg Rasmussen, F., Jones, R., and Öberg, S. (1994). *Phys. Rev. B* **50**, 4378.

Berg Rasmussen, F., Jones, R., Öberg, S., Ewels, C., Goss, J., Miro, J., and Deák, P. (1995). *Mat. Sci. Forum* **196-201**, 791.

Born, M., and Huang, K. (1954). *Dynamical Theory of Crystal Lattices,* Oxford University Press, London.

Breuer, S. J., and Briddon, P. R. (1995). *Phys. Rev. B* **51** 6984.

Breuer, S. J., and Briddon, P. R. (1996). *Phys. Rev. B* **53**, 7819.

Breuer, S. J., Jones, R., Briddon, P. R., and Öberg, S. (1996). *Phys. Rev. B* **53**, 16289

Briddon, P. R., and Jones, R. (1990). *Phys. Rev. Lett.* **64**, 2535.

Briddon P. R., Heggie, M. I., Jones R. (1991). *Mat. Sci. Forum* **83-7**, 457.

Briddon, P. R. (1996). Unpublished.

Collins, A. T., Stanley, M., and Woods, G. S. (1982). *Phil. Mag. A* **46**, 77.

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

Cook, R. J., and Whiffen, D. J. (1966). *Proc. Roy. Soc. A* **295**, 99.

Chen, X. J., Langlois, J. M., and Goddard, W. A. (1995). *Phys. Rev. B* **52**, 2348.

Cheng, Y., Stavola, M., Abernathy, C. R., Pearton, S. J., and Hobson, W. S. (1994). *Phys. Rev. B* **49**, 2469.

Cheong, B. H., and Chang, K, J. (1994). *Phys. Rev. B* **49**, 17436.

Clark, C. D., Kanda, H., Kiflawi, I., and Sittas, G. (1995). *Phys. Rev. B* **51**, 16681.

Clerjaud, B., and Côte D. (1992). *J. Phys. Condens. Matter* **4**, 9919.

Davidson, B. R. Newman, R. C. Bullough, T. J., and Joyce, T. B. (1993). *Phys. Rev. B* **48**, 17106.

Davidson, B. R., Newman, R. C., Kaneto, T., and Naji, O. (1994). *Phys. Rev. B* **50**, 12250.

Davies, G. (1976). *J. Phys. C: Solid St. Phys.* **9**, L537.

Delsole, R., Reining, L., Godby, R. W. (1994). *Phys. Rev. B* **49**, 8024.

Doukhan, J. C., Trepied, L. (1985). *Bull. Mineral.* **108**, 97.

Dunlap, B. I., Connolly, W. J., and Sabin, J. R. (1979). *J. Chem. Phys.* **71**, 4993.

Eggen, B. R., Heggie, M. I,. Jungnickel, G., Latham, C. D., Jones, R., and Briddon, P. R. (1996). *Science* **272**, 87.

Estreicher, S. K., Latham, C. D., Heggie, M. I., Jones, R., and Öberg, S. (1992) *Chem. Phys. Lett.* **196**, 311.

Evans, T., Qi, Z. and Maguire J. (1981). *J. Phys. C.: Solid St. Phys.* **14**, L379.

Ewels, C., Jones, R., and Öberg, S. (1995). *Mat. Sci. Forum* **196-201**, 1297.

Ewels, C. P., Öberg, S., Jones, R., Pajot, B., and Briddon, P. R. (1996a). *Semicond. Sci. and Technol.* **11**, 502.

Ewels, C. P., Jones, R., Öberg, S., Miro, J., and Deák, P. (1996b). *Phys. Rev. Lett.* **77**, 865.

Fushimi, H., and Wada, T. (1996). Private communication.

Goss J., Resende, A., Jones, R., Öberg, S., and Briddon, P. R. (1995). *Mat. Sci. Forum* **196-201**, 67.

Goss, J. P. Jones R., Breuer, S. J., Briddon, P. R., and Öberg, S. (1996a). *Phys. Rev. Lett.*, in press.

Goss, J. P., Jones, R., and Öberg, S. (1996b). Unpublished.

Griggs, D. T., and Blacic, J.D. (1965). *Science* **147**, 292.

Hedin, L. (1969). *Solid State Physics*, edited by F. Seitz, D. Turnbull, and H. Ehrenreich, Academic Press, New York, **23**, 1-180.

Heggie, M., and Jones, R. (1987). *Phil. Mag. Lett.* **55**, 47.

Heggie, M. I., Jones, R., and Umerski, A. (1991). *Inst. Phys. Conf. Series* **117**, 125.

Heggie, M. I., Jones, R., and Umerski, A. (1993). *Phys. Stat. Sol. (a)* **138**, 383.

Heggie, M. I., Briddon, P. R., and Jones, R., (1996a). Unpublished.

Heggie, M. I., Latham, C. D., Maynard, S. C. P., and Jones, R. (1996b). *Chem. Phys. Lett.* **249**, 485.

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

Janak, J. F. (1978). *Phys. Rev. B* **18**, 7165.

Jones, R., and Sayyash, A. (1986). *J. Phys. C: Solid St. Phys.* **19**, L653.

Jones, R. (1987). *J. Phys. C: Solid St. Phys.* **20**, 271.

Jones, R. (1988). *J. Phys C: Solid St. Phys.* **21**, 5735.

Jones, R., and Öberg, S. (1991a). *Phys. Rev. B* **44**, 3673.

Jones, R. (1991b). *Physica* B **170**, 181.

Jones, R., and Öberg S., (1992a). *Phys. Rev. Lett.* **68**, 86.

Jones, R., Briddon P. R., and Öberg S. (1992b). *Phil. Mag. Lett.* **66**, 67.

Jones, R., Umerski, A., Öberg, S. (1992c). *Phys. Rev. B* **45**, 11321.

Jones, R., and Öberg, S. (1992d). *Semicond. Sci. and Technol* **7**, 27.

Jones, R., and Öberg, S. (1994a). *Phys. Rev. B* **49**, 5306.

Jones, R., Goss, J., Ewels, C., Öberg, S. (1994b). *Phys. Rev. B* **50**, 8378.

Jones, R., Öberg, S., Berg Rasmussen F., and Bech Nielsen, B. (1994c). *Phys. Rev. Lett.* **72**, 1882.

Jones, R., Ewels, C., Goss, J., Miro, J., Deák, P., Öberg, S., and Berg Rasmussen, F. (1994d). *Semicond. Sci. and Technol.* **9**, 2145.

Jones, R., and Öberg, S. (1994e). *Mat. Sci. Forum* **143-47**, 253.

Jones, R., Leary, P., Öberg, S., and Torres, V. J. T. (1995a). *Mat. Sci. Forum* **196-201**, 785.

Jones, R., Öberg, S., Goss, J., Briddon, P. R., and Resende, A. (1995b). *Phys. Rev. Lett.* **75**, 2734.

Kajihara, S. A., Antonelli, A., Bernholc, J., and Car, R. (1991). *Phys. Rev. Lett.* **66**, 2010.

Kohn, W., and Sham, L. J. (1965). *Phys. Rev. A* **140**, 1133.

Koppitz, J., Schirmer, O. F., and Seal, M. (1986). *J. Phys. C, Solid St. Phys.* **19**, 1123.

Kozuch, G. M., Stavola, M., Pearton, S. J., Abernathy, C. R., and Lopata, J. (1990). *Appl. Phys. Lett.* **57**, 2561.

Kutzler, F. W., and Painter, G. S. (1992). *Phys. Rev. B* **45**, 3236.

Latham, C. D., Heggie, M. I., Jones, R., and Briddon, P. R. (1994). *Diamond and Related Materials* **3**, 1370.

Leary, P., Jones, R., Öberg, S., Torres, V. J. B. (1996). *Phys. Rev. B*, in press.

Leigh, R. S., and Szigeti, B. (1967). *Proc. Roy. Soc., A* **301**, 211.

Lister, G. M. S., and Jones, R. (1988). Unpublished.

Loubser, J. H. N., and van Wyk, J. (1981). *Proceedings of the Diamond Conference*, Reading, U.K. Unpublished.

Louie, S. G., Froyen, S., and Cohen, M. L. (1982). *Phys. Rev. B* **26**, 1738.

Matsui, M., and Yokoyama, T. (1986). *Inst. Phys. Conf. Series* **79**, 13.

Minaev, N. S., and Mudryi, A. V. (1981). *Phys. Stat. Solidi A* **68**, 561.

Musgrave, M. J. P., Pople, J. A. (1962). *Proc. Roy. Soc.* **A268**, 474.

Öberg, S., Sitch, P. K., Jones, R., and Heggie, M. I. (1995). *Phys. Rev. B* **51**, 13138.

Pederson, M., R., Jackson, K. A., and Pickett, W. E. (1991). *Phys. Rev. B* **44**, 3891.

Perdew, J. P., and Zunger, A. (1981). *Phys. Rev. B* **23**, 5048.

Porezag, D., Frauenheim, Th., Köhler, Th., Seifert, G., and Kaschner, R. (1995). *Phys. Rev. B* **51**, 12947.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1987). *Numerical Recipes*, Cambridge University Press, Cambridge.

Purton, J., Jones, R., Heggie, M., Öberg, S., Catlow, C. R. A. (1992). *Phys. Chem. Minerals* **18**, 389.

Safonov, A. N., Lightowlers, E. C., Davies, G., Leary, P., Jones, R., and Öberg, S. (1996). Unpublished.

Seifert, G., and Eschrig, H. (1985). *Phys. Stat. Sol. B* **127**, 573.

Seifert, G., Eschrig, H., and Bieger, W. (1986). *Z. Phys. Chem.* (Leipzig). **267**, 529.

Sitch, P., Jones, R., Öberg, S., and Heggie, M. I. (1994). *Phys. Rev. B* **50**, 17717.

Sitch, P., Jones, R., Öberg, S., and Heggie, M. I. (1995). *Phys. Rev. B.* **52** 4951.

Slater, J. C. (1960). *Quantum Theory of Atomic Structure*, Vol. 2, McGraw Hill, New York.

Smith, W. V., Sorokin, P. P., Gelles, L. L., and Lasher, G. J. (1959). *Phys. Rev.* **115**, 1546.

Sternschulte, H., Thonke, K., Sauer, R., Münzinger, P. C., and Michler, P. (1994). *Phys. Rev. B* **50**, 14554.

Stoneham, M. (1975). *Defects in Solids*, Oxford University Press, London.

Suezawa, M., Sumino, K., Harada, H., and Abe, T. (1986). *Jpn. J. Appl. Phys.* **25**, L859.

Sutton, A. P., Finnis, M, W., Pettifor, D. G., and Ohta, Y. (1988). *J. Phys. C: Sol. St. Phys.* **21**, 35.

Svane, A. and Gunnarsson, O. (1990). *Phys. Rev. Lett.* **65**, 1148.

Svelto, O. (1976). *Principles of Lasers*, Plenum Press, New York.

Szotek, Z. Temmerman, W. M., and Winter, H. (1993). *Phys. Rev. B* **47**, 11533.

Trombetta, J. M., and Watkins, G. D. (1987). *Appl. Phys. Lett.* **51**, 1103.

Wagner, J., Maier, M., Lauterback, Th., Bachem, K. H., Ashwin, M. J., Newman, R. C., Woodhouse, K., Nicklin, R. and Bradley, R. R. (1992). *Appl. Phys. Lett.* **60**, 2546.

Wagner, J., Bachem, K. H., Davidson, B. R., Newman, R. C., Bullough, T. J., and Joyce, T. B. (1995). *Phys. Rev., B* **51**, 4150.

Wagner, J., Newman, R. C, Davidson, B. R., Westwater, S. P., Bullough, T. J., Joyce, T. B., Latham, C. D., Jones, R., and Öberg, S. (1996). Unpublished.

Watkins, G. D. (1964). *Radiation Damage in Semiconductors*, edited by P. Barach, Dunod, Paris, 97.

Woodhouse, K., Newman, R. C., deLyon, T. J., Woodall, J. M., Scilla, G. J., and Cardone, F. (1991). *Semicond. Sci. and Technol.* **6**, 330.

Vaz de Carvalho, M. H., das Neves A. T. (1987). *J. Phys. C, Solid St. Phys.* **20**, 2713.

Von Barth, U., and Hedin, L. (1972). *J. Phys. C: Solid St. Phys.*, **5**, 1629.

Von Barth, U. (1979). *Phys. Rev. A* **20**, 1693.

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

Yonenaga, I., and Sumino, K. (1989). *J. Appl. Phys.* **65**, 85.

Zheng, J. F., Stavola, M., and Watkins, G. D. (1994). *The Physics of Semiconductors*, edited by D. J. Lockwood, World Scientific, Singapore, 2363.