Theor Chim Acta (1993) 84: 353–361

Theoretica Chimica Acta

Computational considerations for the study of defects in solids*

A. B. Kunz

College of Engineering and Department of Electrical Engineering, Michigan Technological University, Houghton, MI 49931, USA

Received October 1, 1991/Accepted January 7, 1992

Summary. A general method for the study of point and extended defects in non-metals has been formulated and a substantial computer program generated to allow the study of such systems in a routine manner. This method requires only a single ansatz; this is the effect of the defect in question is appreciable only in the immediate proximity of the defect. Beyond this region, the influence of the defect may be obtained from a simple response theory, which may be linear but is not required to be so. This response is manifested as a displacement of the ion cores and by the polarization of these atoms. This situation is considered in a mathematically rigorous extension of the local orbitals method of Adams–Gilbert–Kunz, using the approach defined by Kunz–Klein. This approach ultimately defines the system in terms of building blocks for the system, which may be defined in some arbitrary way. These building blocks form a natural point for parallelization of a computer code, and such has been simply accomplished. Each building block in turn is studied using slightly modified quantum chemical techniques at the Hartree–Fock and Moller–Plesset levels. These techniques are also parallelizable and such has been done. Thus a potential two levels of parallelization may be used here, and this makes possible an ultimate use of large-scale MIMD parallelism.

Key words: Defects – Solids – Non-metals – MIMD parallelism

1. Introduction

Lattice defects in or on crystalline materials, determine many technologically important properties. Reliable computer simulations of such defects are of potential value, and may be expected to contribute to a fundamental understanding of the physical processes that determine the structure and properties of these materials. In the case of point defects, it is attractive to use quantum mechanics to describe the region of the crystal in proximity to the defect, perhaps embed-

---

* This research has been sponsored by the U.S. Navy under the Office of Naval Research under grant ONR-N0001489-J-1601

ding this region in an external potential determined by some auxiliary principle. The hope here is that the response of the lattice to the point defect may then be described by some method which is simpler than the quantum mechanical method used to describe the point defect itself. It is noted parenthetically, that the definition of simpler, as used here, is that it be computationally less expensive to use. Similar considerations apply in a similar way to the case of adsorbates on solid surfaces. Models which may accurately describe the response of an embedding lattice do currently exist. In the present case we begin our development for the case of non-metals. In many studies performed prior to the present for such systems, the use of a classical shell model, based upon point charges, and masses, interacting by simple parameterized potentials has been successful in correlating perfect-lattice equilibrium data with the ground state properties of defects in these systems [1, 2]. Therefore, we begin our study by choosing to think of the embedding lattice in terms of the classical shell model. We find that it is possible to retain the functional form of the shell model, but determine all needed parameters from the quantum mechanical calculation, and to augment this functional form with appropriate angular potentials as well. The region about the defect can be described by means of an unrestricted Hartree–Fock method [3] (UHF). Such a model will in practice not yield sufficient accuracy for our purposes, and is extended by the implementation of a Many-Body Perturbation-Theory method (MBPT). By this choice, we will separate the problems of exchange from those of correlation, rather than combine them as is often the case in a computation based on the density functional method [4].

In the case of producing a technological approach, one must have in mind that the process be computationally accurate and also computationally inexpensive. This latter consideration is needed because one may envision studies of a great many physical situations. In such an event speed and efficiency not only minimize the computational cost but also the tendency of the researcher to fatigue or boredom. The way we are considering to speed up our computations is to invoke parallelism, hopefully in a massive way. The initial attempts for solid systems are described here.

In the case of a cluster embedded in a classical lattice, special care needs to be taken to ensure that mathematical consistency is achieved between the cluster and the embedding lattice. This has been solved formally by the work of Kunz and Klein [5], who achieve this through the introduction of a localizing potential, here called the Kunz–Klein localizing potential or KKLP.

Simulation of a large crystallite or an infinite lattice containing a point defect represented by a cluster and a polarizable embedding lattice is implemented here by means of an energy minimization procedure. That is, one minimizes the total system energy with respect to all parameters that define the lattice and the electronic configuration. For those parts of the lattice described by the shell model, one must minimize the total energy with respect to the positions of the ion cores, and also with respect to the polarization of the ions individually. For the quantum mechanical cluster, energy minimization is carried out with respect to the nuclear positions and also the electronic configuration. In this method it is possible to study states other than the ground state. Since the primary physical outputs are total energies and geometries, spectroscopic data is obtained from total energy differences. Positional variations are carried out initially using the HADES approach as implemented in the ICECAP procedure, or more recently using a Monte Carlo approach [6].

In the next section of this paper, we describe the basic theoretical ideas used in this study. This will include the shell model lattice, the UHF method, the KKLP and the inclusion of correlation via MBPT. The evocation of parallelism may be on two levels, and both are considered here. The first is in a global sense, that is over the chemical building blocks created in the Kunz–Klein partitioning into fragments. Largely the computation of each fragment proceeds independently of each other fragment, although there is some small amount of data interchanged. The second level of parallelism is within the individual fragment. This is also possible to demonstrate, and is done here. Combining both levels may well enable the use of massive parallelism provided a MIMD computer system is available. There are techniques which we consider in our laboratory which may significantly reduce the cost and complexity of such computations.

2. Theoretical methods

In these studies, we assume that we have a system consisting of n electrons and N nuclei. The n electrons have coordinates designated by x_i and mass, m, and charge e. The nuclei have coordinate R_I, and nuclear charge Z_I. In these studies, the Born–Oppenheimer approximation is used and thus the nuclear mass is treated as infinite. The electron coordinate includes spin degrees of freedom. In general lower-case letters refer to electron attributes, while upper-case letter refer to nuclear properties. In this study the atomic system of units is used. That is; Planck's constant, the electronic charge, and the electronic mass are set to unity. Thus, the unit of length is approximately 0.53×10^{-8} cm, and the unit of energy is approximately 27.2 eV. In the usual non-relativistic formalism, the Hamiltonian for the system is:

$$H = -\frac{\hbar^2}{2m}\sum_i \nabla_i^2 - \sum_{i=1}^{n}\sum_{I=1}^{N}\frac{Z_I e^2}{|\vec{r}_i - \vec{R}_I|} + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{e^2}{|\vec{r}_i - \vec{r}_j|} + V_{NN}. \tag{1}$$

Ideally one would like to solve the n-electron Schroedinger equation for this Hamiltonian:

$$H\Psi(\vec{x}_i, \vec{X}_i) = E\Psi(\vec{x}_i, \vec{X}_i) \tag{2}$$

but computational difficulties preclude this. Instead we will resort to a series of approximations beginning with the UHF approximation. In the UHF approximation, the n-electron wavefunction is approximated by an antisymmetrized product of one electron orbitals. These orbitals are chosen to be orthonormal, and to minimize the energy expectation value of the Hamiltonian with respect to the functional form of these orbitals. This set of approximations leads to the system of equations called the UHF equations:

$$F(\varrho)\varphi_i(\vec{x}) = \epsilon_i\varphi_i(\vec{x}) \tag{3}$$

$$F(\varrho) = -\frac{\hbar^2}{2m}\nabla^2 - \sum_{I=1}^{N}\frac{Z_I e^2}{|\vec{r} - \vec{R}_I|} + e^2\int\frac{\varrho(\vec{r}\,'\vec{r}\,')}{|\vec{r} - \vec{r}\,'|}\,d\vec{r}\,' - e^2\varrho(\vec{x}\vec{x}\,')\frac{1}{|\vec{r} - \vec{r}\,'|}\hat{p}(\vec{x}\,'\vec{x}) \tag{4}$$

$$\varrho(\vec{x}\vec{x}\,') = \sum_{i < i_{\text{fermi}}} \varphi_i(\vec{x})\varphi_i^{+}(\vec{x}\,') \tag{5}$$

$$\int \varphi_i^{*}(\vec{x})\varphi_j(\vec{x})\,d\vec{x} = \delta_{ij}. \tag{6}$$

This series of equations may be solved in matrix form using a basis set expansion in terms of contracted gaussian basis orbitals. This procedure is so standard as to require no further discussion here. The one possible key to the parallelization of the matrix equations is to note that the needed integrals over contracted basis functions are all independent of each other. Thus one may parallelize the integral generation simply by partitioning out the needed integrals over the available processors. The use of the integrals in the UHF sequence is also independent in this way. Therefore one may fill a partial Fock matrix on each processor and only amalgamate the matrices at the final step. In this way the parallelization of the Fock problem is rapidly accomplished. My group has developed this process for a Ghost Cube of Suns, an Ametek S2010, and the Intel Hypercube.

The practical problem is to be able to solve this set of equations for extended systems. In the case of the pure, perfect, periodic system, techniques of energy band theory may be used [7]. However, we wish to be able to consider defects as well. There are methods to study some defect cases based upon periodic super-cell methods [8], but in our case the study of charged defects in insulating solids is envisioned. Such studies don't lend themselves well to super-cell methods due to the infinite range of the coulomb potential. Instead, we resort to the older method of local orbitals introduced formally by Adams [9], and Gilbert [10], and given a computational formulation by the author [11]. In this method, we formally divide the system into two parts, the cluster, and its environment. The cluster in practice contains the defect or impurity in question as well as the first few shells of atoms surrounding the defect. The environment contains the remainder of the system.

The Hamiltonian is formally partitioned into two parts, F_A, the cluster Hamiltonian, and V_A, the Hamiltonian for the environment. These are formally:

$$\begin{aligned} F_A = -\frac{\hbar^2}{2m}\nabla^2 - \sum_{I_{\text{in } A}} \frac{Z_I e^2}{|\vec{r}-\vec{R}_I|} + \int \frac{\varrho_A(\vec{r}\,'\vec{r}\,')}{|\vec{r}-\vec{r}\,'|}\,d\vec{r}\,' \\ - e^2 \varrho_A(\vec{x}\vec{x}\,')\frac{1}{|\vec{r}-\vec{r}\,'|}\hat{p}(\vec{x}\,'\vec{x}) \end{aligned} \tag{7}$$

$$V_A \equiv F - F_A. \tag{8}$$

We may further divide the term V_A into two parts. The first is due to the ionic nature of the individual atoms, if any, and is called V_M, and the part due to the non-ionic nature of the atoms, a short range part termed U_A. Having done this we may formally consider the Adams–Gilbert modified UHF equation:

$$[F + \varrho A \varrho]\hat{\varphi}_{Ai} = \hat{\varepsilon}_{Ai}\hat{\varphi}_{Ai}. \tag{9}$$

This equation is simply a canonical transformation on the orignal UHF equation [10], and its solution forms a first order density matrix identical to that of the original UHF equation, and leaves the total system energy unmodified.

In the present implementation, one will choose the arbitrary operator, A, to be $-U_A$. Therefore, we solve the one-electron equation:

$$[F_A + V_M + U_A - \varrho U_A \varrho]\hat{\varphi}_{Ai} = \hat{\varepsilon}_{Ai}\hat{\varphi}_{Ai}. \tag{10}$$

In this implementation the term $\varrho U_A \varrho$ tends to cancel the term U_A in the original UHF operator, and allows the cluster to localize a number of electrons in it. The remainder of the electrons are of necessity delocalized into the environment. One may naturally play the same localization trick on the environment electrons as well.

The term $\varrho U_A \varrho$ is called the KKLP due to its operational effect. The strategy of solution is as follows: Begin with the pure, perfect, periodic lattice. Divide this lattice into some set of natural building blocks. One's intuition will normally dictate a good choice of building blocks. For example, in the case of crystalline NaCl, a natural set of building blocks are the systems Na^+, and Cl^-. These are not free space ions but rather ions self consistently distorted by the crystalline environment. We assume that the orbitals associated with ions situated suitably far from the defect are the same as for the ions of the pure, perfect, periodic crystal. We do not assume that they sit at perfect lattice sites, and allow the potential due to them to be further modified by a polarization contribution.

The implementation of parallelism within the basis Local Orbital procedure is trivial. The building blocks are almost independent of each other. They interact in a mathematical sense only through the knowledge of each other's first order density matrix. Therefore one can parallelize on a grand scale by treating each block in parallel, and exchanging first order density matrix information only as every building block becomes self consistent with the first order density matrix information of the other building blocks. In this way the computational time exceeds the data transfer times by many orders of magnitude. This is even true using a Ghost Cube of Sun workstations connected by ethernet. It is additionally possible to implement parallelization of the individual UHF problems as outlined previously.

In the above discussion, we used the one electron approximation exclusively. This is found to be of inadequate precision for our needs and therefore, we seek to include explicit electron correlation in the cluster computation. This is most easily achieved by the use of a MBPT formalism. This is a natural choice in some ways for a solid system due to the simple fact that the MBPT method is extensive (size-consistent) [13].

The essential features of this approach are demonstrated by consideration of the non-degenerate state case. Extension to a degenerate system can be obtained as well, although the practicalities of implementation are far less simple. In any event, for the cases needed here nondegenerate perturbation theory is adequate. Consider a Hamiltonian, H, which is partitioned into two parts, a zero order Hamiltonian, H_0, whose eigenvalues and eigenvectors are known, and a perturbation, V. Thus:

$$H = H_0 + V \tag{11}$$

and

$$H_0 \Phi_i = W_i \Phi_i. \tag{12}$$

From these solutions we may construct the eigenvalues, and eigenfunctions of the total Hamiltonian. That is we formally solve the correct equation:

$$H\Psi_i = E_i \Psi_i \tag{13}$$

finding that

$$\Psi_i = [1 - (H_0 - W_i)^{-1}(1 - P_I)(E - V - W_i)]^{-1}\Phi_i \tag{14}$$

and

$$E_I = W_I + \langle \Phi_I | V | \Phi_I \rangle + \langle \Phi_I | V(H_0 - W_0)^{-1}(1 - P_I)(-V) | \Phi_I \rangle + \cdots. \tag{15}$$

In the present case we need to properly pick a H_0. This is usually chosen in such calculations to be the sum of the one body UHF operator, as in Moeller–Plesset perturbation theory, however, we do not necessarily know the canonical UHF solution, only a canonical transformation of it. We therefore know only the eigenfunctions and eigenvalues of the Adams–Gilbert–Kunz Eq. (10). We chose the sum of these one body equations to be our zero-order Hamiltonian. This allows a formally tractable solution to be obtained. This solution through second order becomes simply:

$$E_I = \langle \Phi_I | H | \Phi_I \rangle + \sum_{J \neq I} \frac{\langle \Phi_I | V | \Phi_J \rangle^2}{W_I - W_J}. \tag{16}$$

This is the correlation correction used in our work. The invocation of parallelism is simple here in that the MBPT correction at the second order may be made in parallel for each integral. This has been implemented by us and is perhaps the most nearly ideal parallel implementation to date, at least for small-scale parallelism.

3. Parallelization results

The essential consideration initially here is Amdahl's law. This is seen as placing a limit on the value of parallelization, as Amdahl rightly assumes that for a given size problem, when all efficiency is gained from parallelization, there still remains a portion to do that doesn't parallelize. Thus parallelization can only provide a certain potential gain until the parts of the problem that don't parallelize dominate. Thus it is useful to consider the parallelization strategy carefully to ascertain its likely efficiency. In this task one need also consider the points raised by the work of John Gustafson [14]. John observes that the point of parallel development is to enable the solution of problems beyond our current capability. Thus consideration of how the problem parallelizes as a function of problem size is essential.

Consider the current case as an example. There are several possibilities once one has developed a strategy. The typical program has several parts. These may briefly be identified as:

1. NP, these are program parts that don't readily parallelize.
2. NWE, these are program parts which parallelize, but for which the speed gains are not worth the effort.
3. NYD, these are program parts which parallelize, and are worth the effort, but for which the work of parallelizing is not yet done.
4. P, these are program parts for which functioning parallel code has been generated.

To see how this works as a function of problem size, keeping in mind the ultimate aim is to solve problems on a parallel machine which are too large to solve on a current uniprocessor, we examine the lcoal orbital problem as a function of problem size. Since we have two levels of parallelism we must be careful. In the case of the overall local orbital procedure, termed LOPAS, the size element is the number of building blocks. Within a building block, we consider four essential program elements, the symmetry study called LISTER, the generation of molecular integrals over contracted functions called POLY, the self consistent field solution called SCF, and the computation of the correlation energy called MBPT. For all four segments in our strategy the size element is the number of basis functions. There is a fifth element as well, and this is the classical embedding lattice. For this segment, we haven't considered the process of parallelization yet.

Let the size element be N. This is true whether the size refers to the number of building blocks or the basis set size. The problem then breaks into our parts as follows:

	NP	NWE	NYD	P
LOPAS				N
LISTER	N	N^3		
POLY	N	N^2		N^4
SCF	N	N^2	N^3	N^4
MBPT	N			N^5

It is clear from the above that in all cases the portions which parallelize increase in computational intensity far more rapidly with size than do the portions which are not parallelizable or which are not worth the effort. Thus, this is an excellent program series from the aspect of parallel development. The real remaining question is how well we did do. In this our tests are limited by the equipment on hand. The sequence LISTER, POLY, SCF, MBPT was rested on an Intel Hypercube, the Ametek S2010, and a Ghost Cube of Sun Workstations. The most efficient by far was the Ametek S2010. This is partly because a disk drive was available internal to the parallel computer itself and communications to the outside were avoided as were necessary with the other two setups. Unfortunately, the S2010 had only 6 nodes and one internal drive, limiting our ability to fully test the code. The results presented here are for the KrF_2 molecule. A rich basis is used including *s*, *p*, *d*, and *f* orbitals on the Kr and *s*, *p*, and *d* orbitals on the F. The results for the molecule were most satisfactory. Times are total parallel system times and are given in seconds as a function of the number of parallel processors.

The second phase of the test consisted of the study of the ability to parallelize the LOPAS sequence. In this case communication speed is not very important as the ratio of computation to communication is high. Therefore almost any reasonable parallel system works. However, this phase is not a candidate for a large number of processors, as the natural size parameter is the number of building blocks. The additional parallelism being gained within the blocks as outlined above. Thus the ideal processor for this step is itself a parallel machine. In our test then the data was obtained for a Ghost Cube of Sun Sparc 1 Workstations. In this test there are four building blocks, each a Methane molecule but the four cited so that there are four distinct methanes by virtue of environment. The workstations were on various faculty desk tops and faculty use of the stations was not curtailed for the test. In fact the four faculty were unaware that their systems were being used until after the test. The results were:

1 Sparc Station	412 sec.
2 Sparc Stations	204 sec.
4 Sparc Stations	112 sec.

From this result we see that the parallelization of such a sequence is possible at two separate levels. The ideal would be to use both levels simultaneously. Unfortunately, our equipment inventory to date has not allowed this to occur.

Table 1. The computer times for various numbers of processors on a Ametek S2010 are given for the various steps in a LOPAS calculation. Times are total time and are in seconds. Relative times are also given. In this event the 2 processor time is defined as unity

No. Processors	2	4	6
LISTER	502		
	1		
POLY	15942	7342	5017
	1	0.46	0.31
SCF	21156	11149	9471
	1	0.53	0.45
MBPT	14236	7162	4846
	1	0.50	0.34

The results seen in Table 1 for the speedup upon using several processors is not unreasonable. However, there is one odd feature here, this is the slight superlinear speed increase exhibited in the POLY code demonstrated here. We are unable to give a complete understanding of this particular result. The S-2010 system had only one node with a disk drive attached, and this unusual result may in some way reflect upon the efficient use of the I/O subsystem as more use of it is made per unit time in the POLY code sequence. On the other hand, it may simply reflect some not understood peculiarity in the operating system of the S-2010 processor. In any event, the author is convinced that this superlinearity is not some special feature of the coding technique used here. The second place in which some superlinearity is seen is in the LOPAS portion upon running the code on 1 then 2 then 4 workstations. In going from 1 to 2 workstations, the speed increase seems to be superlinear. This may in this case be attributed to the fact that the workstations in question were in use by their owners as normal workstations, as well as a node on a more global parallel machine. Therefore, the times reflect in some measure the use levels of the local person as well as the usage for this test. In this latter case small time deviations are not indicative of performance issues, but rather an indication of test conditions.

4. Conclusions

It is possible to approach the solutions to solid-state problems by use of standard molecular techniques which are slightly modified to include the necessary KKLP. In this way a canonical transformation is used to enable a solution to the solid-state problem in terms of local orbitals, and which also preserves the energy minimization of the SCF scheme. This enables a strategy of efficient computer usage based upon parallelism. This occurs on two levels. The first and largest in granularity is over the chemical building blocks within the LOPAS scheme itself. The second level is within the molecular sequence itself. In this case the independence of integrals over contracted basis functions is used. Both levels may be done simultaneously provided the computational equipment permits. In fact, this may well be a preferred method as the number of processors assigned to a building block can compensate for differences in building block size. Small-scale demonstrations of the effectiveness of these schemes are given, and good speedups upon parallelization are seen.

References

1. Dick BG, Overhauser AW (1958) Phys Rev 112:90
2. Bilz H, Kress W (1979) Phonon dispersion relations in insulators. Springer-Verlag, Berlin
3. Kunz AB, Vail JM (1988) Phys Rev B38:1058
4. Kunz AB, Meng J, Vail JM (1988) Phys Rev B38:1064
5. Kunz AB, Klein DL (1978) Phys Rev B17:4614
6. The ICECAP process has been given its public form in the manuscript: Harding JH, Harker AH, Keegstra PB, Pandey R, Vail JM, Woodward C (1985) Physica B + C131B:151. Later improvements are described in: Kunz AB, Vail JM (1988) Phys Rev B 38:1058. The HADES portion is described: Cade PE, Stoneham AM, Tasker PW (1984) Phys Rev B 30:4621. The LOPAs work is described in its present implementation as: Kunz AB (1982) Phys Rev B 26:2056. The basic Monte Carlo concepts are given in: Binder K (1979) Monte Carlo methods in statistical physics. Topics in Current Physics 7:1, Springer Verlag
7. Callaway JC (1974) Quantum theory of the solid state. Academic Press, NY
8. Stoneham AM (1975) Theory of defects in solids. Oxford Press, Oxford
9. Adams WH (1961) J Chem Phys 34:89
10. Gilbert TL (1964) in: Löwdin PO (ed) Molecular orbitals in chemistry, physics and biology. Academic Press, NY
11. Kunz AB (1969) Phys stat solidi 36:301
12. Pandey R, Kunz AB (1988) J Mater Res 3:1362
13. Thouless DJ (1961) Quantum mechanics of many body systems. Academic Press, NY
14. Gustafson JL (1988) Communications of the ACM 31:532