ONETEP: linear-scaling density-functional theory with plane-waves

P D Haynes[1], **A A Mostofi**[1], **C-K Skylaris**[2] **and M C Payne**[1]

[1]Theory of Condensed Matter, Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK

[2]Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QZ, UK

E-mail: pdh1001@cam.ac.uk

Abstract. This paper provides a general overview of the methodology implemented in ONETEP (Order-N Electronic Total Energy Package), a parallel density-functional theory code for large-scale first-principles quantum-mechanical calculations. The distinctive features of ONETEP are linear-scaling in both computational effort and resources, obtained by making well-controlled approximations which enable simulations to be performed with plane-wave accuracy. Titanium dioxide clusters of increasing size designed to mimic surfaces are studied to demonstrate the accuracy and scaling of ONETEP.

1. Introduction

Over the last decade, first-principles quantum-mechanical calculations based on density-functional theory (DFT) [1, 2] have become established as a means of gaining insight into atomic-scale processes in real materials. While these simulations will always be a complementary tool to experimental techniques rather than a replacement for them, their power is demonstrated by their ability not only to aid in the interpretation of experimental results, but also to lead experiment e.g. through the prediction of the properties of new materials not yet manufactured [3] or the behaviour of matter under conditions unattainable in the laboratory [4].

The plane-wave pseudopotential (PWP) approach [5, 6] has led the way in the development of these methods, being efficient and accurate. Nevertheless the scope of such calculations is limited by the scaling of the computational effort, which increases asymptotically as the cube of the size of the simulation. This means that an eight-fold increase in computing power yields a mere doubling of the number of atoms which can be treated, and restricts the methods to tackling problems involving a few hundred atoms at most. However the complexity of finding the quantum-mechanical ground-state in DFT increases only linearly with system-size. This has prompted the search for linear-scaling or order-N methods which would enable simulations of thousands of atoms to be performed routinely and open up new avenues of application for first-principles quantum-mechanical calculations in areas such as nanotechnology and biology.

This paper is organised as follows. In section 2 two equivalent reformulations of DFT are presented which lead to linear-scaling methods and in section 3 the particular approach implemented in the ONETEP code is outlined. Section 4 contains results and discussion for titanium oxide clusters designed to mimic surfaces to demonstrate the accuracy and scaling of ONETEP.

2. Linear-scaling formulations of DFT

In Kohn-Sham DFT [1, 2] the problem of solving the many-body Schrödinger equation for N interacting particles is reduced to solving N single-particle Schrödinger equations representing a system of N fictitious non-interacting particles with same ground-state density. The effective potential V_{eff} accounts for electron-electron interactions and therefore depends upon the particle density n, so that the following equations must be solved self-consistently:

$$\hat{H}_{\text{KS}}\psi_{j\mathbf{k}\sigma}(\mathbf{r}) = \left\{-\tfrac{1}{2}\nabla^2 + V_{\text{eff}}[n](\mathbf{r})\right\}\psi_{j\mathbf{k}\sigma}(\mathbf{r}) = \varepsilon_{j\mathbf{k}\sigma}\psi_{j\mathbf{k}\sigma}(\mathbf{r})\,, \tag{1}$$

$$n(\mathbf{r}) = \Omega_{\text{cell}} \sum_{j\sigma} \int_{1\text{BZ}} \frac{\mathrm{d}^3k}{(2\pi)^3}\, f_{j\mathbf{k}\sigma} \left|\psi_{j\mathbf{k}\sigma}(\mathbf{r})\right|^2\,, \tag{2}$$

where j, $\mathbf{k}$ and σ label the band index, $\mathbf{k}$-point and spin of the Bloch state ψ respectively. The $f_{j\mathbf{k}\sigma}$ are occupation numbers which at zero temperature must be either zero or unity, and which we will assume to be independent of $\mathbf{k}$ (no partially-filled bands) in the following. The integration over $\mathbf{k}$-space in (2) is over the first Brillouin zone (1BZ) and Ω_{cell} is the unit cell volume. Solutions to (1) may be obtained via a constrained Rayleigh-Ritz variational procedure.

It is immediately apparent that this formulation of DFT cannot yield a linear-scaling method. At least N solutions to (1) are required (those with the lowest eigenvalues $\varepsilon_{j\mathbf{k}\sigma}$) and each of these solutions will in general extend over a volume which is proportional to N. The set of solutions therefore contains an amount of information proportional to N^2, and any manipulation of this data must require a computational effort which at best scales in the same manner. However the requirement that the solutions be mutually orthogonal results in the asymptotic N^3 scaling.

Wannier functions [7] are defined by a unitary transformation (Fourier series) of the Bloch states which maintains orthogonality,

$$w_{j\mathbf{R}\sigma}(\mathbf{r}) = \sqrt{\tfrac{\Omega_{\text{cell}}}{(2\pi)^3}} \int_{1\text{BZ}} \mathrm{d}^3k\, \exp(-\mathrm{i}\mathbf{k}\cdot\mathbf{R})\, \psi_{j\mathbf{k}\sigma}(\mathbf{r})\,, \tag{3}$$

where $\mathbf{R}$ is a lattice vector labelling a periodic image of the unit cell in which the Wannier function is centred. This transformation may be generalised by allowing a further unitary transformation among the bands, which can be different at each $\mathbf{k}$-point.

The single-particle density-matrix is constructed from the Bloch states or Wannier functions:

$$\rho(\mathbf{r},\mathbf{r}') = \Omega_{\text{cell}} \sum_{j\sigma} f_{j\sigma} \int_{1\text{BZ}} \frac{\mathrm{d}^3k}{(2\pi)^3}\, \psi_{j\mathbf{k}\sigma}(\mathbf{r})\psi^*_{j\mathbf{k}\sigma}(\mathbf{r}') = \sum_{j\sigma} f_{j\sigma} \sum_{\mathbf{R}} w_{j\mathbf{R}\sigma}(\mathbf{r}) w^*_{j\mathbf{R}\sigma}(\mathbf{r}') \tag{4}$$

and the equivalent condition to the orthogonality of the Bloch states and Wannier functions is idempotency i.e. $\rho^2 = \rho$. In contrast to the extended Bloch states, both the Wannier functions and density-matrix reflect the locality or "nearsightedness" [8] of quantum mechanics (that the observable properties of a given region depend only weakly on perturbations in a spatially distant region) through their decay properties which are related via (4). Both analytical [9, 10] and numerical [11] studies have established that for insulating systems both quantities decay exponentially, whereas the decay in metals is algebraic. See the review by Goedecker [12] for more detail.

Exponential decay is the key to linear-scaling methods: although the quantity of information in the set of Wannier functions and density-matrix is still proportional to N^2, the information which is significant in determining the observable properties of a system scales only linearly with N. One can therefore exploit this decay and work directly with a set of Wannier functions or density-matrix which has been truncated to retain only the physically significant information. The equivalent equations to (1,2) can then be solved using a variational procedure, imposing the condition of orthogonality on the Wannier functions [13, 14, 15, 16, 17] or idempotency on the density-matrix [18, 19, 20]. A comparison of these methods can be found in [12].

3. The ONETEP approach

In addition to linear scaling, a prerequisite in the design of ONETEP was to provide controlled accuracy. The ONETEP method [21] is formulated in terms of the density-matrix which, along the lines of [22], is represented in separable form,

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta\sigma\sigma'} \phi_{\alpha\sigma}(\mathbf{r}) K^{\alpha\beta}_{\sigma\sigma'} \phi_{\beta\sigma'}(\mathbf{r}')\,, \tag{5}$$

in terms of a density kernel K and a set of local orbitals ϕ, also known as support functions but which are referred to as nonorthogonal generalised Wannier functions (NGWFs) [23] in ONETEP to reflect the relationship expressed in (4).

Truncation of the density-matrix is achieved by imposing two spatial cut-offs. The first is applied to the NGWFs, which are strictly localised within atom-centred spherical regions. The second is applied to the density kernel, which is made sparse by setting to zero those matrix elements corresponding to NGWFs centred on atoms which are far apart. These approximations are independently controlled by varying the radii of the spherical regions and the cut-off distance for the density kernel. The shorter these cut-offs, the more information is discarded from the density-matrix and the faster the calculation, albeit at the expense of restricting the variational freedom and therefore the accuracy of the solutions. In practice the cut-offs are increased until physical properties such as the total energy are converged to an acceptable degree.

In addition to the spatial cut-offs, the accuracy is also controlled by the quality of the basis set used to expand the NGWFs. In other methods real-space grids with finite-difference [22, 24] or multigrid [25, 26] techniques, B-splines or blip functions [27], localised spherical-waves [28], numerical atomic orbitals [29, 30, 31, 32] and Gaussians [33, 34] have been proposed. In ONETEP a set of periodic cardinal sine or PSINC functions [23, 35] are used which are equivalent to a set of plane-waves and which therefore inherit their accuracy (particularly with regard to the kinetic energy [36, 37]) and the ability to improve the basis set completeness systematically via a single parameter, the energy cut-off.

Linear scaling in ONETEP is obtained through the use of fast Fourier transforms (FFTs) within a box whose size does not increase with system-size, but is related only to the radii of the NGWF spheres. This "FFT box" technique has been shown to be a highly accurate approximation [23, 38]. In addition, the FFT box gives an indication of the likely cross-over of ONETEP with traditional PWP codes such as CASTEP [39] i.e. the system-size at which the linear-scaling method beats the traditional method (which will always be faster for small systems). Both methods invest a large fraction of the total computational effort in performing FFTs. In the case of ONETEP these are nearly all done in the FFT box, whereas in CASTEP they are all done in the whole simulation cell. In ONETEP the FFT box is typically $20 \times 20 \times 20$ Å^3 and these dimensions do not increase with system-size and are generally the same for a wide variety of systems. This corresponds to the volume occupied by a few hundred atoms in a solid, but the cross-over is far more favourable for less densely-packed systems such as biological molecules, nanotubes or systems involving surfaces. In CASTEP, vacuum requires the same level of description as regions where atomic binding occurs. However in ONETEP, where there are no atoms there are no FFT boxes and a saving is made. This can reduce the cross-over by an order of magnitude for isolated molecules, clusters and polymers.

The optimisation of both the density kernel and the NGWFs is achieved through a combination of a density-matrix minimisation method [20] based on the purifying transformation of McWeeny [40] and a penalty-functional method [41]. The NGWF optimisation also uses a preconditioning scheme [35, 42] which ensures that the number of iterations in the solution of (1,2) does not increase with system-size so that the computational effort for the whole calculation is proportional to N. As a result of the distribution of atomic and simulation cell data (charge density and potential) ONETEP also exhibits excellent scaling on parallel computers [21].

4. Titanium oxide clusters

In this section the application of ONETEP to titanium oxide clusters designed to represent surfaces is considered. In traditional PWP methods, surfaces are modelled using slabs constructed in periodic supercells. In addition to ensuring that the slab contains sufficient layers the slabs must also be adequately separated to avoid spurious interactions between periodic images. While this option is also possible with ONETEP (which can in fact accommodate large vacuum regions between slabs while incurring minimal additional computational expense) the alternative approach of modelling surfaces by isolated clusters is also possible.

Clusters ranging in size from ten to 200 atoms have been studied using norm-conserving pseudopotentials [43] in separable form [44], a local density approximation [45, 46] for exchange and correlation and an 800 eV cut-off for the PSINC basis set. The NGWF radii for titanium and oxygen were all set at 3.175 Å and four NGWFs were associated with each atom.

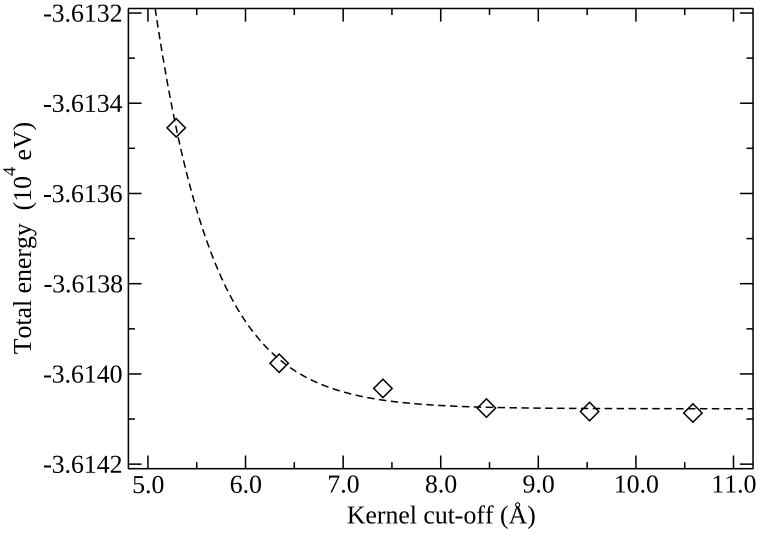

Figure 1. Convergence of the total energy of a $Ti_{38}O_{76}$ cluster with respect to the density kernel cut-off

Figure 1 shows how the total energy of a $Ti_{38}O_{76}$ cluster converges as the density kernel cut-off is increased. The trend line shown is a best fit to an exponential decay which reflects the expected behaviour of the density-matrix, although the convergence in practice is not smooth but occurs in jumps as particular matrix elements are included when the cut-off matches the distance between the relevant atoms. In this system, the total energy is converged to 10^{-3} eV per atom for a kernel cut-off of 8 Å.

Table 1. Comparison of the energetics of a CO molecule located at a distance d from a Ti_3O_6 cluster calculated with ONETEP and CASTEP

	Total energy (eV)		Energy difference (eV)
	$d = 2.000$ Å	$d = 2.252$ Å	
CASTEP	-3419.694	-3419.355	0.339
ONETEP	-3415.708	-3415.361	0.347
% error	0.1	0.1	2.3

In order to assess the accuracy of ONETEP in determining physical properties such as binding energies, the interaction of a carbon monoxide molecule with a small Ti_3O_6 cluster was studied. Table 1 shows the energies obtained from two configurations illustrated in figure 2. The difference in the total energies calculated by ONETEP and CASTEP is about 0.4 eV per atom but is mostly accounted for by the variational restriction originating from the truncation radii chosen for the NGWFs. Equivalent cut-offs were used for the basis sets and identical pseudopotentials were used. The only difference between the calculations was the use of a $12 \times 12 \times 12$ Å^3 cell for CASTEP but a $26 \times 26 \times 26$ Å^3 cell for ONETEP. When the energy differences between two configurations are compared the error is less than 1 meV per atom and this demonstrates the fact that physical properties (which depend upon energy differences rather than absolute total energies) can be calculated accurately even when total energies are not fully converged.

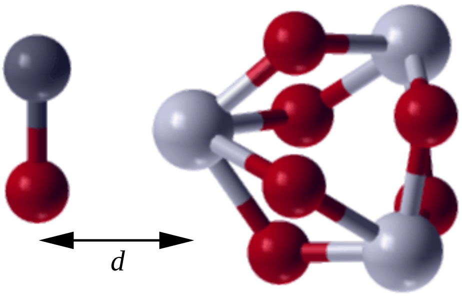

Figure 2. Illustration of the system used to study the interaction of a CO molecule and a Ti_3O_6 cluster

Finally the linear scaling of the ONETEP is demonstrated in figure 3. The time per iteration (on a Sun Fire V40z server with four 2.2 GHz single-core Opteron CPUs) scales linearly with system-size in accord with results on other systems [21]. The number of iterations fluctuated by up to 20% but did not change systematically with system-size. These results therefore support the claim of linear-scaling with system-size for the entire computational effort.

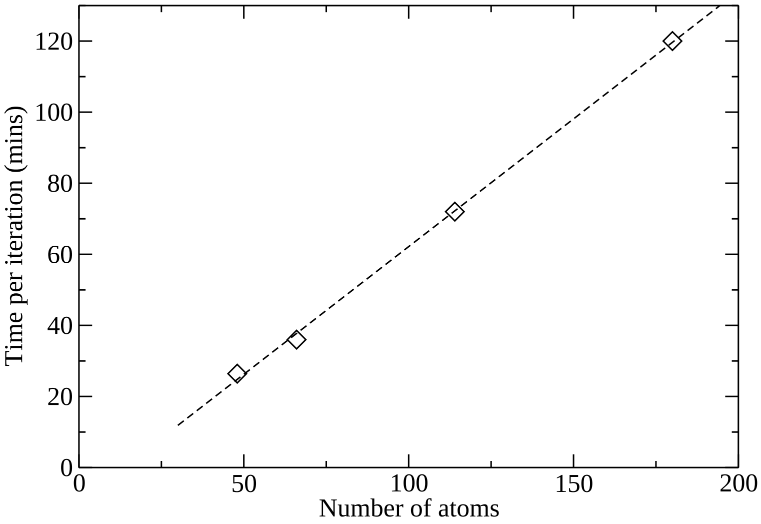

Figure 3. Scaling of the time per iteration against system-size represented by the number of atoms in the cluster

Acknowledgments

PDH acknowledges the support of Sidney Sussex College, Cambridge and AAM similar support from Christ's College, Cambridge. C-KS was supported by a Royal Society University Research Fellowship. The development of ONETEP was greatly helped by resources at the Cambridge-Cranfield High Performance Computing Facility and equipment grants from the Royal Society for PDH and AAM.

References

[1] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** 864

[2] Kohn W and Sham L J 1965 *Phys. Rev.* **140** 1133

[3] Côté M, Haynes P D and Molteni C 2002 *J. Phys.: Condens. Matter* **14** 9997

[4] Alfè D, Gillan M J and Price G D 1999 *Nature* **401** 462

[5] Ihm J, Zunger A and Cohen M L 1979 *J. Phys.* C **12** 4409

[6] Denteneer P J H and van Haeringen W 1985 *J. Phys.* C **18** 4127

[7] Wannier G 1937 *Phys. Rev.* **52** 191

[8] Kohn W 1996 *Phys. Rev. Lett.* **76** 3168

[9] Kohn W 1959 *Phys. Rev.* **115** 809

[10] des Cloizeaux J 1964 *Phys. Rev.* **135** A685 and A698

[11] Ismail-Beigi S and Arias T 1999 *Phys. Rev. Lett.* **82** 2127

[12] Goedecker S 1999 *Rev. Mod. Phys.* **71** 1085

[13] Mauri F, Galli G and Car R 1993 *Phys. Rev.* B **47** 9973

[14] Ordejón P, Drabold D, Grumbach M and Martin R 1993 *Phys. Rev.* B **48** 14646

[15] Mauri F and Galli G 1994 *Phys. Rev.* B **50** 4316

[16] Ordejón P, Drabold D A, Martin R M and Grumbach M P 1995 *Phys. Rev.* B **51** 1456

[17] Kim J, Mauri F and Galli G 1995 *Phys. Rev.* B **52** 1640

[18] Li X-P, Nunes R W and Vanderbilt D 1993 *Phys. Rev.* B **47** 10891

[19] Daw M S 1993 *Phys. Rev.* B **47** 10895

[20] Nunes R W and Vanderbilt D 1994 *Phys. Rev.* B **50** 17611

[21] Skylaris C-K, Haynes P D, Mostofi A A and Payne M C 2005 *J. Chem. Phys.* **122** 084119

[22] Hernández E and Gillan M J 1995 *Phys. Rev.* B **51** 10157

[23] Skylaris C-K, Mostofi A A, Haynes P D, Diéguez O and Payne M C 2002 *Phys. Rev.* B **66** 035119

[24] Hernández E, Gillan M J and Goringe C M 1996 *Phys. Rev.* B **53** 7157

[25] Fattebert J L and Bernholc J 2000 *Phys. Rev.* B **62** 1713

[26] Nardelli M B, Fattebert J L and Bernholc J 2001 *Phys. Rev.* B **64** 245423

[27] Hernández E, Gillan M J and Goringe C M 1997 *Phys. Rev.* B **55** 13485

[28] Haynes P D and Payne M C 1997 *Comput. Phys. Commun.* **102** 17

[29] Sankey O F and Niklewski D 1989 *Phys. Rev.* B **40** 3979

[30] Kenny S D, Horsfield A P and Fujitani H 2000 *Phys. Rev.* B **62** 4899

[31] Junquera J, Paz Ó, Sánchez-Portal D and Artacho E 2001 *Phys. Rev.* B **64** 235111

[32] Anglada E, Soler J M, Junquera J and Artacho E 2002 *Phys. Rev.* B **66** 205101

[33] White C A, Johnson B G, Gill P M W and Head-Gordon M 1996 *Chem. Phys. Lett.* **253** 268

[34] Strain M C, Scuseria G E and Frisch M J 1996 *Science* **271** 51

[35] Mostofi A A, Haynes P D, Skylaris C-K and Payne M C 2003 *J. Chem. Phys.* **119** 8842

[36] Skylaris C-K, Mostofi A A, Haynes P D, Pickard C J and Payne M C 2001 *Comput. Phys. Commun.* **140** 315

[37] Skylaris C-K, Diéguez O, Haynes P D and Payne M C 2002 *Phys. Rev.* B **66** 073103

[38] Mostofi A A, Skylaris C-K, Haynes P D and Payne M C 2002 *Comput. Phys. Commun.* **147** 788

[39] Segall M D, Lindan P J D, Probert M J, Pickard C J, Hasnip P J, Clark S J and Payne M C 2002 *J. Phys.: Condens. Matter* **14** 2717

[40] McWeeny R 1960 *Rev. Mod. Phys.* **32** 335

[41] Haynes P D and Payne M C 1999 *Phys. Rev.* B **59** 12173

[42] Teter M P, Payne M C and Allan D C 1989 *Phys. Rev.* B **40** 12255

[43] Hamann D R, Schlüter M and Chiang C 1979 *Phys. Rev. Lett.* **43** 1494

[44] Kleinman L and Bylander D M 1982 *Phys. Rev. Lett.* **48** 1425

[45] Ceperley D M and Alder B J 1980 *Phys. Rev. Lett.* **45** 566

[46] Perdew J P and Zunger A 1981 *Phys. Rev.* B **23** 5048