LARGE-SCALE AB INITIO CALCULATIONS

Tsuyoshi MIYAZAKI

National Institute for Materials Science,
1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

Rathin CHOUDHURY, David BOWLER and Michael GILLAN

Dept. of Physics and Astronomy, University College London,
Gower Street, London WC1E 6BT, UK

Density functional theory (DFT) has become a standard tool for modelling materials. But conventional methods are very inefficient for large complex systems, because the memory requirements scale as N^2 and the cpu requirements as N^3 (N is the number of atoms). We report recent progress in the development of the CONQUEST code, which performs $O(N)$ (linear-scaling) DFT calculations on parallel computers, and has a demonstrated ability to handle systems of over 10,000 atoms. The code is based on the strategy of minimising the total energy with respect to the Kohn-Sham density matrix, and the practical techniques for implementing this strategy are outlined. The code can be run at different levels of precision, ranging from empirical tight-binding, through *ab initio* tight-binding, to full *ab initio*, and techniques for calculating ionic forces in a consistent way at all levels of precision will be presented. Illustrations are given of practical CONQUEST calculations on semiconductor surface reconstructions. The outlook for future large-scale work on surface nanostructures will be sketched.

1. INTRODUCTION

The aim of this paper is to summarize very recent progress in techniques for performing *ab initio* calculations on very large systems, using methods based on density-functional theory (DFT) and pseudopotentials.[1] The key to progress in this area is the principle of "near-sightedness",[2] i.e. the spatial localization of the density matrix, which is the basis for linear-scaling, also called $O(N)$ techniques,[3,4] in which the number of computer operations and the computer storage needed to perform an electronic total-energy calculation are proportional to the number of atoms in the system. The CONQUEST DFT code[5–11] is designed to perform this kind of calculation, and its $O(N)$ capabilities have been tested on systems of up to 16,000 atoms,[10] but up to now it has been limited to rather simple systems. We report here the developments which now make it possible to do $O(N)$ DFT calculations on non-trivial materials problems using the CONQUEST code.

The DFT-pseudopotential method[1] has been an established part of the materials modeller's toolkit for something like 20 years. It is routinely used by hundreds of groups worldwide for study-

ing problems ranging all the way from surface catalysis to planetary interiors, and from aqueous solution chemistry to semiconductor nanostructures. It is used as a matter of course for materials containing elements from the whole periodic table, including the rare-earths and actinides. Although it is well recognised that DFT methods may sometimes suffer from significant inaccuracies, particularly for strongly correlated systems, DFT results are widely used as the point of reference for validating more approximate methods, such as *ab initio* tight-binding or empirical tight-binding. But in spite of its widespread use, traditional DFT-pseudopotential methods are difficult to apply to systems of more than a few hundred atoms, because the computer effort increases at least as fast as N^2, and ultimately as N^3, as the number of atoms N increases. This means that there are many important areas of materials science that are currently out of range of traditional methods, for example large-scale nanostructures, such as quantum dots and nanowires, cracks and voids, and large biomolecules such as proteins and DNA. This situation has stimulated a major effort over the past 10 years to develop $O(N)$ methods for electronic-structure methods in general,[12–17] and the DFT-pseudopotential method in particular.[5,6,11,18–22]

The origin of the poor scaling of traditional methods is well understood. These methods generally work with electronic orbitals that are eigenfunctions of the Hamiltonian, which extend over the entire system. Both the number of orbitals and the amount of information in each orbital are proportional to N, so the computer effort must go at least as N^2. The need to handle scalar products between all pairs of orbitals brings the asymptotic scaling to N^3. It is also well understood that this poor scaling is unnecessary. The amount of information needed to specify the electronic structure does not really scale as N^2, it scales as N. To exploit this fact, all that is necessary is to work with the density matrix $\rho(\mathbf{r},\mathbf{r}')$; since this decays to zero as $|\mathbf{r}-\mathbf{r}'|\rightarrow\infty$, the linear scaling of the information content is made apparent.[2] Equivalently, one can work with "localised orbitals", i.e. linear combinations of the eigenfunctions constructed so that they differ significantly from zero only in spatially localised regions.[12,13]

These ideas form the basis of a range of $O(N)$ reformulations of electronic-structure methods, including tight-binding,[3] DFT-pseudopotentials,[5,6,11,18,19,22] Hartree-Fock,[16] post-Hartree-Fock, and quantum Monte Carlo.[17] The CONQUEST code is a practical implementation of the $O(N)$ ideas for the DFT-pseudopotential technique. Later in the paper, we point out its connections with the SIESTA code,[18,19] which also has $O(N)$ capabilities.

Demonstrations of the practical $O(N)$ scaling of CONQUEST on massively parallel computers were reported several years for simple silicon systems.[10] It was shown that the scaling is almost indistinguishable from linear over a very wide range of N. However, unreliable stability in the ground-state search and in electronic self-consistency limited the effectiveness of the code. A further obstacle has been the need to achieve exact consistency between atomic forces and the total energy,

which is essential for both structural relaxation and for dynamical simulation. Major improvements in the stability of the code have already been reported so here we emphasise recent progress in the calculation of forces.

The paper is organised as follows. In Sec. 2, we summarise briefly the main techniques used in CONQUEST, focusing on the hierarchy of precisions that can be used. Sec. 3 outlines how we have implemented the calculation of forces throughout the hierarchy. Practical results on the reconstruction of Si surfaces are presented in Sec. 4 to illustrate the present capabilities of the code, and we conclude in Sec. 5 with comments on future directions.

2. SUMMARY OF CONQUEST TECHNIQUES

In DFT,[1] the total energy $E_{\rm tot}$ is a sum of contributions associated with electronic kinetic energy $E_{\rm kin}$, electron-pseudopotential energy $E_{\rm ps}$ (usually, a sum of local and non-local parts), Hartree energy $E_{\rm Har}$, exchange-correlation energy $E_{\rm xc}$ and ion-ion Coulomb energy $E_{\rm ion}$:

$$E_{\rm tot} = E_{\rm kin} + E_{\rm ps} + E_{\rm Har} + E_{\rm xc} + E_{\rm ion} \ . \tag{1}$$

Our starting point for the development of $O(N)$ DFT has been the observation that $E_{\rm tot}$ can be expressed in terms of the Kohn-Sham density matrix $\rho(\mathbf{r},\mathbf{r}')$, and that the self-consistent ground state is obtained by minimising $E_{\rm tot}$ with respect to ρ, subject to the condition that ρ is idempotent.[5,6] The idempotency condition means that ρ is a projector (the projector onto the subspace of occupied states), or equivalently that its eigenvalues are either 0 or 1. If ρ is to be approximated, the condition of idempotency can be replaced by that of 'weak' idempotency,[12] meaning that all eigenvalues are in the interval $[0,1]$.

In order to proceed, we assume, without significant loss of generality, that $\rho(\mathbf{r},\mathbf{r}')$ is *separable*, so that it can be represented in the form:

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\lambda\mu} \phi_\lambda(\mathbf{r}) K_{\lambda\mu} \phi_\mu(\mathbf{r}') \ , \tag{2}$$

where the space spanned by the functions $\phi_\lambda(\mathbf{r})$'s contains the occupied subspace. The ground state search is then performed by variation of both the $\phi_\lambda(\mathbf{r})$ and of the matrix elements $K_{\lambda\mu}$, subject to weak idempotency.[5,6] In practice, we take the $\phi_\lambda(\mathbf{r})$ to be 'localised orbitals': they are functions that are freely varied, but are non-zero only within spherical regions (so-called 'support regions') centred on the atoms. There is considerable freedom in choosing the number of ϕ_λ on each atom, but there are natural choices.

The practical implementation of this scheme raises two major questions: (i) what basis set to use to represent the variable localised orbitals; (ii) how to impose weak idempotency. The basis-set question is as old as quantum mechanics, and many ways have been proposed to represent localised

orbitals, including numerical representation on a grid,[23] spherical waves,[20] pseudo-atomic basis sets,[18,19,22] and the B-spline method.[7] At present in CONQUEST, two options are implemented: B-splines, which are mathematically simple and systematically improvable, much like plane-waves; and pseudo-atomic orbitals, which are economical, and very often give accurate results. We envisage that other options will be added as required.

There are also several techniques for imposing weak idempotency. At present, we use a combination of the techniques of Li, Nunes and Vanderbilt (LNV)[14] and Palser and Manolopoulos,[24] both of which are closely related to McWeeny's 'purification' scheme.[25] In the LNV technique, the density matrix K is represented in terms of an 'auxiliary' density matrix L as:

$$K = 3LSL - 2LSLSL \ , \tag{3}$$

where S is the overlap matrix for localised orbitals: $S_{\lambda\mu} = \langle\phi_\lambda|\phi_\mu\rangle$. Minimisation of the total energy with respect to L automatically drives K towards idempotency. To achieve $O(N)$ operation, the minimisation is performed with a spatial cut-off on the L-matrix, so that $L_{\lambda\mu} = 0$ when the distance between the centres of the support-functions ϕ_λ and ϕ_μ exceeds a chosen cut-off R_L. Again, other methods for imposing idempotency could be implemented with little effort.

As an alternative to searching for the ground state by varying L, CONQUEST also has the option of obtaining the ground state directly by diagonalisation. Of course, this is an $O(N^3)$ operation, and it will be appropriate to run in this mode only for small systems. Nevertheless, it is extremely useful to be able to use diagonalisation, because it allows one to test the $O(N)$ errors incurred with a given cut-off on the L-matrix.

The ground-state search is organised into three loops. In the innermost loop, the ground state is determined for fixed support functions and fixed electron density, either by varying L or by diagonalisation. In the middle loop, self-consistency is achieved by systematically reducing the electron-density residual, i.e. the difference between the input and output density in a given self-consistency cycle.[26] In the outer loop, the ϕ_λ are varied. This organisation corresponds to a hierarchy of approximations. If only the inner loop is used, we get the scheme known as non-self-consistent *ab initio* tight binding (NSC-AITB), which is a form of the Harris-Foulkes approximation.[27–30] If the inner two loops are used, we get self-consistent *ab initio* tight binding (SC-AITB). If all loops are used, we have full *ab initio*. In this last case, we recover the exact DFT ground state as the region radius $R_{\rm reg}$ and the L-matrix cut-off R_L are increased. For non-metallic systems, the evidence so far is that accurate approximations to the ground state are obtained with quite modest values of the cut-offs.[6,19]

The scheme we have outlined is closely related to the methods used in SIESTA.[18,19] The main differences are: (i) the SIESTA code has not up to now allowed the localised orbitals ϕ_λ to vary, and

instead the ϕ_λ's are represented by fixed PAO's; (ii) idempotency is imposed using the techniques developed independently by Mauri *et al.*[12,31] and by Ordejon *et al.*;[13] (iii) the technique of 'neutral-atom potentials'[18,19] allows calculation of matrix elements to be performed very efficiently.

CONQUEST was written from the outset as parallel code, and a large part of the development effort has been concerned with techniques for achieving good parallel scaling. The parallelisation techniques have been described in detail elsewhere,[7,10,11] so we give only a brief summary. There are three main types of operation that must be carefully distributed across processors:

- the storage and manipulation of localised orbitals, e.g. the calculation of $\phi_\lambda(\mathbf{r})$ on the integration grid starting from blip- or PAO-coefficients, and the calculation of the derivatives of $E_{\rm tot}$ with respect to these coefficients, which are needed for the ground-state search;
- the storage and manipulation of elements of the various matrices (H, S, K, L, etc...);
- the calculation of matrix elements by summation over domains of points on the integration grid.

Efficient parallelisation of these operations, and the elimination of unnecessary communication between processors, depend heavily on the organisation of both atoms and grid points into small compact sets, which are assigned to processors.[10] When the code runs in $O(N)$ mode, matrix multiplication takes a large part of the computer effort, and we have developed parallel multiplication techniques[10] that exploit the specific patterns of sparsity on which $O(N)$ operation depends.

3. IONIC FORCES

In any electronic-structure scheme that is designed for structural relaxation and dynamical simulation of materials, the algorithms for calculating the forces $\mathbf{F}_i$ on the ions must be the exact derivatives of the ground-state energy $E_{\rm tot}$ with respect to ionic positions $\mathbf{R}_i$, so that $\mathbf{F}_i = -\nabla_i E_{\rm tot}$. A well-known advantage of the DFT-pseudopotential scheme is that it is straightforward in principle to achieve this exact relationship between energy and forces. Force algorithms within the various self-consistent and non-self-consistent schemes used in CONQUEST have been extensively discussed in the literature. Nevertheless, we have found it necessary to re-examine the calculation of forces, in order to develop a scheme that works in a unified way for all levels of the hierarchy of approximations, and which also works equally well for both the diagonalisation and $O(N)$ modes of operation (see previous Section). We summarise here the force algorithms that have very recently been implemented in CONQUEST – a full report on this will be published elsewhere (Miyazaki *et al.*, to be published).

At the empirical TB level, the ionic force is a sum of the band-structure part $\mathbf{F}_i^{\mathrm{BS}}$ and the pair-potential part $\mathbf{F}_i^{\mathrm{pair}}$, the former being given by:[29]

$$\mathbf{F}_i^{\mathrm{BS}} = -2\mathrm{Tr}\left[K\nabla_i H - J\nabla_i S\right] , \tag{4}$$

where K and J are the density matrix and energy matrix respectively.[29] It is readily shown that in the $O(N)$ scheme of LNV, and in some other $O(N)$ schemes, the same formula for $\mathbf{F}_i^{\mathrm{BS}}$ is the exact derivative of the $O(N)$ total energy. In the LNV scheme, K is given by eqn (3), and J by:

$$J = -3LHL + 2LSLHL + 2LHLSL . \tag{5}$$

In NSC-AITB (Harris-Foulkes), the forces can be written in two equivalent ways. The way that corresponds most closely to empirical TB is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{BS}} + \mathbf{F}_i^{\Delta\mathrm{Har}} + \mathbf{F}_i^{\Delta\mathrm{xc}} + \mathbf{F}_i^{\mathrm{ion}} , \tag{6}$$

where $\mathbf{F}_i^{\mathrm{BS}}$ is given by exactly the same formula as in empirical TB. The contributions $\mathbf{F}_i^{\Delta\mathrm{Har}}$ and $\mathbf{F}_i^{\Delta\mathrm{xc}}$, which arise from the double-counting Hartree and exchange-correlation parts of the NSC-AITB total energy, have been discussed elsewhere.[29] The final term $\mathbf{F}_i^{\mathrm{ion}}$ come from the ion-ion Coulomb energy. This way of writing $\mathbf{F}_i$ expresses the well-known relationship between NSC-AITB and empirical TB that in the latter the pair term represents the sum of the three contributions $\Delta\mathrm{Har} + \Delta\mathrm{xc}+$ ion-ion. The alternative, and exactly equivalent, way of writing $\mathbf{F}_i$ in NSC-AITB is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{ps}} + \mathbf{F}_i^{\mathrm{Pulay}} + \mathbf{F}_i^{\mathrm{NSC}} + \mathbf{F}_i^{\mathrm{ion}} . \tag{7}$$

Here, $\mathbf{F}_i^{\mathrm{ps}}$ is the "Hellmann-Feynman" force exerted by the valence electrons on the ion cores; $\mathbf{F}_i^{\mathrm{Pulay}}$ is the Pulay force that arises in any method where the basis set depends on ionic positions; $\mathbf{F}_i^{\mathrm{NSC}}$ is a force contribution associated with non-self-consistency, and is expressed in terms of the difference between output and input electron densities; $\mathbf{F}_i^{\mathrm{ion}}$, as before, is the ion-ion Coulomb force. Exactly the same formulas represent the exact derivative of E_{tot} in both diagonalisation and $O(N)$ modes.

In both SC-AITB and full AI, the force formula is:

$$\mathbf{F}_i = \mathbf{F}_i^{\mathrm{ps}} + \mathbf{F}_i^{\mathrm{Pulay}} + \mathbf{F}_i^{\mathrm{ion}} , \tag{8}$$

which differs from the second version of the NSC-AITB formula eqn (7) only by the absence of the non-self-consistent contribution $\mathbf{F}_i^{\mathrm{NSC}}$, as expected.

The above hierarchy of force formulas has been implemented in CONQUEST, and extensive tests have ensured that the total energy and the forces are exactly consistent within rounding-error precision.

4. ILLUSTRATIVE RESULTS

To illustrate the ability of CONQUEST to address non-trivial problems, and to show how the levels of the hierarchy of approximations can work together, we have performed relaxations of the reconstructed Si(001) surface (see Fig. 1). We have used the PAO and B-spline basis sets available in the code, and have performed relaxations using non-self-consistent AITB, self-consistent AITB, and full *ab initio*. We compare the results against those of two other standard codes: the SIESTA code[19] for PAO comparisons; and the VASP code[32] for plane-wave comparisons.

When using the PAO basis set in CONQUEST we have used a "single zeta" (SZ) set, and relaxed with both NSC-AITB and SC-AITB methodologies. For the B-splines, we have used a grid spacing equivalent to a plane wave basis with a 140 eV cutoff, using the full *ab initio* methodology. For SIESTA, which uses PAOs, we used a SZ set and a more fully converged DZP ("double zeta plus polarisation") set. For VASP, we used a plane-wave cutoff of 150 eV. All calculations used the same unit cell sizes and k-point sampling. The results are shown in Table 1. Note that the dimer angle is measured relative to the horizontal (i.e. the (001) surface).

Method	Basis	Bond length (Å)	Bond angle
Siesta(NSC)	SZ	2.50	15.9°
Conquest(NSC)	SZ	2.50	14.5°
Siesta(SC)	SZ	2.41 and 2.49	11.7° and 31.2°
Conquest(SC)	SZ	2.41 and 2.49	10.2° and 33.5°
Conquest(SC)	B-spline	2.37	22.8°
Siesta(SC)	DZP	2.40	19.9°
VASP	PW	2.41	19.7°

Table 1: Comparison of CONQUEST predictions with those of SIESTA and VASP for the relaxed structure of the reconstructed Si (001) surface (see Fig. 1). Comparisons of dimer bond length and buckling angle are given with SZ and B-spline basis sets in CONQUEST, and SZ and DZP basis sets in SIESTA; VASP used the conventional plane-wave basis set.

We see that the SZ basis sets, although somewhat crude, give useful semi-quantitative predictions. The inclusion of self consistency changes the buckling angle of the surface dimers, and shortens the dimer bond length; it also induces an unphysical distortion, in which the symmetry of the surface is lost, as shown by the two sets of results for both CONQUEST and SIESTA. The SIESTA results for the DZP basis show that this comes very close to plane-wave results. The CONQUEST B-spline results (full *ab initio*) show excellent agreement with the VASP results. We note that the exact consistency of forces and total energy in CONQUEST makes structural relaxation for this kind

of problem very straightforward.

We have also performed NSC-AITB $O(N)$ relaxations of the same system, and find a bond length of 2.42 Å and a dimer angle of 15°. These results are remarkably independent of cut-off on the L matrix (they remain essentially the same for $R_L > 7$ Å). We show an illustrative surface structure in Figure 1.

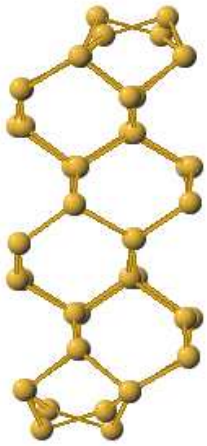

Figure 1: Relaxed structure of the Si (001) surface from $O(N)$ CONQUEST calculations.

5. DISCUSSION AND CONCLUSIONS

The development of practical and reliable $O(N)$ DFT techniques is a challenge that has been taken up by several research groups. We have argued here that there is a strong advantage in setting up these techniques so that they can be applied at different levels of precision, going from minimal-basis tight binding all the way through to full *ab initio*. We have also shown how the calculation of ionic forces can be performed consistently at all levels of this hierarchy, both in $O(N)$ and in diagonalisation modes of operation. Our illustrative CONQUEST calculations on the reconstruction of the Si (001) surface demonstrate excellent numerical agreement with other DFT methods, including the plane-wave technique. We are now applying the CONQUEST code to more ambitious large-scale problems, including the reconstruction of Ge overlayers on Si (001) and the adsorption of metal clusters on oxide surfaces.

ACKNOWLEDGMENT

The CONQUEST project is partially supported by ACT-JST. DRB is supported by a Royal Society University Research Fellowship, and RC by an EPSRC studentship.

REFERENCES

1) M. C. PAYNE, M. P. TETER, D. C. ALLAN, T. A. ARIAS, J. D. JOANNOPOULOS, Rev. Mod. Phys. **64** (1992) 1045.

2) W. KOHN, Phys. Rev. Lett. **76** (1996) 3168.

3) D. R. BOWLER, M. AOKI, C. M. GORINGE, A. P. HORSFIELD, D. G. PETTIFOR, Modell. Simul. Mater. Sci. Eng. **5** (1997) 199.

4) S. GOEDECKER, Rev. Mod. Phys. **71** (1999) 1085.

5) E. HERNANDEZ, M. J. GILLAN, Phys. Rev. B **51** (1995) 10157.

6) E. HERNÁNDEZ, M. J. GILLAN, C. M. GORINGE, Phys. Rev. B **53** (1996) 7147.

7) C. M. GORINGE, E. HERNÁNDEZ, M. J. GILLAN, I. J. BUSH, Comput. Phys. Commun. **102** (1997) 1.

8) D. R. BOWLER, M. J. GILLAN, Comput. Phys. Commun. **120** (1999) 95.

9) D. R. BOWLER, I. J. BUSH, M. J. GILLAN, Int. J. Quantum Chem. **77** (2000) 831.

10) D. R. BOWLER, T. MIYAZAKI, M. J. GILLAN, Computer Physics Communications **137** (2001) 255.

11) D.R.BOWLER, T.MIYAZAKI, M.J.GILLAN, J. Phys.:Condens. Matter **14** (2002) 2781.

12) F. MAURI, G. GALLI, R. CAR, Phys. Rev. B **47** (1993) 9973.

13) P. ORDEJÓN, D. DRABOLD, M. GRUMBACH, R. MARTIN. Phys. Rev. B **48** (1993) 14646.

14) X.-P. LI, R. W. NUNES, D. VANDERBILT, Phys. Rev. B **47** (1993) 10891.

15) A. D. DANIELS, J. M. MILLAM, G. E. SCUSERIA, J. Chem. Phys. **107** (1997) 425.

16) M. CHALLACOMBE, J. Chem. Phys. **110** (1999) 2332.

17) A. J. WILLIAMSON, R. Q. HOOD, J. C. GROSSMAN, Phys. Rev. Lett. **87** (2001) 246406.

18) P. ORDEJON, E. ARTACHO, J. M. SOLER, Phys. Rev. B **53** (1996) R10441.

19) J. M. SOLER, E. ARTACHO, J. D. GALE, A. GARCÍA, J. JUNQUERA, P. ORDEJÓN, D. SÁNCHEZ-PORTAL, J. Phys.:Condens. Matter **14** (2002) 2745.

20) P. D. HAYNES, M. C. PAYNE, Phys. Rev. B **59** (1999) 12173.

21) C. K. GAN, P. D. HAYNES, M. PAYNE, Phys. Rev. B **63** (2001) 205109.

22) T. OZAKI, K. TERAKURA, Phys. Rev. B **64** (2001) 195126.

23) J. L. FATTEBERT, J. BERNHOLC, Phys. Rev. B **62** (2000) 1713.

24) A. H. PALSER, D. E. MANOLOPOULOS, Phys. Rev. B **58** (1998) 12704.

25) R. MCWEENY, Rev. Mod. Phys. **32** (1960) 335.

26) D. D. JOHNSON, Phys. Rev. B **38** (1988) 12807.

27) J. HARRIS, Phys. Rev. B **31** (1985) 1770.

28) W. FOULKES, R. HAYDOCK, Phys. Rev. B **39** (1989) 12520.

29) O. F. SANKEY, D. J. NIKLEWSKI, Phys. Rev. B **40** (1989) 3979.

30) A. P. HORSFIELD, A. M. BRATKOVSKY, J. Phys.: Condens. Matter **12** (2000) R1.

31) J. KIM, F. MAURI, G. GALLI, Phys. Rev. B **52** (1995) 1640.

32) G. KRESSE, J. HAFNER, Phys. Rev. B **54** (54) 11169.