A divide-and-conquer linear scaling three-dimensional fragment method for large scale electronic structure calculations

2008 J. Phys.: Condens. Matter 20 294203

(http://iopscience.iop.org/0953-8984/20/29/294203)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **20** (2008) 294203 (8pp) doi:10.1088/0953-8984/20/29/294203

A divide-and-conquer linear scaling three-dimensional fragment method for large scale electronic structure calculations

Zhengji Zhao[1], **Juan Meza and Lin-Wang Wang**[2]

Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

E-mail: lwwang@lbl.gov

Received 24 January 2008, in final form 13 February 2008
Published 24 June 2008
Online at stacks.iop.org/JPhysCM/20/294203

Abstract

We present a new linear scaling *ab initio* total energy electronic structure calculation method based on a divide-and-conquer strategy. This method is simple to implement, easy to parallelize, and produces accurate results when compared with direct *ab initio* methods. The new method has been tested on nanosystems with up to 15 000 atoms using up to 8000 processors.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

In theoretical material science and nanoscience research, there are many cases where thousands or tens of thousands of atoms need to be simulated. It could be a molecular dynamics simulation for water absorption on a metal oxide surface, or catalytic processes for nanostructure growth. It could also be charge density self-consistent calculations for a large nanocrystal with tens of thousands of atoms for determining its internal electric field and surface state coupling to the internal states. For all of these calculations, *ab initio* density functional theory (DFT), especially with the use of the local density approximation (LDA), is the method of choice. However, due to the O(N^3) computational scaling [1] of the direct LDA method, it can only be applied to about one to two thousand atoms, even for the largest supercomputers available today [2]. In addition, future increases in supercomputer power will be due more to an ever larger number of processors and computing cores, instead of increasing CPU speeds. Unfortunately, the parallelization of direct DFT methods might have a limit on the order of 10 000 processors, due to a communication bottleneck [2]. Thus, both the total computational cost and the limits of parallelization call for a change in the direct LDA algorithm to linear scaling O(N) methods [3] for large system simulations. Indeed, for systems with more than 500 atoms most of the O(N) methods already become faster than direct LDA methods.

Currently, there are many O(N) methods, and many of them are discussed in this special journal issue. A common O(N) algorithm is based on localized orbitals [4]. However, there are some technical difficulties in this approach. One difficulty is the possible existence of local minima in the total energy function introduced by the restriction of the wavefunctions on the local orbital manifold. Obviously these local minima could cause numerical convergence problems. Special methods and algorithms have been devised to overcome these problems [5]. Another issue is that most of the local orbital methods are naturally represented by a localized basis set (either atomic basis set or real space grids). It is therefore not straightforward to use plane-wave basis sets to represent the localized orbitals [6]. Compared to real space grids [7], the plane-wave basis and its convergence properties have been more thoroughly studied, and it is more widely used in material science simulations. On the computational side, the overlap between neighboring local orbitals poses a challenge for code parallelization. A closely related method to the localized orbital method is the truncated density matrix method [8]. The truncated density matrix method might have avoided the local minimum problem, but it is costly to represent the matrix based on real space grids. As a result, it is mostly represented by atomic orbitals. Another approach is to first construct a localized basis set from a plane-wave or real space grid basis, then use this localized basis set to

[1] Present address: National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA.
[2] Author to whom any correspondence should be addressed.

represent the density matrix. In the energy minimization, one then optimizes the density matrix coefficients and the localized basis set simultaneously. This approach is exemplified by the CONQUEST project [9].

Another approach for realizing O(N) scaling is a divide-and-conquer method. In a divide-and-conquer method, a large system is divided into small pieces (fragments), and each fragment is calculated independently. The fragment results are placed together to give the total energy and the charge density of the whole system. The critical issue here is how to put the fragments together without introducing artificial boundary effects. One of the earliest methods in this approach was introduced by Yang [10]. In Yang's method, spatial partition functions are applied to the charge densities of the fragments when generating the charge density of the whole system. Thus, only the central parts of the fragment charge densities (and their corresponding kinetic energy density) are used. One technical problem is that the total energy cannot be expressed in a variational formula. Furthermore, in order to reach charge neutrality, a global Fermi energy has to be used for the occupation of the fragment wavefunctions, thus allowing charge transfer between fragments. In this regard, the fragments are treated almost like metallic pieces. Another issue is how to partition the kinetic energy where different ways of partitioning might lead to different results [11]. Nevertheless, this divide-and-conquer method has been used to calculate systems with tens of thousands of atoms in molecular dynamics simulations [11].

In this paper, we present a new O(N) method based on the divide-and-conquer approach. We call it a linear scaling three-dimensional fragment (LS3DF) method. This method has been reported briefly in a previous paper [12], but here more technical details will be provided. Compared with Yang's method, we used a different scheme to patch the fragment charge densities. Instead of using spatial partition functions, we use positive and negative fragments. By judiciously combining the positive and negative fragments, the artificial boundary effects will cancel out. Our method has the following features: (1) its accuracy increases exponentially with the fragment size, and very accurate results can be obtained with relatively small fragments; (2) its formalism and implementation are straightforward. It can be implemented easily using an existing *ab initio* code; (3) since the fragment wavefunction calculations are independent for different fragments, it can be parallelized easily; (4) it can be applied to *ab initio* methods other than DFT.

In this paper, we have only tested our method on covalently bonded systems with a band gap. One feature in our current implementation of the method is that after some artificial surface passivation, each fragment is an insulating system with a band gap. This can be easily done by using partial charge pseudo-hydrogen atoms for covalently bonded systems (where a pseudo-hydrogen atom will be placed at the center of a cut-off bond). Although further tests are needed, we do expect the method will work for ionic systems with band gaps. Usually artificial atoms can be used to provide the electrons to fill the ionic close shell orbitals and at the same time to provide local charge neutrality with its nuclei charge. For example, we have tried this for an Ag_2S ionic crystal for different surfaces; the use of an artificial hydrogen atom passivates the surfaces well. A more interesting case will be for metallic systems. There surface passivation is not necessary, but the influence of the surface can have a longer range. Actual tests in the future are necessary to determine how accurate this method can be for metallic systems.

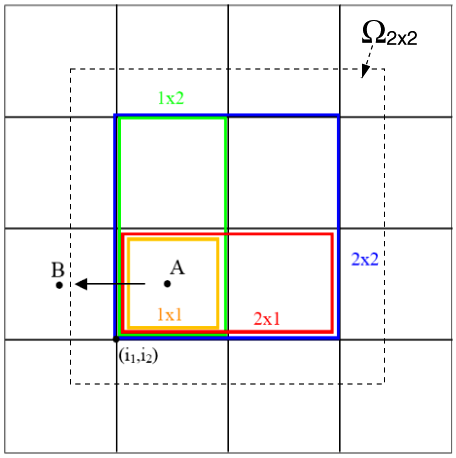

Figure 1. A schematic view of the division of the space into fragments. In this figure, $m_1 \times m_2 = 4 \times 4$.

2. Formalism

Like many other O(N) methods, our method is based on the nearsightedness of the quantum mechanical effects [13]. The central premise is that the total energy of a given system can be split into two parts: the electrostatic energy part and the quantum mechanical energy part (e.g, the kinetic energy and exchange correlation energy). While the electrostatic energy is long range and must be solved via a global Poisson equation, the computationally expensive quantum mechanical energy is short range [13] and can be solved locally. Thus, the system can be divided into small fragments, and the quantum mechanical energy can be obtained through the summation of these fragments, while the electrostatic energy can be calculated from the total charge density of the whole system. Our special division and patching scheme is illustrated in figure 1 using two dimensions for clarity. In figure 1, a two-dimensional periodic supercell is divided into $m_1 \times m_2$ small pieces. At each $m_1 \times m_2$ grid point (i_1, i_2), we can define four fragments with their sizes being 1×1, 1×2, 2×1 and 2×2, respectively. If we use S to denote these sizes, then each fragment F can be specified as (i_1, i_2, S). Now, if quantum energies E_F and charge densities $\rho_F(r)$ of all the fragments F have been calculated, then the total quantum energy of the whole system is calculated as $E = \sum_F \alpha_F E_F$, and the total charge density as $\rho(r) = \sum_F \alpha_F \rho_F(r)$. Here, $\alpha_F = \pm$ is the sign of the fragment F. $\alpha_F = 1$ if its corresponding $S = 1 \times 1$ or 2×2, and $\alpha = -1$ if its corresponding $S = 1 \times 2$ or 2×1.

To understand the above formula, we can check each point inside a fragment (point A in figure 1). Note that each spatial

point will be included in 3^2 fragments: four 2×2 fragments, two 2×1 fragment, two 1×2 fragments and one 1×1 fragment. After the above $\pm$ cancellations, it will be covered by only one fragment, which is what is needed to contribute to one copy of the whole system. We can now check for each boundary point. A boundary can be defined with a direction, thus a boundary from A to B will be different from a boundary from B to A. We have used an arrow in figure 1 to represent a directional boundary (from A to B). A point on this directional boundary is covered by six fragments (two 2×2, two 1×2, one 1×2 and one 1×1 fragments), with equal numbers of positive and negative signs. Since all these pieces have the same (directional) boundary at this point, and given the nearsightedness, their charge density will be the same near this point. As a result, the boundary effects will cancel out. The same is true for the corner effect.

The above scheme can be extended to a three-dimensional system in a straightforward way. Here, a periodic supercell is divided into $m_1 \times m_2 \times m_3$ fragments, and from each grid point corner (i_1, i_2, i_3) there are eight fragments, with sizes $S(\alpha_S)$ equal to $1\times1\times1(-)$, $1\times1\times2(+)$, $1\times2\times1(+)$, $2\times1\times1(+)$, $1\times2\times2(-)$, $2\times1\times2(-)$, $2\times2\times1(-)$, and $2\times2\times2(+)$. The same formula $E = \sum_F \alpha_F E_{\rm F}$ and $\rho(r) = \sum_F \alpha_F \rho_F(r)$ can be used, where the fragment $F = (i_1, i_2, i_3, S)$, and this formula has the same property of canceling out all the surface, edge and corner effects.

In the above scheme (figure 1), we have used the size of two grid points in the $m_1 \times m_2 \times m_3$ grid for the $2 \times 2 \times 2$ fragment in each direction. Actually, it is possible to use a smaller size for the '$2 \times 2 \times 2$' fragment. Then the smaller fragments can be defined as the overlapping areas of the '$2 \times 2 \times 2$' fragments originated from neighboring grid point (i_1, i_2, i_3). Further tests are needed to find out whether this will save computing time for the same accuracy of calculations.

To carry out our LS3DF scheme shown in figure 1, we first divide our three-dimensional supercell into an $M = m_1 \times m_2 \times m_3$ grid. Atoms located inside a grid cube are assigned to the corresponding fragments. Thus, the division of the whole system into fragments is automatic. No prior knowledge about the physical system is necessary. As a result, sometimes one fragment can consist of a few physically separated clusters. But this will not affect the applicability and accuracy of this method. One critical point is to passivate the artificial surface created by this division, so that each fragment is still an insulating system. For covalently bonded systems, this can be done automatically by placing pseudo-hydrogen atoms with partial charges at the centers of the cut-off bonds [15]. This automatic procedure works well for all covalent bonding systems. A fragment is defined by the atoms (including the passivation atoms) plus a buffer vacuum region as indicated in by the dashed lines in figure 1. The fragment spatial domains can be denoted as Ω_F, where $F = (i_1, i_2, i_3, S)$ is the fragment index. Each fragment can be treated as an open system, or a periodic system with a periodic cell Ω_F. Due to the vacuum buffer region, there is no difference whether we treat it as an open system or a periodic system. In our method, we will treat it as a periodic system so that plane-wave basis can be used to describe the fragment wavefunctions. The fragment wavefunction $\psi_{F,i}(r)$s are defined only within the fragment domain Ω_F, and i is the wavefunction index.

Now we can write the total energy $E_{\rm tot}$ of the system as a variational form in terms of the fragment wavefunctions $\psi_{F,i}(r)$:

$$E_{\rm tot} = \sum_F \alpha_F \sum_i O(\epsilon_{F,i}, E_{\rm F}) \int \psi^*_{F,i}(r)[-\tfrac{1}{2}\nabla^2]\psi_{F,i}(r){\rm d}^3r + V_{\rm ion}(r)\rho_{\rm tot}(r){\rm d}^3r + \frac{1}{2}\int \frac{\rho_{\rm tot}(r)\rho_{\rm tot}(r')}{|r-r'|}{\rm d}^3r{\rm d}^3r' + \int \epsilon_{xc}(\rho_{\rm tot}(r))\rho_{\rm tot}(r){\rm d}^3r + \sum_F \alpha_F \int \Delta V_F(r)\rho_F(r){\rm d}^3r \tag{1}$$

where the total charge density $\rho_{\rm tot}$ is calculated as

$$\rho_{\rm tot}(r) = \sum_F \alpha_F \rho_F(r), \tag{2}$$

and the fragment charge density $\rho_F(r)$ is calculated as

$$\rho_F(r) = \sum_i O(\epsilon_{F,i}, E_{\rm F})|\psi_{F,i}(r)|^2 \qquad \text{for } r \in \Omega_F \tag{3}$$

where $O(\epsilon_{F,i}, E_{\rm F})$ is the Fermi–Dirac occupation function based on the overall Fermi energy $E_{\rm F}$ and the fragment wavefunction eigenenergy $\epsilon_{F,i}$.

In equation (1), $V_{\rm ion}(r)$ is the total ionic potential. The term $\Delta V_F(r)$ is an additional surface passivation potential that is only nonzero near the boundary of the fragment. For different fragments sharing the same boundary B, their $\Delta V_F(r)$ at that boundary B should be the same. Due to the fragment cancellations (the $\sum_{F'} \alpha_{F'}\rho_{F'}(r)$ for F' sharing the same boundary B should be small), the net value of the last term in equation (1) should be small. The amplitude of this term can be used as a measure for the accuracy of this method.

The total energy $E_{\rm tot}$ is a variational minimum (or maximum, depending on the sign of α_F) with regard to $\psi_{F,i}(r)$, subject to the orthonormal constraints:

$$\int_{\Omega_F} \psi^*_{F,i}(r)\psi_{F,j}(r){\rm d}^3r = \delta_{i,j}. \tag{4}$$

As a result, we can derive the fragment Kohn–Sham equation from $\delta E_{\rm tot}/\delta\psi^*_{F,i}(r) = \alpha_F O(\epsilon_{F,i}, E_{\rm F})\epsilon_{F,i}\psi_{F,i}(r)$, which gives us

$$[-\tfrac{1}{2}\nabla^2 + V_F(r)]\psi_{F,i}(r) = \epsilon_{F,i}\psi_{F,i}(r), \tag{5}$$

and

$$V_F(r) = V_{\rm tot}(r) + \Delta V_F(r) \qquad \text{for } r \in \Omega_F, \tag{6}$$

where $V_{\rm tot}(r)$ is the usual LDA total potential calculated from $\rho_{\rm tot}(r)$ by solving a global Poisson equation for the whole system. The global charge density self-consistency can be achieved iteratively using the usual potential mixing scheme [1] for $V_{\rm tot}(r)$. The overall computational flow is illustrated in figure 2 in comparison with the direct LDA method. We have used a plane-wave expansion for the

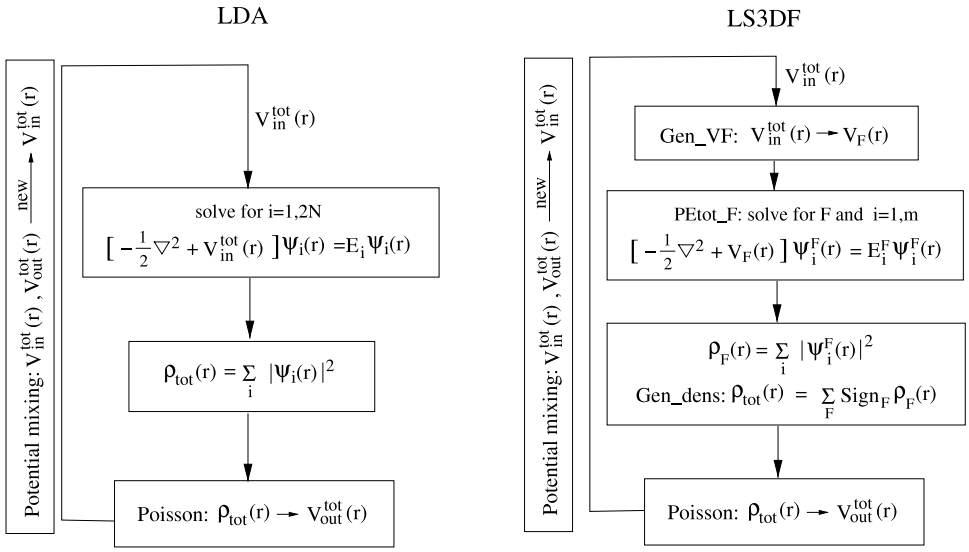

Figure 2. The computational flow charts for the direct LDA method and the LS3DF method. In the LS3DF method, the first box corresponds to equation (6) in the text, the second box corresponds to equation (5), and the third box corresponds to equations (2) and (3). The self-consistent iteration potential mixing schemes in LDA and LS3DF are the same.

wavefunctions $\psi_{F,i}(r)$ and norm conserving pseudopotentials for the Hamiltonian. Equation (5) (the second box in the LS3DF flow chart in figure 2) is solved using a conjugated gradient method based on the plane-wave code, PEtot [16]. After the charge self-consistency is reached, due to the variational principle, atomic forces can be calculated using the Hellman–Feynman theory. Of practical importance is the observation that the calculations of equation (5), the computationally most expensive step in figure 2, can be carried out independently for each fragment, which makes the overall computation trivially parallelizable. In the above formalism, we have used a Fermi–Dirac occupation function $O(\epsilon_{F,i}, E_F)$ in the summation over fragment wavefunction index i. This is necessary if the overall system is metallic. However, if the system is an insulator, and with proper surface passivation each fragment is also an insulator, then $O(\epsilon_{F,i}, E_F)$ is a sharp step function for index i, with N_F ($2N_F$ being the number of electrons in a fragment) occupied states, and the rest of the states unoccupied. In this case, $O(\epsilon_{F,i}, E_F)$ does not depend sensitively on E_F, and the summation over i with $O(\epsilon_{F,i}, E_F)$ can be replaced by a summation over i up to N_F without the use of $O(\epsilon_{F,i}, E_F)$.

One technical issue in our method is how to calculate the passivation potential $\Delta V_F(r)$ (so the resulting fragments remain insulators). The pseudo-hydrogen atoms placed at the surface of a fragment will make the fragment an insulator with a fragment potential $V_F(r)$ if the fragment is calculated self-consistently by itself. However, what we need in equation (6) is to change the total potential $V_{tot}(r)$ into the fragment potential $V_F(r)$ by adding a surface passivation potential $\Delta V_F(r)$. Thus, $\Delta V_F(r)$ is not just the hydrogen potential, it has to be something more which can change the $V_{tot}(r)$ in the buffer region into a vacuum-like potential. To solve this problem, we have used the sum of atomic charge densities to construct a (non-self-consistent) $\rho_{F,atom}(r)$, $\rho_{tot,atom}(r)$. From these charge densities and atomic pseudopotentials, we can calculate the corresponding $V_{F,atom}(r)$ for the fragment LDA potential, and $V_{tot,atom}(r)$ for the total system LDA potential using the LDA formula. Based on this potential, we have calculated the surface passivation potential as

$$\Delta V_F(r) = V_{F,atom}(r) - V_{tot,atom}(r) \qquad \text{for } r \in \Omega_F. \quad (7)$$

Furthermore, to assure that $\Delta V_F(r)$ at a given boundary B is the same for all the fragments sharing this common boundary, we have taken the average among all the fragments sharing this boundary. A typical passivation potential $\Delta V_F(r)$ is shown in figure 3. Also shown in figure 3 are the fragment potential $V_F(r)$ and the global total potential $V_{tot}(r)$ as in equation (6). Note that in an actual self-consistent calculation $\Delta V_F(r)$ is fixed throughout the calculation, and the generation of $\Delta V_F(r)$ does not take much time.

As discussed in the introduction, our approach is similar to Yang's divide-and-conquer method in that they both divide the system into small pieces in three dimensions. Our LS3DF method can also be compared to the fragment molecular orbital (FMO) method [14]. FMO is specifically designed for biological systems where a long molecule chain is divided into many small segments (monomers). In the FMO method, all monomers and monomer–monomer pairs are calculated to take into account the artificial effects caused by breaking the covalent bonds between the neighboring monomers. The breaking of the chain molecule to monomers is basically an one-dimensional event. As a result, only monomer–monomer pairs need to be calculated. In contrast, in LS3DF we have three-dimensional fragments with different sizes in a spatially compact form. If we identify our smallest $1 \times 1 \times 1$ fragment with the monomers in FMO, then we have calculated up to eight monomer clusters (the $2 \times 2 \times 2$ fragments). As a result of the spatial division (instead of focusing on molecule chains), the LS3DF has a rigorous cancelation of the boundary effects. As we will show, the error in LS3DF drops rapidly as the fragment size increases.

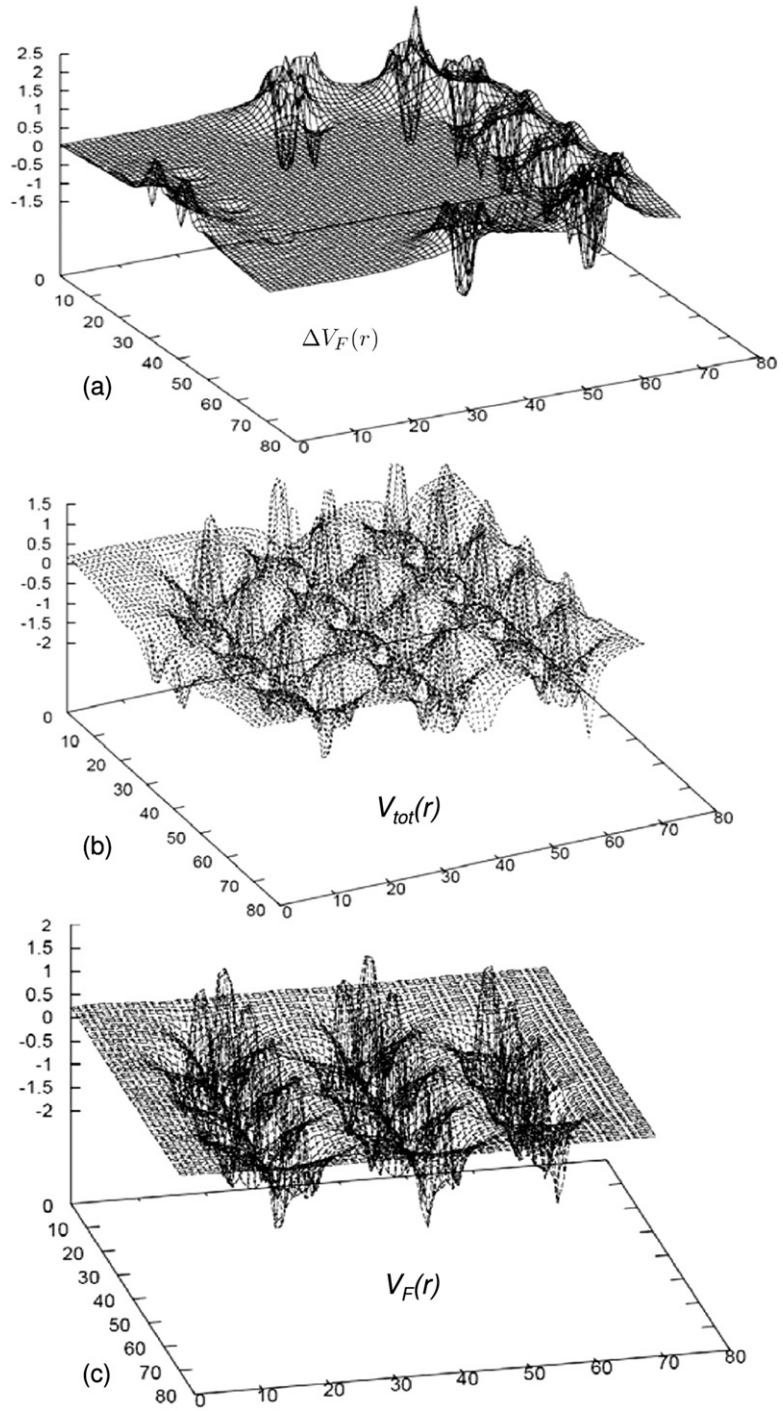

Figure 3. The surface passivation potential $\Delta V_F(r)$ (a); the whole system total potential $V_{\rm tot}(r)$ (showing only the portion for $r \in \Omega_F$) (b); and the fragment potential $V_F(r)$ (c). Note that $\Delta V_F(r) + V_{\rm tot}(r) = V_F(r)$, and in the self-consistent iterations $\Delta V_F(r)$ is fixed while $V_{\rm tot}(r)$ and $V_F(r)$ keep changing.

3. Numerical results

As a numerical test, we first show a comparison between the LS3DF method and the direct LDA method. We use a $Si_{235}H_{104}$ quantum dot (QD) with surface hydrogen passivation. Norm conserving pseudopotentials are used and the plane-wave basis set cutoff is 35 Ryd. The size of the $1 \times 1 \times 1$ grid space is a, the lattice constant of the diamond structure. Thus, the smallest fragment has eight Si atoms, and a 0.5 a buffer space on each side. The total energy difference between the LS3DF method and the direct LDA method is 3 meV/atom. The average charge density difference is 0.2%, and the average atomic force difference is 5×10^{-5} au. We also tested Si slabs and rods, and CdSe quantum dots. The errors for these tests are similar to the above Si QD results. We also calculated the quantum dot polarization under an

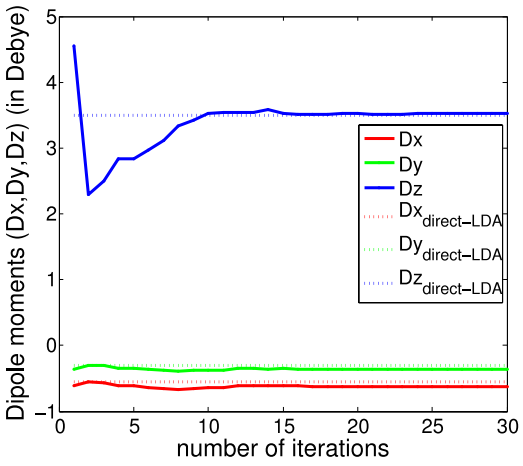

Figure 4. The convergence of the dipole moment as a function of self-consistent iteration steps for a CdSe quantum dot.

Table 1. The convergence of the LS3DF results in comparison with direct LDA results for bulk Si calculations. The fragment sizes 0.5 a, 1 a, 1.5 a correspond to 8, 64, 216 Si atoms in the $2 \times 2 \times 2$ fragments respectively. ΔE is the total energy error, $\Delta\rho$ is the total charge density error.

Fragment size	0.5 a	1 a	1.5 a
ΔE (meV/at)	30	2.9	4.0
$\Delta\rho$	1.1%	0.14%	0.08%
$\sum_F \alpha_F \int \Delta V_F \rho_F dr$ (meV/at)	213	5.5	1.0

external electric field in a Si quantum dot using the above 1.0 a fragment size. The LS3DF and direct LDA differences for the response charge and total induced dipole moment are both about 2%.

We next test the effects of fragment size on the LS3DF error. For this test, we have calculated the bulk Si with the LS3DF method. The bulk system is chosen so we can have the exact result for the direct LDA method. Our tests are shown in table 1. From table 1 we can see that the LS3DF errors drop rapidly as the fragment size increases from 0.5 a to 1.5 a. We note that the total energy error does increase a bit in going from 1.0 a to 1.5 a. This is due to the use of negative fragments. Because of this, although the LS3DF total energy is variational with regard to fragment wavefunctions $\psi_{F,i}(r)$, it is not an upper limit of the converged total energy; it does not approach the converged energy monotonically from above when the size of the fragment increases. This does pose a challenge for how to estimate the accuracy of the LS3DF calculations. A more robust test is based on the charge density error, which drops rapidly with the fragment size. In table 1, we have also shown the last term in equation (1). As we discussed before, this term should be canceled out from different fragments. This is indeed true. If all α_F are all set to unity, then this term can be tens of thousands of times larger.

Another direct way to test the accuracy of the LS3DF method is to compare the physical properties calculated by the LS3DF method and the direct LDA method. One very sensitive property is the total dipole moment of a quantum dot. As we know the total energy error depends on the second order of the wavefunction and charge density errors, while the dipole

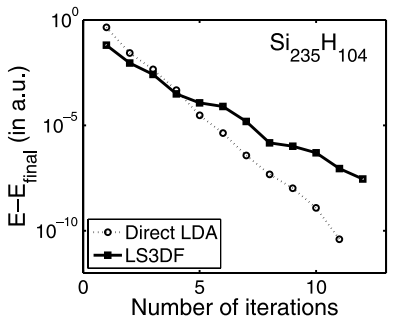

Figure 5. Self-consistent convergence curves for LS3DF and direct LDA methods.

moment error depends on the first order of the wavefunction and charge density errors. We calculated the dipole moments for a small 178-atom wurtzite CdSe quantum dot. In the charge density self-consistent calculations, for both LS3DF and direct LDA methods, we have solved the Poisson equation using an open boundary condition, thus there is no neighboring dipole–dipole interaction due to the use of a periodic supercell. Using a $1 \times 1 \times 1$ fragment of 12 Cd + Se atoms, the LS3DF z-direction (c-axis) dipole moment was computed to be 3.52 au, while the LDA result was 3.49 au. The absolute difference (0.03 au) is much smaller than the error introduced by using different pseudopotentials. Figure 4 shows the convergence of the dipole moments through the self-consistent iterations. This can be compared with the total energy convergence shown in figure 5.

As discussed in the introduction, one common problem for some of the O(N) methods is the existence of local minima and the resulting slow total energy convergence. Figure 5 shows the convergence of the self-consistent iterations in LS3DF in comparison with the direct LDA result. As one can see, the LS3DF method has a convergence rate similar to the direct LDA method. Thus, the LS3DF method does not have the convergence problem seen in other methods. This is understandable because we are using the same potential mixing scheme as the direct LDA method (figure 2), and the whole system charge to potential formula (Poisson equation and LDA exchange correlation formula) is the same for these two methods, and the potential to charge response function is essentially the same as we have tested by applying an external electric field discussed before. Besides, since the calculation of fragment wavefunction is independent for different fragments, the convergence of the Schrodinger's equation (5) is fast using the traditional conjugate gradient method. In figure 6, we show the computational cost as the flop counts of the LS3DF method and the direct LDA method. As we can see, these two methods cross around 500 atoms when a 1 a fragment size is used. This crossover is similar to the crossover reported by many other O(N) methods. In the flow chart of figure 2, the Kohn–Sham equations for different fragments (equation (5)) are solved independently. As a result, this step (the second box of the LS3DF flow chart in figure 2) can be parallelized trivially. The time spent on the global Poisson equation is less than 5% of the total computational time. As a result, we have been able to achieve an excellent linear scaling up to thousands

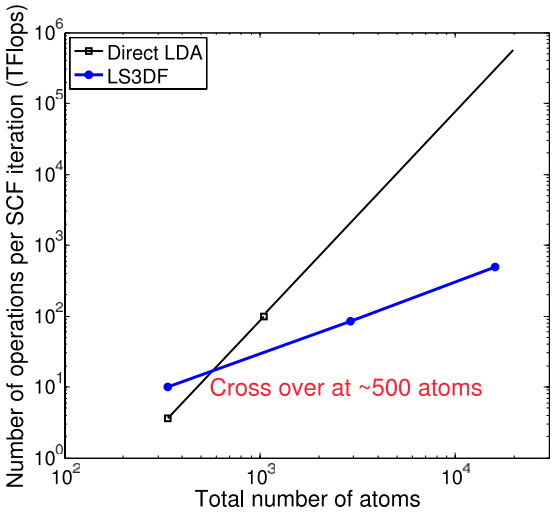

Figure 6. The computation flops needed for one step in self-consistent calculation. A typical 1 *a* fragment size for the smallest fragment is used for the LS3DF calculations. The flop counts are measured on an IBM-SP power3 machine using the profiling tool IPM.

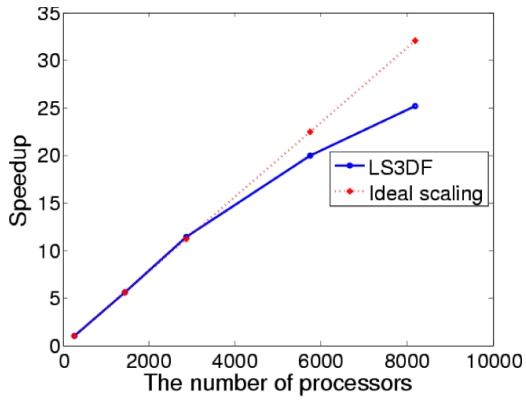

Figure 7. The speed-up of the LS3DF method versus the number of processors. The speed-up is measured starting from 256 processors. The system calculated is a CdSe quantum dot with about 2000 atoms. The smallest fragment contains 12 Cd and Se atoms. 32 processors are used to calculate each fragment. The computation is done on a Cray XT4 computer.

of processors. In figure 7, we show the speed-up of the LS3DF method with the number of processors for a 2000 atom CdSe quantum dot. The method scales well up to 6000 processors. In this test, 32 processors are used to calculate each fragment. We are working on other parts of the code (e.g. the first and third boxes in the LS3DF flow chart of figure 2) and the file I/O to further improve the performance of the code.

To demonstrate the power of the LS3DF method, we have calculated a 15 000 atom Si quantum dot. It takes about 30 min for one self-consistent iteration using 2048 processors on an IBM-SP power3 machine. The Si quantum dot charge is shown in figure 8. By comparison, if a direct LDA method had been used, it could take many weeks using a similar number of processors [17]. We have also used the LS3DF method to study the total dipole moments of CdSe quantum dots [12], and found

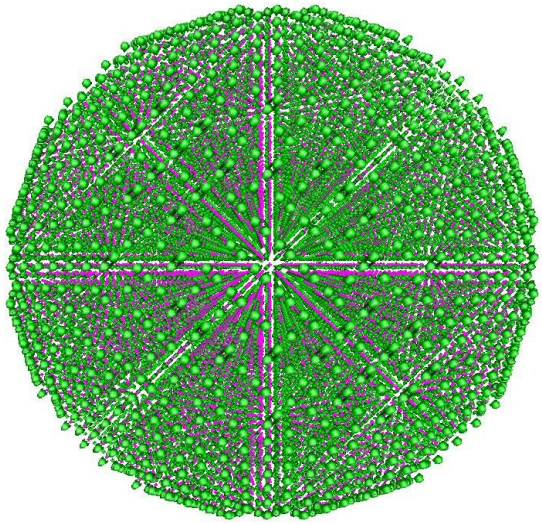

Figure 8. The charge density isosurface (green) of a 15 843 Si atom quantum dot calculated using the LS3DF method. The smallest dots (pink) represent the individual Si atoms.

significant dipole moments of those quantum dots which can change the localization of internal electron states.

Finally, we would like to comment on the prefactor of our linear scaling method, comparing to other methods reported in this special journal issue. Accurate comparison is difficult because different codes use different wavefunction representations, and there are many other factors affecting the overall speed of the code. Thus, here we can only provide a rough estimation. The reported crossovers with direct LDA calculation for localized orbital and density matrix methods are also about 500 atoms. Since these crossovers are similar to ours for similar accuracy, we can deduce that the LS3DF method should be as efficient as those O(N) methods [5] because the direct LDA method is the same. If the direct LDA method is not the same (e.g. due to the use of real space grid, or atomic basis sets), for a fair comparison we can also change the corresponding method in our LS3DF calculation for the fragment wavefunctions. This should not change the crossover size, thus the estimation of the relative efficiency. One can also estimate the computational cost (for a system with 2N electrons) as follows. First, note that for an accurate LS3DF calculation it is not necessary to have small quantum confinement effects for the fragment wavefunctions. What is important is for one side of the fragment to have a small effect on the other side in terms of charge density and kinetic energy density. Such charge density effects should have a similar decay length as the localized Wannier function, because the densities can be calculated from the sum of the Wannier function square. This is often observed in nanocrystal calculations. While an individual eigenstate wavefunction might be significantly different from its bulk counterpart even far away from the surface due to quantum confinement effects, the total charge density approaches its bulk value quickly, right after a bond length away from the surface. In practice, we do find that the 64 atom $2 \times 2 \times 2$ fragment size is similar to the orbital size needed in the localized orbital method for similar accuracies [5]. Most of the LS3DF computational cost is in

the computation of the $2 \times 2 \times 2$ fragments. There are $M = m_1 \times m_2 \times m_3$ such fragments, each with $16N/M$ electrons. Thus, in total, there will be $8N$ fragment wavefunctions for the total calculation each in a domain of $\Omega_{2\times2\times2}$. In the localized orbital method, the number of localized wavefunctions is about N–$2N$ depending on the implementation [5]. Thus, from this count, our method does have a larger number of fragment wavefunctions than the localized orbitals. But this can be compensated by the fast iterative convergence of the fragment wavefunctions (equation (5)) due to the wavefunction decoupling among the fragments. In a localized orbital method, the change of the local orbital from one site might affect the local orbitals at far away sites due to the orthogonalization among the localized orbitals. This will slow down its iterative convergency. Besides, in our wavefunction calculation for each fragment, the $O(N^3)$ step is not dominating yet due to the relative small fragment sizes. Thus, this is a computational sweet spot for the fragment calculations. Lastly, as discussed before, one can reduce the size of the '$2 \times 2 \times 2$' fragment (smaller than the 2 grid size of the $m_1 \times m_2 \times m_3$ grid) to see whether it can speed up the calculation. Overall, it is not clear at this stage which $O(N)$ method will be the fastest. It might very well be that the answer depends on the physical problem to be solved. A further study on this topic is needed, especially after all the methods become more mature.

4. Conclusion

We have presented a new divide-and-conquer linear scaling three-dimensional fragment method for *ab initio* electronic structure calculations. We have presented the technical details for how the method is implemented, and described its performance. We demonstrated that this method can be used to calculate the electronic structures of large nanosystems, and that this method can be used to achieve excellent computational efficiencies for massively parallel machines. In terms of programming, the new method can be adapted from an existing *ab initio* package relatively easily; it can also be used for other quantum mechanical methods, and not just the density functional theory. The results of this new method are very accurate when compared with the direct LDA method. In addition, it has a variational formalism; therefore, the calculation of atomic forces is straightforward using Hellman–Feynman theory.

Acknowledgments

This work was supported by the DMSE/BES/SC and MICS/ASCR/SC of the US Department of Energy under contract No DE-AC02-05CH11231. It used the resources of the National Energy Research Scientific Computing Center (NERSC) and the National Center for Computational Sciences (NCCS).

References

[1] Payne M C, Teter M P, Allan D C, Arias T A and Joannopoulos J D 1992 *Rev. Mod. Phys.* **64** 1045
[2] Gygi F, Yates R K, Lorenz J, Draeger E W, Franchetti F, Ueberhuber C W, de Supinski B R, Kral S, Gunnels J A and Sexton J C 2005 *Proc. 2005 ACM/IEEE Conf. on Supercomputing*
[3] Goedecker G 1999 *Rev. Mod. Phys.* **71** 1085
[4] Galli G and Parrinello M 1992 *Phys. Rev. Lett.* **69** 3547
[5] Fattebert J-L and Gygi F 2006 *Phys. Rev.* B **73** 115124
[6] Skylaris C K, Mostofi A A, Haynes P D, Dieguez O and Payne M C 2002 *Phys. Rev.* B **66** 35119
[7] Skylaris C K, Dieguez O, Haynes P D and Payne M C 2002 *Phys. Rev.* B **66** 73103
[8] Li X P, Nunes R W and Vanderbilt D 1993 *Phys. Rev.* B **47** 10891
[9] Bowler D R, Choudhury R, Gillan M J and Miyazaki T 2006 *Phys. Status Solidi* b **243** 989
[10] Yang W 1991 *Phys. Rev. Lett.* **66** 1438
[11] Shimojo F, Kalia R K, Nakano A and Vashishta P 2005 *Comput. Phys. Commun.* **167** 151
[12] Wang L W, Zhao Z and Meza J 2008 *Phys. Rev.* B at press
[13] Kohn W 1996 *Phys. Rev. Lett.* **76** 3168
[14] Kitaura K, Ikeo E, Asada T, Nakano T and Uebayasi M 1999 *Chem. Phys. Lett.* **313** 701
[15] Li J and Wang L W 2005 *Phys. Rev.* B **72** 125325
[16] http://hpcrd.lbl.gov/linwang/PEtot/PEtot.html
[17] Chelikowsky J 2005 *APS Bull.* **51** 612