First Principles Calculations using Density Matrix Divide-and-Conquer within the SIESTA Methodology

B O Cankurtaran^{1,2}**, J D Gale**^{1} **and M J Ford**^{2}

^{1} Nanochemistry Research Institute, Department of Applied Chemistry, Curtin University of Technology, GPO Box U1987, Perth, WA 6845, Australia

^{2} Institute for Nanoscale Technology, University of Technology, Sydney, 15 Broadway, Sydney 2007, Australia

E-mail: `mike.ford@uts.edu.au`

Abstract. The density matrix divide-and-conquer technique for the solution of Kohn-Sham density functional theory has been implemented within the framework of the SIESTA methodology. Implementation details are provided where the focus is on the scaling of the computation time and memory use, in both serial and parallel versions. We demonstrate the linear scaling capabilities of the technique by providing ground state calculations of moderately large insulating, semi-conducting and (near)metallic systems. This linear scaling technique has made it feasible to calculate the ground state properties of quantum systems consisting of tens of thousands of atoms with relatively modest computing resources. A comparison with the exisiting order-N functional minimization (Kim-Mauri-Galli) method is made between the insulating and semi-conducting systems.

Submitted to: *J. Phys.: Condens. Matter*

1. Introduction

Electronic structure calculations, based on first principles quantum mechanics, provide reliable physical and chemical descriptions of atomistic, molecular and crystal systems. However, practical calculations are often limited to fairly small systems ($<$ 500 atoms) due to both theoretical difficulties and limitations in available computational resources. The theoretical difficulties arise from the high order, $O(N^2)$ and greater, scaling which is inherit within all *ab initio* quantum mechanical methods in the absence of approximations, where N is a measure of the system size and usually most critically depends on the number of basis functions.

To date, density functional theory (DFT) [1] has proven to be a reliable and efficient choice in the study of small to medium quantum systems. Although the approximation of the exchange-correlation functional in Kohn-Sham theory leads to deviations from experiment, the reproduction of many physical properties is sufficient for practical use and often deviations are systematic, thereby increasing the level of confidence in interpreting the results. A further feature of DFT is that it is ameanable to expression through a wide variety of basis functions such as planewaves [2], Gaussians [3], wavelets [4], grids [5], B-splines [6], psincs [7], and numerical orbitals [8]. In this present article we will focus on the use of real-space localized orbital methods, while recognising this is just one of many possible approaches.

Solution of the Kohn-Sham equations consists of two key steps - the construction of the Hamiltonian and the attainment of the self-consistent field, including the determination of the orthogonal Kohn-Sham states. In the worst case scenario, construction of the DFT Hamiltonian matrix can scale as $O(N^4)$ due to the Coulomb term, though it was recognised that the use of density fitting [9] in an auxiliary basis can reduce this to $O(N^3)$ at most. Diagonalization of the Hamiltonian matrix will similarly scale as $O(N^3)$. Thus the building and diagonalization of the Hamiltonian matrix are considered to be the major bottlenecks of any conventional implementation. Although DFT is considered relatively efficient it is still computationally prohibitive for the study of systems consisting of atom numbers in the thousands and greater. To overcome this barrier, techniques have been developed and employed to reduce the scaling of the computational cost to the linear regime, $O(N)$ (order-N). In the same way, memory usage must also scale linearly, instead of as $O(N^2)$, in order to avoid another potential bottleneck.

The key to achieving linear-scaling is to enforce locality in all phases of the calculation. If the basis functions are strictly local in real space then the construction of the Hamiltonian readily becomes order-N. Only the Coulomb energy requires special consideration, but can be constructed with linear-scaling through approaches such as fast multipole methods [10], or full multigrid methods [11]. Due to the locality, the Hamiltonian matrix, and in general the overlap matrix, become sparse and thus the memory naturally scales linearly too. In the present work, we will focus on the SIESTA methodology [12] to define the Hamiltonian and overlap matrices, while noting that there are many similarities to the approach embodied within the PLATO code [13]. Here norm-conserving pseudopotentials are used to replace the core electrons and nuclei with a non-local potential, while the valence states are expanded in numerical pseudo atomic orbitals (PAOs) [14]. These PAOs are the numerical solutions to the atomic pseudized problem, represented as a tabulation on a radial grid and multiplied by the appropriate spherical harmonic. In order to make the basis functions strictly local, the atomic problem is solved within a confining potential that becomes infinite, either instantaneously, or asymptotically, at a given radius [15]. Thus the approximation is contained within the basis function, rather than the Hamiltonian, as opposed to methods where the Hamiltonian is made sparse through thresholding of integrals involving infinitely ranged basis sets [16]. Further details of the construction of the Hamiltonian, as well as the extension of the SIESTA approach to include greater radial variational freedom, can be found elsewhere.

Of course, enforcing locality in the Hamiltonian is a necessary, but not sufficient condition for a method to be order-N. It is also necessary to replace matrix diagonalization with an approach to obtaining the self-consistent density that enforces localized solutions without explicit orthogonalization of all Kohn-Sham states. This exploits the fact that states are known to decay exponentially in materials with a band gap, while even metals exhibit power-law decay. One of the first linear scaling methods to be proposed in this context for DFT was the divide-and-conquer (D&C) approach, proposed by Yang in 1991 [17, 18] and then subsequently reformulated for use within the density-matrix framework in 1995 [19]. This method reduces the $O(N^3)$ scaling inherit with the diagonalization of the Hamiltonian matrix to the linear scaling regime by using partition functions to subdivide the electron density of the complete system. Each subsystem is then solved separately and the electron charge density of each subsystem is found. The sum of the corresponding contributions from all subsystems is used to obtain the total electron density and the energy of the system. This is possible due to the fact that the electron density is a local property within DFT.

Following the proposal of the divide and conquer approach, there was extensive interest in other linear-scaling approaches within the field. This included methods based on functional minimization with respect to localized Kohn-Sham states [20], while avoiding explicit orthogonalization, and techniques that operate directly on the density matrix with sparsity imposed [21],[22],[23]. In the present implementation of the SIESTA methodology the Kim-Mauri-Galli (KMG) functional [24] is generally employed to determine the electronic states under the imposition of a fixed electronic chemical potential (i.e. Fermi level). At this point it is appropriate to consider the merits and demerits of the different approaches. Firstly, the divide and conquer approach suffers from the problem of duplication of effort. As will be seen when the details of the method are presented in the subsequent section, any given matrix element will appear in the Hamiltonian of many localized states and similar eigenstates will be generated in numerous cases since they will contribute to different subsystems. Hence, the overlap of subsystems leads to repetition that increases the prefactor of the linear-scaling and consequently the cross-over point at which the linear-scaling algorithm out performs matrix diagonalization can be raised. Turning to consider the KMG approach, the use of functional minimization eliminates the duplication of effort present in divide and conquer. However, the KMG method is subject to difficulties of its own. Because the algorithm works at constant chemical potential, rather than fixed number of electrons, it is necessary to *a priori* specify the Fermi level to lie within the band gap. If this is not the case, then the method diverges. For wide gap insulators this is rarely an issue since there is considerable margin for error when guessing the chemical potential to use, whereas for a semiconductor or small gap system it becomes a matter of trial and error. To complicate things further, the Fermi level is a function of the density matrix and therefore will change during the self-consistent field iterations, leading to the potential need to adjust the chemical potential at each cycle during the early stages of SCF convergence. Consequently, the most practical scheme for utilizing the KMG method is to perform a small number of iterations of SCF using conventional diagonalization in order to obtain a good approximation to the density matrix and to locate the band gap, and then use this information to initialize the order-N method. This approach is particularly advantageous when performing first principles molecular dynamics or geometry optimization of complex structures, where the cost of the initial few cycles of diagonalization becomes insignificant relative to the number of subsequent SCF iterations.

Although being one of the earliest so-called order-N methods, D&C has been relatively neglected until recently [25] within the condensed matter physics field, though it has found significant use within the chemistry community due to the greater focus on localized basis sets and semi-empirical QM methods [26, 27, 28]. A few researchers have extended the D&C method for the applicability of it to large molecular dynamics simulations using the frozen density approach [29, 30] and to solid state systems [31, 32]. It could be argued that the situation with regard to the prefactor of divide and conquer is not as severe as it might be on current computers for two reasons. Firstly, there exist highly machine optimized routines for serial diagonalization on most platforms that have made diagonalization as competitive as it is for moderately sized problems. Secondly, the simplicity of the scheme lends itself to two tier parallelism, with distributed memory schemes for the division of the subsystems over processes, while each diagonalization may be parallelized over a smaller number of nodes using a shared memory paradigm. This approach will be particularly well suited to modern multi-core machines. When these factors are combined with the robust nature of divide and conquer with respect to the size and position of the band gap there is reason to believe that reappraisal of the D&C scheme is in order.

Here, we report our implementation of the D&C technique within the SIESTA code [12]. When coupled with the linear combination of numeric atomic orbitals within the SIESTA methodology, our results suggest that D&C can prove to be a very efficient first principles quantum mechanical calculation method. By incorporating D&C within SIESTA, we have taken advantage of the linear scaling associated with numerical orbitals, in the sparse matrix representation, when constructing the Hamiltonian matrix. Hence, we have provided a robust fully linear scaling solution to DFT calculations.

2. Density-Matrix Divide-and-Conquer Theory

The D&C scheme is related to the principle that the electronic structure for a particular region of a quantum system, to a good approximation, only depends significantly on the external potential due to nearby sub-systems, while those further away are rapidly screened with increasing distance. This principle was formalised and coined "near-sightedness" by Kohn [34]. The divide and conquer method, first proposed by Yang [17, 18], was arguably the first practical linear-scaling scheme for first principles methods and while it precedes the work of Kohn, it builds on the prior knowledge of localization through construction of Wannier functions [35, 36].

The D&C method involves dividing a system into a set of smaller overlapping subsystems. The speedup in calculation time occurs because each subsystem is solved separately with a cost that no longer depends on the size of the global problem. The individual subsystems are coupled to each other by a common Fermi energy allowing electrons to flow until equilibrium is achieved. The obtained electronic information for each subsystem is then combined in a specific way so as to provide an approximation to the global (complete system) density matrix.

Our implementation treats each subsystem as consisting of a core region that is surrounded by a buffer region, as per the original work of Yang [17]. The atom(s) found in the core region are those whose localized electronic states are to be determined, while the atoms within the buffer region are required to correctly describe the electronic states of the core atoms within the local subsystem. For the purposes of the present work, we shall focus on the situation where the core region holds one atom, while the buffer region can include as many atoms as required. Each atom in the system will become a core atom of a single subsystem. The size of the buffer region depends on the decay length within the material of interest and controls the degree of deviation from the unrestricted Kohn-Sham solutions. Within the SIESTA methodology, an initial guideline as to the radius needed is given by the distance at which the Hamiltonian matrix elements go exactly to zero (which will always be greater than the equivalent distance for the overlap matrix as a consequence of the matrix elements arising from the pseudopotential). However, the buffer size may need to exceed this distance since there is no guarantee that the density matrix will decay at the same rate as the Hamiltonian. Despite this, it is found that using smaller buffer radii than the Hamiltonian cutoff can also produce reasonable qualitative results for certain systems, as will be shown in section 4.1.1.

Although, the present focus is on the situation where there is a subsystem centred on each individual atom this need not be the case. For example, where atoms are closely linked, such as in a functional group or small covalent molecule, this entity could be treated with a single subsystem. The benefit of this is that the computational cost is lowered by a factor related to the number of core atoms per subsystem. In the limit where serial diagonalization dominates, the cost will be reduced by the third power of the number of atoms combined per core (assuming all have the same number of basis functions per atom). The disadvantage is that in a system with an evolving geometric structure then there is a greater risk of discontinuities in the potential energy surface should a functional group dissociate and the subsystems are dynamically updated, while if the membership of the subsystems remains fixed then the quality of the electronic structure would be a non-uniform function of the nuclear configuration. Although, not reported here, we have attempted to remedy this problem by smoothing out the boundaries of the subsystem Hamiltonian matrix but have only achieved a small correction in the final total energies. Further work is required to alleviate this problem. Having a subsystem centred on each atom represents the conservative option that minimizes such errors, at an increased computational cost.

2.1. Formulation

The formulation described as follows is based on the density-matrix version of the D&C method [19]. Here, the density-matrix is the primary entity in the formulation, the focus of D&C is to estimate the global density-matrix from the sum of contributions from all subsystem density matrices.

Within D&C the global density matrix is divided up into individual subsystem density-matrices weighted by a normalized partition function;

$$\sum_\alpha \mathbf{P}^\alpha_{ij} = 1, \tag{1}$$

where α is the subsystem index, i and j are orbital indices. The partition function, $\mathbf{P}^\alpha_{ij}$ is defined by a Mulliken type [37] weight matrix (suitable for subsystems consisting of one core atom),

$$\mathbf{P}^\alpha_{ij} = \begin{cases} 1 & \text{if } i \in \alpha \text{ and } j \in \alpha \\ 1/2 & \text{if } i \in \alpha \text{ and } j \ni \alpha \ . \\ 0 & \text{if } i \ni \alpha \text{ and } j \ni \alpha \end{cases} \tag{2}$$

Defining the Kohn-Sham one electron density;

$$\rho(\mathbf{r}, \mathbf{r}') = 2 \sum_m^{N/2} \psi_m(\mathbf{r}) \psi_m(\mathbf{r}') = \sum_{ij} \rho_{ij} \phi_i(\mathbf{r}) \phi_j(\mathbf{r}'), \tag{3}$$

where electron density is defined in the space of the Kohn-Sham orbitals, $\{\psi_m(\mathbf{r})\}$. The density matrix, ρ_{ij}, is defined in the atomic orbital space, $\{\phi_i(\mathbf{r})\}$, and is given by the linear coefficients, $\{C_{im}\}$, as follows:

$$\rho_{ij} = 2 \sum_m^{N/2} C_{im} C_{jm}. \tag{4}$$

We can then divide the density matrix into subsystem contributions. The density matrix is then a sum of contributions from all subsystems, weighted by the partition matrix:

$$\rho_{ij} \equiv \sum_\alpha \mathbf{P}^\alpha_{ij} \rho_{ij} = \sum_\alpha \rho^\alpha_{ij}. \tag{5}$$

The local nature of the density matrix allows each subsystem density matrix contribution to be approximated by;

$$\rho_{ij}^{\alpha} = 2\mathbf{P}_{ij}^{\alpha} \sum_{m} f_{\beta}(\epsilon_F - \epsilon_m^{\alpha}) C_{im}^{\alpha} C_{jm}^{\alpha} \tag{6}$$

where f_β is the Fermi function approximating an occupation number, β is the inverse electronic temperature, ϵ_F is the Fermi energy common to all subsystems and ϵ_m is the orbital energy.

The Fermi energy needs to be found iteratively so that the global density matrix yields the correct number of electrons, N;

$$N = \sum_{ij} \rho_{ij} \mathbf{S}_{ij} = \sum_{ij} \left(2 \sum_{\alpha} \mathbf{P}_{ij}^{\alpha} \sum_{m} f_{\beta}(\epsilon_F - \epsilon_m^{\alpha}) \times C_{im}^{\alpha} C_{jm}^{\alpha} \right) \mathbf{S}_{ij} \tag{7}$$

3. Implementation

In the present work we have combined the density-matrix D&C scheme with the SIESTA methodology [12] for the linear-scaling construction of the Hamiltonian and overlap matrices. Given the use of localized PAOs as basis functions within the SIESTA methodology, this is a natural combination to achieve linear-scaling for large systems with relatively modest resources. The following sections contain a description of the key aspects of the present methodology.

3.1. Algorithm

The general overview of the D&C implementation within the SIESTA code is shown in a flowchart in Figure 1. The flowchart has been appropriately marked to indicate which parts of the code involve the original SIESTA routines (straight line), parallel communication (dashed line) and the present D&C module (dotted line). The algorithm begins by reading the spatial locations of all atoms and options to perform the DFT run. Once the atom specifics have been read into SIESTA it will distribute the atom information across the compute nodes according to a domain decomposition algorithm (see section 3.3). In short, each compute node will be responsible for a subset of orbitals localized in a region of space and all the corresponding electronic information pertaining to those orbitals. Each node then generates the elements of the Hamiltonian and overlap matrices that it is uniquely responsible for; if in serial mode the complete matrices are stored on the single node. The D&C section of the code then begins from this point. If it is the first SCF cycle, the system will be divided into subsystems. This entails creating a list structure to store the orbital information for each subsystem with distinguishing lists for the core and buffer atoms. If running in parallel, the matrix elements belonging to buffer orbitals that reside on other compute nodes need to be communicated to the nodes with ownership of subsystems requiring that data. Because of the spatial locality of the domain decomposition, the number of compute nodes to be communicated with should remain constant or decrease as the system size increases, according to whether the number of processors employed scales with the system size or remains fixed, respectively.

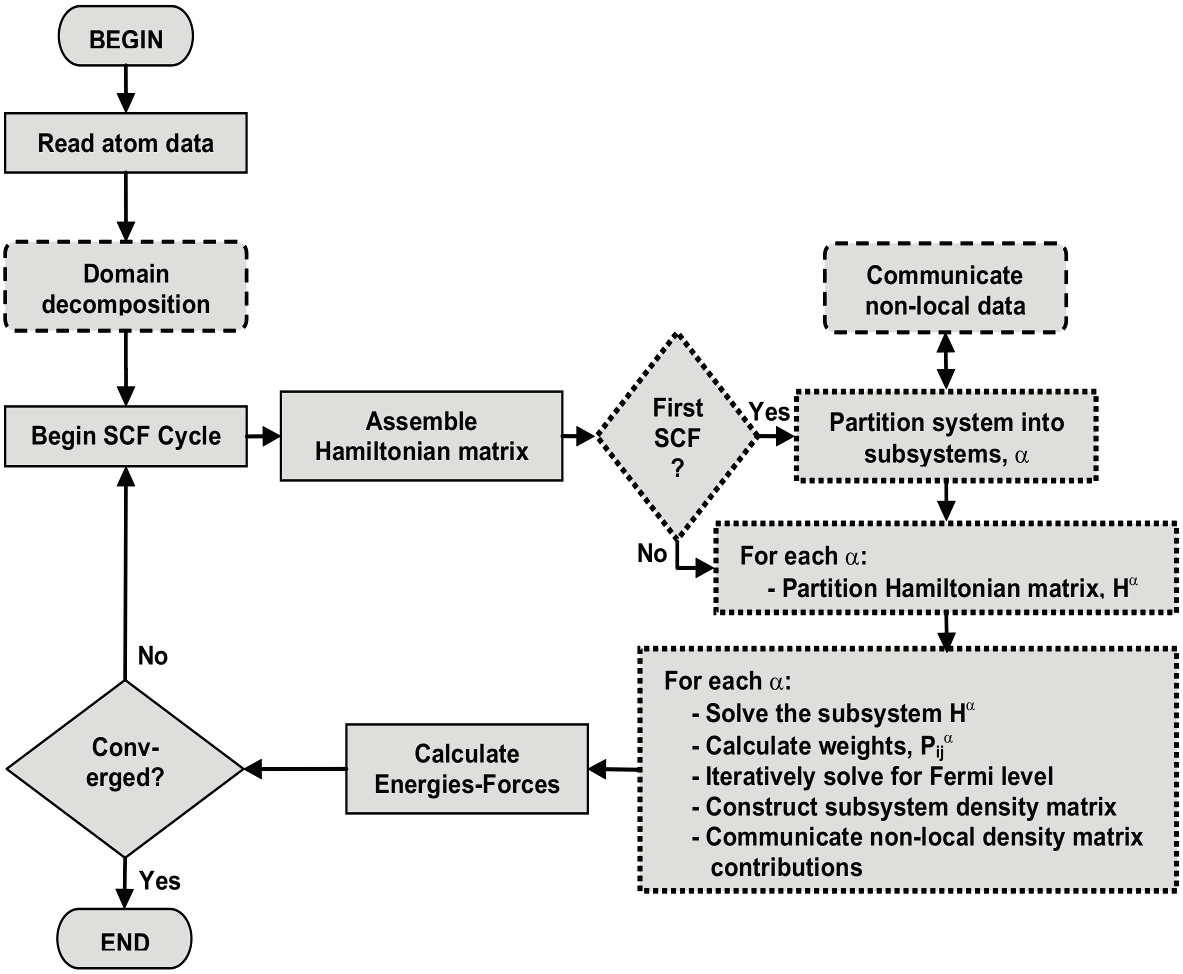

BEGIN
Read atom data
Domain decomposition
Begin SCF Cycle
Assemble Hamiltonian matrix
First SCF ?
Yes
No
Communicate non-local data
Partition system into subsystems, α
For each α:
- Partition Hamiltonian matrix, H^{α}
For each α:
- Solve the subsystem H^{α}
- Calculate weights, P_{ij}^{α}
- Iteratively solve for Fermi level
- Construct subsystem density matrix
- Communicate non-local density matrix contributions
Calculate Energies-Forces
Conv-erged?
No
Yes
END

Figure 1. Schematic outlining the major implementation sections and process flow for the implementation of D&C within the SIESTA code.

The solution for the global density matrix proceeds by first solving the generalized eigenvalue problem for each subsystem, calculating the partition weights (equation 2) and other values that will benefit from caching. Once the eigenvalues of all subsystems are known, the Fermi energy is found by iterative variation until equation 7 is satisfied. Having determined the Fermi energy, the global density matrix is found by calculating the density matrices for each subsystem and then combining the contributions multiplied by the previously calculated partition weights.

3.2. Memory Considerations

When using D&C for large systems, the amount of memory used by the process must be manageable and scale linearly with system size. For D&C to be practical for very large systems only the information that is absolutely required should be stored. A large part of task is already accomplished within SIESTA since all matrices that represent orbital based information (such as the Hamiltonian, overlap and density matrices) are stored in a sparse matrix representation [] as a 1-D array of non-zero valued elements. Because of the strict spatial locality of basis functions, the sparsity patterns for the Hamiltonian and overlap matrix are known *a priori* and fixed for any given nuclear configuration, while the density matrix is assumed to adopt the same sparsity pattern as the Hamiltonian. This use of sparse arrays ensures that the SIESTA methodology, by default, is linear scaling in memory usage, except when diagonalization is employed. Here dense matrix algebra is used locally for compatibility with standard eigensolution routines. Diagonalization is typically used in cases where the system size is below the cross-over point at which linear-scaling solution becomes advantageous, as well as in cases where the details of the band structure for a material are to be determined.

The D&C implementation, as has been described in section 3.1, can consume large amounts of memory for large systems. This is due to the fact that each subsystem must store 2-dimensional arrays for the subsystem Hamiltonian matrix, the subsystem overlap matrix, the subsystem eigenvector solutions and the subsystem density-matrix. However, the subsystem Hamiltonian and overlap matrices are not in use by the time it comes to construct the density matrix, reducing the peak memory use. In the algorithm where the computational effort is minimized, the eigenstates of all subsystems must be stored simultaneously since they cannot be used in the construction of the local density matrix until the global Fermi energy is known. When the number of subsystems is large and the subsystem sizes are considerable this can lead to a prohibitive amount of memory usage.

To overcome this issue, an alternate algorithm has been implemented that counters this problem, if so desired. It is accomplished by using a single allocation of memory for each matrix (Hamiltonian, overlap, eigenvectors and density matrix) that is large enough to store the information for the largest subsystem. That is, instead of storing matrices for each subsystem only one set of matrices are stored and reused for each subsystem. This reduces the memory usage from $N_p N_{orb}^H N_{orb}^S N_{orb}^{eig}$ to $N_{maxorb}^H N_{maxorb}^S N_{maxorb}^{eig}$, where the subscript *maxorb* denotes the use of the maximum number of orbitals found within any of the subsystems. Using this memory conserving option leads to the memory usage scaling in a sub-linear fashion, but does increase the computing time required for each SCF iteration, since the subsystem Hamiltonian and overlap matrices will need to be diagonalized twice (the first time just requiring determination of the eigenvalues) if no caching of eigenvectors for later use can be performed. Depending on whether the calculation time is dominated by the diagonalization step, this can have a significant influence on the time required for the SCF cycle. On average there is 50% increase in computing time and the worst case scenario will yield a doubling of the prefactor.

If memory usage is the key bottleneck, then it can be reduced to the absolute minimum required by computing all eigenvalues for the subsystems on the fly as required. Given that the eigenvalues are needed at each iteration of the Fermi energy solution, this likely to make this algorithm uncompetitive as it would increase the prefactor by at least an order of magnitude, if not more. Memory reduction can also be achieved by grouping atoms together to form subsystems (i.e. multiple core atoms per partition), since this reduces the total number of eigenstates to be stored by eliminating some duplication.

3.3. Parallelization

The parallel version makes use of the load balancing scheme included within the SIESTA package for the KMG order-N method, namely a domain decomposition algorithm to distribute the atoms amongst the compute nodes. The domain decomposition algorithm divides the unit cell into right-angled sections of side lengths as close to being equal while remaining commensurate with the lattice vectors. It then allocates each non-empty section (i.e. each section with a non-zero atom count) to a node. The allocation is conducted in a way so as to try to achieve a balanced number of atoms per node. This process could be further refined by accounting for the neighbour density in order to achieve improved load balancing. The contributions to the Hamiltonian, overlap and density matrices from each atom are then stored on the corresponding compute nodes. When using conventional diagonalization routines within SIESTA a block-cyclic orbital decomposition (either 1-D or 2-D) scheme is used to enable compatability with the ScaLAPACK [39] parallel eigensolvers.

Because of the use of spatial locality during the parallel construction and solution for each subsystem, the only global communication occurs during the determination of the Fermi energy. Here the eigenvalues and weights are stored on the node responsible for that particular subsystem. For every trial value of the chemical potential, the occupancy of each subsystem must be determined and a global summation performed to determine the total number of electrons before iteratively refining the Fermi level. Once the Fermi energy is converged then the overall density matrix can be constructed by purely local communication.

4. Results

Calculations have been performed on a range of different systems in order to examine the performance of the present combination of D&C with the SIESTA methodology. The examples chosen include insulating, semi-conducting and near-metallic systems in order to demonstrate the varied application of D&C. The specific test cases are a linear alkane chain, C_nH_{2n+2}, for the insulating system, previously studied by Warschkow et al [40], bulk silicon for the semi-conducting system, and a single walled (5,5) armchair carbon nanotube for the near-metallic system. The linear scaling and the rate of convergence of the total energy to the Kohn-Sham energy when increasing the partition radius are studied. By increasing the partition radius, this implies increasing the number of buffer atoms in the buffer region. This is reported as an increase in the buffer region radius surrounding the core atom (subsystem centre). As with all tests in this study, each subsystem contains a single core atom surrounded by a buffer region. With this type of partitioning the number of subsystems equals the number of atoms within the system.

The scaling of the calculation time is shown by plots of the time required to complete the first SCF cycle and the section of the first SCF cycle only relevant to the D&C module. The first SCF cycle incorporates the building of the Hamiltonian and overlap matrices (handled by the SIESTA code) and the diagonlization and building of the global density matrix (handled by the D&C module). For comparison, the performance of the Kim-Mauri-Galli order-N solver already implemented within SIESTA is examined for the polymer and bulk silicon. Due to the inherent difficulties of achieving convergence, when working at fixed chemical potential, for the near-metallic nanotube the KMG algorithm was not examined for this case.

The calculations for the semi-conducting bulk silicon system were performed using the memory conservation scheme, as described in section 3.2. The remaining calculations were performed using the algorithm in which the eigenvectors for each subsystem are stored during the computation of the Fermi level.

Calculations were performed on a 32 processor SGI Altix machine (1.5 GHz) with 64 GB of RAM. All calculations were run on a single processor, except those in Section 4.4 where the parallel performance of the code for a bulk silicon system consisting of 21952 atoms is examined.

4.1. Insulating System

4.1.1. Linear Alkane Chain The example of an insulating system studied here is the 1-D periodic linear alkane chain, C_nH_{2n+2}, where the number of formula units per unit cell, n, has been varied. This system should provide a favourable case for all linear-scaling methods as a closed-shell, wide gap, material with low dimensionality. The calculations were carried out using a 150 Rydberg cut-off for the real-space integration grid used to represent the density, an energy shift of 0.02 Rydberg for the PAO orbital confinement, and a density matrix convergence criteria of 1x10^{-4} for self-consistency. The Perdew-Burke-Ernzerhof (PBE) [38] form of the generalized gradient approximation (GGA) was used for the exchange-correlation (XC) functional. The dependence of the D&C method on the basis set and the buffer region size is examined for various length alkane chains in Table 1. The table shows the energy difference per atom between the D&C calculated total energy and the conventional SIESTA calculated total energy, $(E_{dc} - E_{siesta})/n$, computed by diagonalization.

The errors found for all basis sets and buffer region sizes are relatively small. Given that the numbers quoted are the absolute differences in energy, any relative energies would exhibit even smaller discrepancies. Furthermore, even for the smallest buffer region size any error is likely to be small at the level of the accuracy of DFT. As the quality of the basis set is improved from SZ to DZ, the discrepancy in the energy increases, while inclusion of polarization functions actually leads to a reduction in error, at least for smaller buffer regions. While such variations will be sensitive to the details of the construction of the basis functions, such as the split-norm for radial degrees of freedom, the important conclusion is that there unlikely to be a strong influence on the

Table 1. Energy differences per formula unit (eV) between diagonalization and divide and conquer as a function of buffer region size and basis set for the C_nH_{2n+2} alkane chain.

Number of Atoms	Buffer Region (Å)	Basis Set			
		SZ[a]	SZP[b]	DZ[c]	DZP[d]
192	5.0	4.285E-03	2.705E-03	-1.661E-02	4.170E-03
	7.5	6.0765E-04	3.164E-04	-9.237E-04	8.031E-05
	10.0	-7.074E-07	6.057E-06	-4.656E-05	4.705E-05
384	5.0	4.288E-03	2.705E-03	-1.661E-02	4.167E-03
	7.5	6.076E-04	3.164E-04	-9.237E-04	8.030E-05
	10.0	-7.075E-07	6.063E-06	-4.656E-05	4.705E-05
768	5.0	4.286E-03	2.705E-03	-1.661E-02	5.258E-03
	7.5	6.074E-04	3.164E-04	-9.151E-04	1.026E-04
	10.0	-7.075E-07	6.061E-06	-4.656E-05	4.705E-05

[a] Single-zeta. [b] Single-zeta + polarization. [c] Double-zeta. [d] Double-zeta + polarization.

Table 2. Force differences per formula unit per Angstrom (eV/Å) between diagonalization and divide and conquer as a function of buffer region size and basis set for the C_nH_{2n+2} alkane chain.

Number of Atoms	Buffer Region (Å)	Basis Set			
		SZ[a]	SZP[b]	DZ[c]	DZP[d]
192	5.0	4.62E-02	-7.97E-03	-8.15E-02	-1.03E-01
	7.5	-1.24E-03	-1.77E-03	2.20E-03	-4.74E-03
	10.0	3.50E-05	5.00E-05	6.10E-05	-9.91E-04
384	5.0	4.67E-02	-8.00E-03	-8.15E-02	-1.03E-01
	7.5	-1.24E-03	-1.77E-03	2.20E-03	-4.74E-03
	10.0	3.50E-05	5.00E-05	6.10E-05	-9.91E-04
768	5.0	4.65E-02	-7.92E-03	-8.15E-02	-1.02E-01
	7.5	-2.62E-03	-1.59E-03	2.02E-03	-4.73E-03
	10.0	3.50E-05	5.00E-05	6.10E-05	-9.91E-04

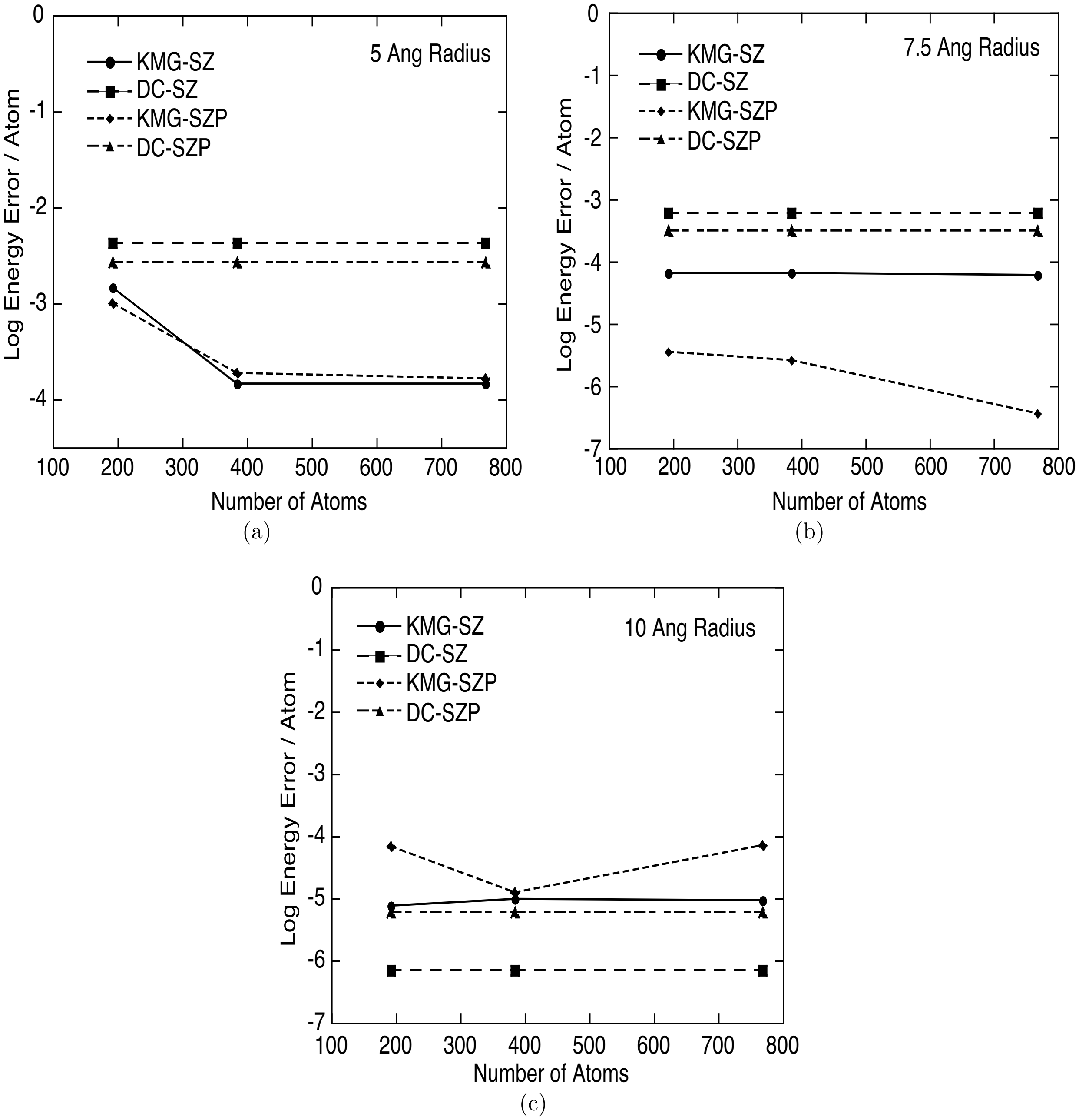
[a] Single-zeta. [b] Single-zeta + polarization. [c] Double-zeta. [d] Double-zeta + polarization.

convergence behaviour of the D&C method.

As is to be expected, the errors decrease in size as the buffer region radius is increased. Table 1 shows that even a small buffer region radius of 5.0 Å is adequate for this system, regardless of basis set size, even though the buffer region is smaller than the maximum orbital interaction range of 7.3030 Å (for single-zeta, SZ) to 7.4416 Å (for DZP). The errors in the calculated forces are shown in table 2. The errors in the forces are larger than the total energy errors, however, this is to be expected. As with the total energy errors, the errors in the force decrease as the buffer region is increased. The size of the errors for the 10.0 Å buffer region indicate that molecular dynamic simulations are a possibility with the D&C scheme, as long as the buffer region is an adequate size.

For comparison to the present D&C results, we have also performed calculations on this model system using the Kim-Mauri-Galli order-N functional. The same localization radius has been applied to the Wannier functions within the KMG approach as for the partition radius in the D&C technique. Consequently, both methods are attempting to find localized solutions with the same confinement constraint. The methods differ though in that the KMG approach contains a further approximation in that inverse of the overlap matrix is represented by a series expansion, usually truncated at first order. The errors in the total energy relative to full diagonalization are shown as their logarithms in Figure 2 for both KMG and D&C. For D&C the order of magnitude of the error is relative constant as a function of increasing system size, while that for KMG decreases. This behaviour is likely to be, at least in part, a consequence of the increased sparsity of the overlap matrix leading to the additional approximation within the KMG scheme improving. Interesting, for the smaller radii of confinement for the eigenstates the KMG yields a lower error in the total energy than the D&C scheme, which is somewhat unexpected, though the situation reverses for a radius of 10.0 Å.

The scaling of the calculation time of SZ basis set calculations for increasing supercell dimensions of the C_nH_{2n+2} alkane chain is shown in Figure 3(a). The graph shows the timing contribution of the D&C module section to the first SCF cycle. The graph clearly exhibits linear scaling of the calculation time as the system size is increased for all buffer region sizes (i.e. the diagonalization of the Hamiltonian matrix and the assembly of the global density matrix are all linear scaling processes). Athough, not shown here, the scaling is found to be linear regardless of basis set sizes, as expected. It is also possible to analyze the prefactor associated with the buffer region radius for this simple case. For radii of 5.0 Å, 7.5 Å, and 10.0 Å, the number of orbitals within the partition centred on a carbon atom is 42, 66 and 90, respectively, for a single-zeta basis set. When the slopes of the lines in Figure 3 are compared against these numbers, it appears that the prefactor scales approximately as the second power of the number of orbitals in the partition, as opposed to the theoretical maximum of a cubic scaling. Figure 3(a) shows a comparison of calculation time with the KMG order-N method. A direct comparison is not appropriate in this case as the KMG method generally has differing times for each SCF iteration. The first few SCF iterations take the longest time and as the caluclation progesses through the SCF steps each iteration takes less time. The figure displays the timings for the contribution to the KMG order-N method for the first SCF iteration and the average time for all SCF iterations compared with the calculation time for the D&C section of the first SCF cycle. The calculation times for the C_nH_{2n+2} alkane chain does not differ too much between the order-N methods.



(a)

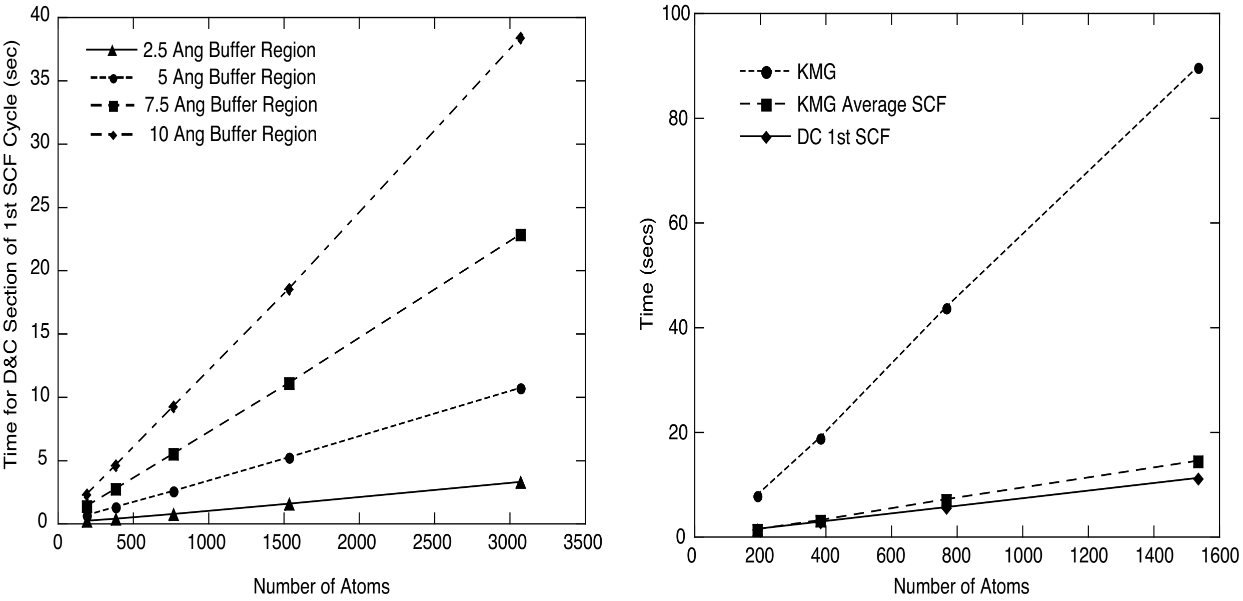
(b)

(c)

Figure 2. Comparisons of the errors per atom in the total energy between the D&C method and the KMG method for the C_nH_{2n+2} alkane chain with varying lengths. The D&C method exhibits a constant error as a function of the system size, while for the KMG method, the error becomes constant as the system size is increased. a) Using a 5.0 Å radius for the buffer region (D&C) and the Wannier function radius (KMG). b) Using a 7.5 Å radius for the buffer region (D&C) and the Wannier function radius (KMG). c) Using a 5.0 Å radius for the buffer region (D&C) and the Wannier function radius (KMG).

4.2. Semiconducting System

Bulk silicon has been chosen as the test case for the semiconducting system, having been previously widely studied using linear-scaling methods. The calculation was performed



(a) (b)

Figure 3. The CPU time scaling as a function of the number of atoms per supercell for a linear alkane chain, C_nH_{2n+2}. a) The D&C contribution to the first SCF iteration. b) A comparison between the KMG method and the D&C method. The KMG method's first SCF and average SCF iteration calculation times are shown.

using a 40 Rydberg cut-off for the real space integration grid used to represent the density, an energy shift of 0.01 Rydberg for the PAO orbital confinement, and a density matrix convergence criteria of 1x10^{-3}. The interation ranges within the Hamiltonian matrix vary from 9.3843 Å for the SZ basis set to 9.3843 Å for the DZP basis set. Again the PBE functional was used for the XC energy and potential. As in the insulating case, we have calculated the energy difference per atom between the D&C total energy and that obtained via full system diagonalization, see Table 3, as a function of basis set and buffer region size for a supercell consisting of 512 atoms.

As before, no dependence was found on the basis set used and that by increasing the subsystem size (i.e. the buffer region) the error in the total energy is reduced, with one exception discussed below. Commensurate with the smaller band gap and higher dimensionality of this system, the errors in the total energy are larger than in the insulating polymer case. Consequently, larger buffer regions are required to capture the decay length of the Wannier functions accurately. However, the use of partitions shorter than the interaction range of the Hamiltonian is still acceptable for at least qualitative results. Because the sparsity pattern of the density matrix in SIESTA is determined by that of the Hamiltonian, the computational penalty for using a large buffer radius only becomes particularly pronounced once this length scale is exceeded.

There is one discrepancy in the results, for the 8.0 Å buffer region size and DZP basis set the error in the total energy, -1.34620E-01 eV, is larger than errors found for decreasing buffer region sizes. In changing the radius from 7 to 8 Å two extra shells

Table 3. Energy differences (eV/atom) between divide and conquer and diagonalization for a bulk silicon supercell consisting of 512 atoms as a function of buffer radius and basis set size.

Number of Atoms	Buffer Region	Basis Set			
	(Å)	SZ[a]	SZP[b]	DZ[c]	DZP[d]
512	6.0	-4.879E-02	9.306E-03	5.570E-02	-7.512E-02
	7.0	1.751E-02	-9.124E-03	9.001E-02	-2.960E-02
	8.0	1.320E-02	-4.685E-03	3.115E-02	-1.346E-01

[a] Single-zeta. [b] Single-zeta + polarization. [c] Double-zeta. [d] Double-zeta + polarization.

of silicon atoms are included within the buffer region, comprising 28 atoms, as opposed to a single shell for the first transition. This demonstrates that the convergence with respect to buffer region is not guaranteed to be smooth and fluctations are likely to be particularly pronounced when all atoms are symmetry equivalent due to the extent of mixing in the bands on the system.

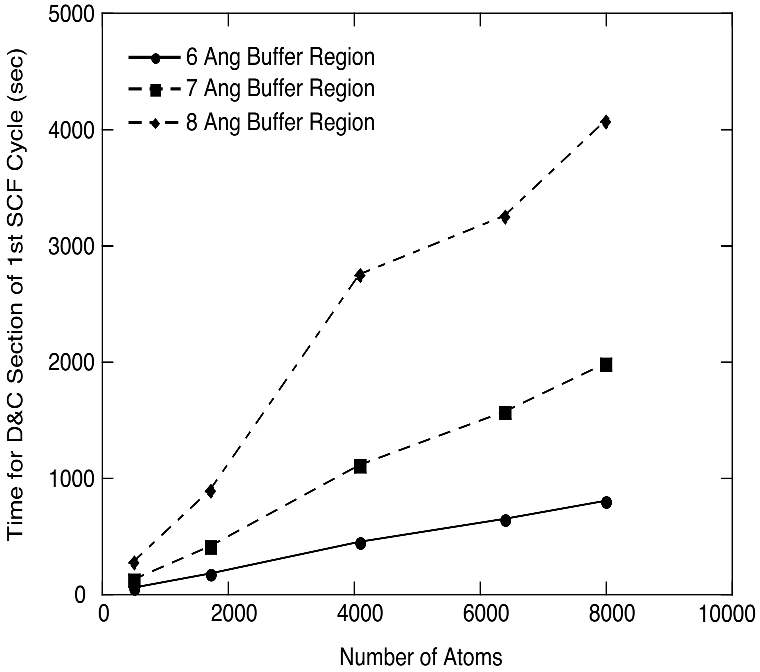


Figure 4. The CPU time scaling of a series of varying sized bulk Silicon. The contribution of the D&C section of the ocde to the first SCF iteration is shown.

The scaling performance of this system (with increasing atom numbers) is shown in Figure 4. The graph shows the calculation time for the D&C section of the first SCF cycle. The calculations examine the scaling from 512 atoms to 8000 atoms using the SZ basis set. For the 6.0 Å buffer region size, linear-scaling is evident with increasing system size. The 7.0 Å buffer region size calculations show linear-scaling beyond 4096 atoms but deviate below this. This behaviour is even more evident with the 8.0 Å buffer region size calculations, where there is approximately $O(N^3)$ scaling for the system sizes examined up to 4096 atoms and near linear-scaling for larger supercells. This discrepancy between 6400 atoms and 8000 atoms can not be currently resolved. We are assuming that it due to hardware issues and not the D&C method itself, as there is no indication form the other results that linear-scaling should not occur. The abscence of linear-scaling for small system sizes is due to the larger buffer region radii being greater than half the supercell length, based on a lattice constant of 5.43 Å for a single unit cell of silicon. Within this regime, each partition includes nearly all the atoms of the supercell and so the cubic scaling of the diagonalization for the partitions dominates. Once the unit cell length becomes greater than the buffer region diameter there is a progressive transition to the expected linear-scaling until the crossover point is reached at which divide and conquer becomes more efficient.

A comparison in calculation times between the D&C method and the KMG method is not reported here as the actual time to calculate the first SCF iteration within the KMG method is very large. Computing resources were not available for this comparison. We can report that for the KMG method a time of 4.73 hours was required for the first SCF iteration to complete for a Silicon system consisting of 512 atoms and using a SZ basis set with a Wannier radius of 6.0 Å. This time is well above the time required for a complete calculation (i.e. till convergence) for the same system using the D&C method. Although, the time for each SCF iteration will reduce in the KMG method the benefits of using the D&C method for this semiconducting system are noticeable.

4.3. Near Metallic System

This last test case was chosen to demonstrate the applicability of the D&C method for (near)metals. We have chosen a (5,5) armchair single walled carbon nanotube (SWNT) for this purpose. The calculations were performed using the PBE functional with a 100 Rydberg cutoff for the density integration mesh, 0.02 Rydberg for the PAO energy shift and a density matrix convergence criteria of 1x10^{-4}. The resulting interaction ranges within the Hamiltonian vary from 7.3030 Å for the SZ basis set to 7.4416 Å for the DZP basis set.

As in the previous two cases we have calculated the variation of the error in the total energy with respect to different basis sets and buffer region sizes. The test system consisted of a 1000 atoms within the one-dimensional supercell. The results are summarized in Table 4. The trends in the total energy with partition radius are less well defined for the present system, as would be expected to the longer decay length. For

Table 4. Energy difference (eV/atom) between divide and conquer and diagonalization as a function of buffer region radius and basis set quality for a single walled (5,5) carbon nanotube.

Number of Atoms	Buffer Region	Basis Set			
	(Å)	SZ[a]	SZP[b]	DZ[c]	DZP[d]
1000	5.1121	1.194E-02	1.100E-03	-3.409E-02	-7.250E-02
	5.8424	-8.730E-03	-3.894E-03	-2.499E-02	-3.111E-02
	7.3030	2.272E-03	-1.335E-03	-1.315E-02	-1.225E-02

[a] Single-zeta. [b] Single-zeta + polarization. [c] Double-zeta. [d] Double-zeta + polarization.

the DZ and DZP basis sets the error does consistently decrease with increasing radius, though slowly, while for the SZ basis set the absolute magnitude decreases, but with the sign oscillating. For the SZP there is no apparent convergence within the range of radii examined and a more extensive exploration of larger radii is required. Despite the lack of a clear and rapid decay in error with radius, the magnitude of the difference from the full diagonalization results, per atom, is comparable to that of thermal energy at ambient conditions and so higher levels of convergence may not be required for all calculations.

Figure 5 shows the scaling of the calculation times of the D&C section which contributes to the first SCF cycle with increasing system size. The SZ basis set was used for all the timing calculations. For all buffer region sizes the scaling is indeed found to be linear.

To reduce the error in the total energy larger buffer region sizes are required. The timing results show that by increasing the buffer region slighty, as shown by the transition from a region radius of 5.8 Å to 7.3 Å , this will increase the calculation time considerably. This requirement of a larger buffer region will inhibit the use of the D&C method for small metallic systems. The so called crossover point, where it is computational beneficial to use the D&C method rather than conventional techniques, is pushed out to larger problems, which makes the use of the D&C method really only applicable to fairly large near-metallic systems. Using different partition schemes that produce smaller numbers of the subsystems can remedy this problem and further work is in progress in this area.

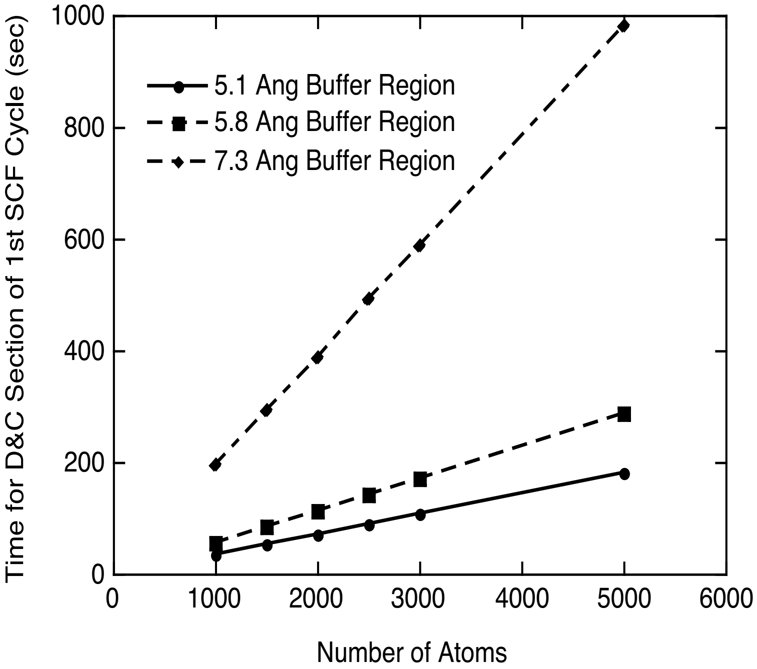


Figure 5. The CPU time scaling of a series of varying length (5,5) single walled carbon nanotube. The contribution of the D&C section implemented within the code to the first SCF iteration is shown.

4.4. Parallelisation

The parallel performance of the D&C implementation was tested on the bulk silicon system for a supercell containing 21902 atoms. Using a single-zeta basis set, 40 Rydberg mesh cut-off for the integration grid, a PAO energy shift of 0.02 Rydberg and a buffer region radius of 6.08 Å, the test examined the parallel performance from 1 processor to 32 processors.

All calculations were executed using the memory conservation option (see section 3.2). Figure 6 shows that the speedup gained from using larger numbers of processors is nearly perfect relative to the calculation time for a single processor. For 32 processors, the speed up of 31.78 times is very close to the ideal value of 32. This indicates that the computational effort is indeed dominated by the diagonalization of the subsystems, which is embarrassingly parallel, while the computational of the Fermi energy and build of the Hamiltonian matrices, where communication is required, represents a small overhead. Similar results were obtained by Pan *et al* [33] with their parallel implementation of the D&C method.

It should be noted that for this specific case the load balancing is perfect, i.e. in all cases each compute node has an equal number of subsystems of equal size due to the high symmetry of the problem. This is an important factor in contributing to the near perfect speedup. However, perfect load balancing will not always occur in practice with the present scheme for systems with inhomogeneous density or atom type distributions.

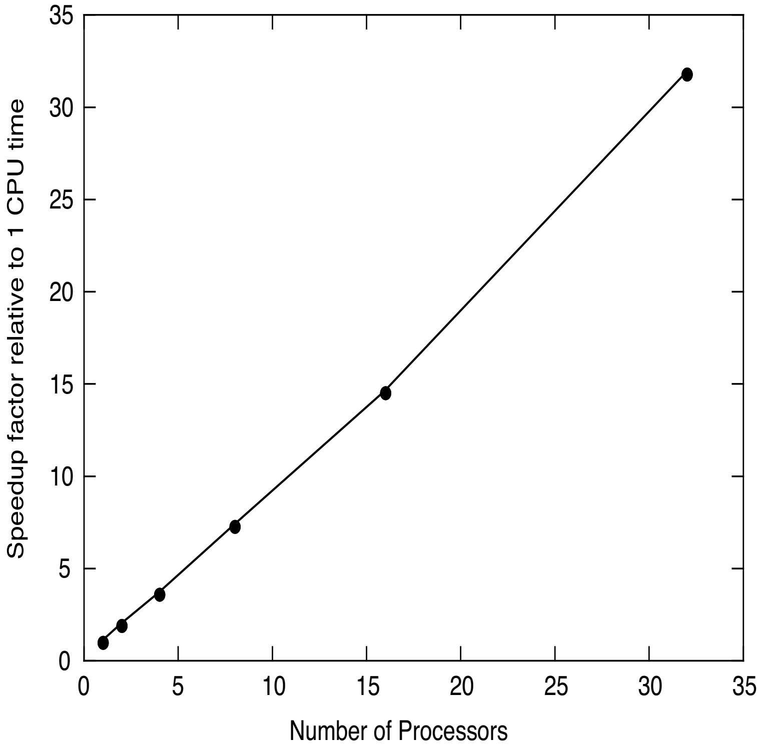


Figure 6. Parallel performance of the D&C implementation when studying a bulk silicon supercell containing 21952 atoms. Shown here is the speedup acquired when increasing the number of processors relative to a single processor calculation.

Further refinement of the implementation is under way to ensure improved load balance for all systems.

5. Concluding Remarks

We have successfully combined the density-matrix D&C scheme with the SIESTA methodology for computing the Hamiltonian and overlap matrices. Our implementation exhibits linear-scaling within the D&C scheme, provided the dimensions of the physical system exceed those of the allowed range for the localized states. The applicability to a variety of systems with varying band gaps has been demonstrated, including a near-metallic carbon nanotube. This scheme will allow practical electronic structure calculations of very large systems, consisting of thousands to tens of thousands of atoms, with relatively modest computational resources. While the results of the divide and conquer scheme are comparable to those currently obtained with the Kim-Mauri-Galli algorithm in SIESTA, the robustness of the approach leads to it being advantageous for systems with small band gaps, and therefore a valuable alternative approach to achieving linear-scaling within the SIESTA methodology. When executed in parallel for large systems the divide and conquer approach exhibits near perfect speedup, providing there is appropriate load balancing.

References

[1] Kohn W and Sham LJ 1965 *Phys. Rev.* **140** A1133

[2] Payne MC, Teter MP, Allan DC, Arias T and Joannopoulos JD 1992 *Rev. Mod. Phys.* **64** 1045-1097

[3] Andzelm J and Wimmer E 1992 *J. Chem. Phys.* **96** 1280-1303

[4] Cho K, Arias TA, Joannopoulos JD and Lam PK 1993 *Phys. Rev. Lett.* **71** 1808-1811

[5] Chelikowsky JR, Troullier N, Wu K and Saad Y 1994 *Phys. Rev. B* **50** 11355-11364

[6] Hernandez E and Gillan MJ 1997 *Phys. Rev. B* **55** 13485

[7] Skylaris C-K, Haynes, Mostofi AA and Payne MC 2005 *J. Chem. Phys.* **122** 084119

[8] Delley B 1990 *J. Chem. Phys.* **92** 508-517

[9] Sambe H and Felton RH 1975 *J. Chem. Phys.* **62** 1122-1126

[10] Greengard L and Rokhlin V 1987 *J. Comput. Phys.* **73** 325

[11] Merrick MP, Iyer KA and Beck TL 1995 *J. Phys. Chem.* **99** 12478-12484

[12] Soler JM, Artacho E, Gale JD, Garcia A, Junquera J, Ordejon P and Sanchez-Portal D 1992 *J. Phys. Cond. Mat.* **14** 2745-2780

[13] Kenny SD, Horsfield AP and Fujitani H 2000 *Phys. Rev. B* **62** 4899-4905

[14] Sankey OF and Niklewski DJ 1989 *Phys. Rev. B* **40** 3979

[15] Junquera J, Paz O, Sanchez-Portal D and Artacho E 2001 *Phys. Rev. B* **64** 235111

[16] Challacombe M *J. Chem. Phys.* **110** 2332-2342

[17] Yang W 1991 *Phys. Rev. Lett.* **66** 1438-41

[18] Yang W 1991 *Phys. Rev. B.* **44** 7823-26

[19] Yang W and Lee TS 1991 *J. Chem. Phys.* **103** 5674-78

[20] Mauri F, Galli G and Car R 1993 *Phys. Rev. B* **47** 9973-9976

[21] Li XP, Nunes RW and Vanderbilt D 1993 *Phys. Rev. B* **47** 10891

[22] Palser AHR and Manolopoulos DE 1998 *Phys. Rev. B* **58** 12704-12711

[23] Niklasson AMN 2002 *Phys. Rev. B* **66** 155115

[24] Kim J, Mauri F and Galli G 1995 *Phys. Rev. B* **52** 1640-1648

[25] Shimojo F, Kalia RK, Nakano A and Vashishta P 2005 *Comp. Phys. Comm.* **167** 151-164

[26] Dixon SL and Merz Jr KM 1996 *J. Chem. Phys.* **104** 6643-6649

[27] Dixon SL and Merz Jr KM 1996 *J. Chem. Phys.* **107** 879-893

[28] Lee TS, Lewis JP and Yang W 1998 *Comp. Mat. Science* **12** 259-277

[29] Lee TS and Yang W 1997 *Int. J. Quant. Chem.* **69** 397-404

[30] Ermolaeva MD, van der Vaart A and Merz Jr KM 1999 *J. Phys. Chem. A* **103** 1868-1875

[31] Zhu T, Pan W and Yang W *Phys. Rev. B* **53** 12713-12724

[32] Zhu T, Pan W and Yang W *Theor. Chem. Acc.* **96** 2-6

[33] Pan W, Lee TS and Yang W 1998 *J. Comp. Chem.* **19** 1101-1109

[34] Kohn W 1996 *Phys. Rev. Lett.* **76** 3168-3171

[35] Kohn W 1993 *Chem. Phys. Lett.* **208** 167-172

[36] Geller MR and Kohn W 1993 *Phys. Rev. B* **48** 14085-14088

[37] Mulliken RS 1962 *J. Chem. Phys.* **36** 3428

[38] Perdew JP, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865

[39] Blackford LS, Choi J, Cleary A, D'Azevedo E, Demmel, J, Dhillon I, Dongarra J, Hammarling S, Henry G, Petitet A, Stanley K, Walker D and Whaley RC 1997 *ScaLAPACK Users' Guide* Society for Industrial and Applied Mathematics, ISBN 0-89871-397-8

[40] Warschkow O, Dyke JM and Ellis DE 1998 *J. Comp. Phys.* **143** 70-89

Acknowledgments

This work was supported by the Australian Research Council, the University of Technology, Sydney and Curtin University of Technology. Computing resources were provided by the Australian Centre for Advanced Computing and Communication (ac3), iVEC, and the national facility at the Australian Partnership for Advanced Computing (APAC). JDG would like to thank the Government of Western Australia for a Premier's Research Fellowship.