Band Parallelism in CASTEP: Scaling to More Than 1000 Cores

P.J. Hasnip[1] *M.I.J. Probert*[1] *K. Refson*[2] *M. Plummer*[3] *M. Ashworth*[3]
[1] University of York (UK); [2] STFC, Rutherford Appleton Laboratory (UK);
[3] STFC, Daresbury Laboratory (UK)

May 1, 2009

ABSTRACT: CASTEP is the UK's premier quantum mechanical materials modeling code. We describe how the parallelism is implemented using a 3-layer hierarchical scheme to handle the heterogeneously structured dataset used to describe wave functions. An additional layer of data distribution over quantum states (bands) has enhanced the scaling by a factor of 4-8, allowing many important scientific calculations to efficiently use over a thousand cores on the HECToR XT4 service.

KEYWORDS: CASTEP, density functional theory, hierarchical parallelism

1 Introduction

CASTEP[1] is a density-functional theory[2] program for calculating the groundstate electronic charge density of periodic systems of electrons and nuclei. It is widely used in materials science, chemistry, physics, engineering and, increasingly, molecular biology research, in both industrial and academic research institutions.

The main computational effort in a CASTEP calculation is the solution of a set of generalised Kohn-Sham[3] eigenvalue equations corresponding to different sampling points ('k-points') of the reciprocal-space Brillouin zone of a periodic simulation cell,

$$\hat{H}_k \psi_{bk}(\vec{r}) = E_{bk} \hat{S}_k \psi_{bk}(\vec{r}) \tag{1}$$

where $\hat{H}_k$ is the Hamiltonian at k-point k, $\{\psi_{bk}(\vec{r})\}$ are the eigenstates, known as *bands*, $\{E_{bk}\}$ are the corresponding band-energies (i.e. the eigenvalues), and $\hat{S}_k$ is the overlap operator required when using ultrasoft pseudopotentials to describe the atoms.

The Hamiltonian is not a fixed operator but contains terms that depend on the electronic charge density n, which is given by

$$n(\vec{r}) = \sum_{bk} f_{bk} \int |\psi_{bk}(\vec{r})|^2 d^3k \tag{2}$$

where f_{bk} is the occupancy of band b at k-point k. Unless the system is spin-polarised, each band can hold up to two electrons; thus for an insulating system with N_e electrons, the lowest $\frac{1}{2}N_e$ bands are fully occupied ($f_{bk} = 2$) and all higher bands are unoccupied ($f_{bk} = 0$), so do not contribute to the density.

Since $\hat{H}_k = \hat{H}_k[n]$, the equations for different k-points are not completely independent, but interact via this electronic charge density $n(r)$. For the periodic systems CASTEP is designed to simulate, the density shares the same periodicity as the simulation system and so it is natural to express it in a Fourier basis. The basis functions are plane-waves whose wave-vectors are reciprocal lattice vectors of the simulation cell, the so-called G-vectors. The bands themselves need not be quite periodic, but they can always be written as the product of a constant phase factor $e^{i\vec{k}.\vec{r}}$ and a periodic function:

$$\psi_{bk}(\vec{r}) = e^{i\vec{k}.\vec{r}} \left(\sum_G c_{Gbk} e^{i\vec{G}.\vec{r}} \right) = \sum_G c_{Gbk} e^{i(\vec{G}+\vec{k}).\vec{r}} \tag{3}$$

It is these arbitrary phases, the 'k-points', that we sample, approximating the integral in equation 2 by a summation over discrete k-vectors drawn uniformly from the first Brillouin zone.

The summation over G-vectors in equation 3 is truncated so that only plane-waves whose G-vectors satisfy $|\vec{G} + \vec{k}| \leq G_{cutoff}$ are included in the basis set; thus the G-vectors of the included plane-waves lie within a sphere in reciprocal-space, centred on the origin. Care must be taken to ensure that all computed properties are converged with respect to both the k-point sampling and G_{cutoff}.

The aim of this computation is to compute the N_b bands with the lowest eigenvalues (where N_b is usually slightly more than $\frac{1}{2}N_e$) in order to determine the groundstate of the system – that is, the state which has the lowest energy. Once the groundstate has been determined, the density n is the groundstate electronic charge density which, together with the bands ψ_{bk}, can be used to compute almost any chemical, optical, mechanical or electronic property of the system. In this way programs such as CASTEP have been used successfully to predict new crystal structures, interpret NMR spectra, find new catalysts, investigate corrosion and crack propagation and develop new drugs, amongst many other applications.

1.1 Computational implementation

In the plane-wave basis, each band is completely defined by the complex coefficients $\{c_{Gbk}\}$. All of the bands together, across all of the k-points, make up the total Kohn-Sham wavefunction. CASTEP is written in Fortran 95, and it is this wavefunction that is the fundamental data object. CASTEP's `wave` module defines this as a derived datatype, a simplified version of which is given below:

```
type, public :: wavefunction
  complex(kind=dp), dimension(:,:,:) :: coeffs
  integer,          dimension(:)     :: waves_kp
  integer                            :: nbands
  integer                            :: nkpts
end type wavefunction
```

where the `coeffs` array holds all of the complex coefficients c_{Gbk}, is allocatable, and is dimensioned:

```
wvfn%coeffs(1:max_waves,1:nbands,1:nkpts)
```

The `wave` module also contains subroutines to perform basic operations on wavefunction data objects, such as addition, dot-products, orthogonalisation and so on.

Thus for a given k-point, each band is a vector of length `max_plane_waves`$=N_G$, and the Hamiltonian operator is a matrix of size $N_G \times N_G$.

Since the number of plane-waves can be $O(100,000)$ or more, the Hamiltonian matrix is too large either to store or apply explicitly; however it may be expressed as the sum of three terms,

$$\mathrm{H} = \mathrm{H}_{\mathrm{KE}} + \mathrm{H}_{\mathrm{local}} + \mathrm{H}_{\mathrm{nl}} \tag{4}$$

where H_{KE} is diagonal in the plane-wave basis (G-space); $\mathrm{H}_{\mathrm{local}}$ is diagonal in real-space (and is the term which depends on the density n); and H_{nl} is a low-rank correction. Each of these three terms uses only $O(N_G)$ storage.

Whenever CASTEP has to apply the Hamiltonian to the bands, it applies H_{KE} and H_{nl}, then Fourier-transforms the bands to real-space to apply $\mathrm{H}_{\mathrm{local}}$, and then back-transforms to G-space to recover $\mathrm{H}\psi$ in the plane-wave basis.

Since the Hamiltonian is so large, diagonalising it directly is too expensive to be practical, and yields far more eigenstates than the lowest few we require; hence the search for the groundstate is an iterative procedure to compute just the lowest N_b eigenstates. A trial set of bands is improved by converting the eigenvalue problem into a minimisation problem, and computing an appropriately preconditioned gradient[4]. CASTEP then seeks an improved set of eigenstates in the space spanned by the initial trial set and these preconditioned gradient directions.

In the density mixing (DM) methodology[5] this improvement of the approximate eigenstates is performed for a *fixed* Hamiltonian. A new density is then constructed from the improved eigenstates, but rather than using this directly, CASTEP employs a density mixer to produce a new estimate for the true groundstate density (see Figure 1). Finally the new Hamiltonian is constructed, and the process is repeated until the change in the total energy of the system falls below a given threshold.

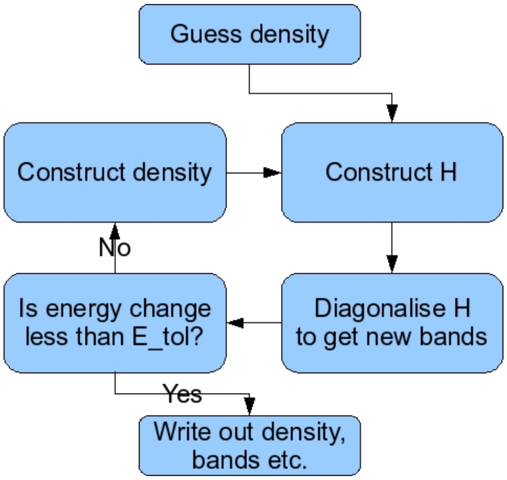

Figure 1: Diagram showing the basic CASTEP self-consistent field (SCF) cycle for the density mixing algorithm.

1.2 Existing parallel strategies

Some CASTEP calculations involve computing properties from several related groundstate calculations, e.g. path-integral molecular dynamics simulations. Such calculations are trivial to parallelise, since each groundstate calculation is completely independent of the others. At this crude level of parallelism CASTEP creates a 'task farm', and gives a different groundstate calculation to each 'farm'. Aside from the occasional load-balancing issues, these calculations scale perfectly at this level of parallelism, but only certain kinds of calculation can exploit it; of more interest, therefore, is the parallelisation of the basic CASTEP groundstate calculation, which is the focus of this work.

Since the bands at different k-points are almost independent of each other, it is natural to distribute the data over the k-points. Calculations scale almost perfectly with this distribution of data. As the size of simulation system is increased the Brillouin zone, which is a region of reciprocal space, decreases. Since it is this Brillouin zone that the k-points sample, the number of k-points required for a given accuracy also decreases with increasing system size, and for large systems N_k is O(1). This means that on HPC machines it is rarely possible to distribute the data and workload sufficiently using k-point parallelism alone.

A further distribution strategy used by CASTEP is to distribute the data by the G-vectors. Since the number of plane-waves is often large, and grows approximately linearly with simulation cell size, this enables efficient data distribution and load-balancing. The 3D G-vector grid is distributed by columns, so that each processing element (PE) holds one or more complete columns of the data. The 3D Fourier transform is then performed as three 1D Fourier transforms, with an explicit data transposition in between each transformation. Unfortunately these transpositions involve all-to-all communications across the PEs containing the G-vectors, and these communication costs limit the number of PEs that may be used efficiently.

In this way the PEs in a parallel computation are arranged logically in a 3D PE grid, with tasks farmed along one axis, k-points distributed along another, and G-vectors distributed along the third axis. By combining both k-point and G-vector parallelism CASTEP has demonstrated excellent scaling properties from 1 to O(100) PEs across many different computer architectures. Calculations capable of using the task farming parallelism can extend this trivially to O(1000) PEs.

1.3 Computational Costs

The dominant costs for CASTEP simulations of large systems are the orthogonalisation of the trial bands to each other, which is necessary in the iterative diagonalisation scheme to prevent all the trial bands from collapsing to the one with the lowest eigenvalue. Similar operations are also used to diagonalise the Hamiltonian in the subspace of the current trial bands. These operations involve computing an overlap matrix between all pairs of bands, which scales as $N_G N_b^2 N_k$ for a system with N_G G-vectors, N_b bands and N_k k-points. This operation is distributed naturally over G- and k-points, with only a single reduction over G-vectors required to ensure all PEs have the complete overlap matrix for each of their k-points. N_G and N_b both grow linearly with system size but N_k decreases, and for very large systems $N_k = 1$; thus the asymptotic scaling with system size is cubic.

Once the overlap matrix has been computed for a given k-point, it is used to compute an appropriate band-transformation matrix. Depending on the desired operation this may require matrix inversion or diagonalisation, but in either case the operations scale as N_b^3. This work is not distributed over the G-vectors, only the k-points, so as larger and larger systems are studied these operations will start to become expensive. In CASTEP these operations are performed using standard LAPACK subroutine calls, and in this work Cray's LibSci version 10.2.1 was used.

Once the appropriate transformation matrix has been computed, the transformation is applied to the bands. This operations also scales as $N_G N_b^2 N_k$, but is distributed naturally over the G-vectors and k-points with no communications required between PEs.

The Fourier transforms are carried out using conventional Fast Fourier Transform (FFT) libraries, and the time to FFT the whole wavefunction scales as $N_G \ln(N_G) N_b N_k$. Although the scaling with N_G and N_b is favourable compared to that of orthogonalisation, the all-to-all communications required for the data transpositions limit the number of PEs that can be used efficiently.

The length of the FFTs in CASTEP is typically small, and even for large simulation cells it will usually be O(100); however the total number of FFTs required in a simulation can be extremely large. The vast number of messages coupled with their short length means that in CASTEP these communication stages are usually latency-bound, rather than bandwidth-bound. In this work FFTW 3.1.1 was used throughout.

Because of the problems with the all-to-all communications, CASTEP's communication in the transpose stage has been optimised for machines with multiple PEs per node[6] by partitioning the communications explicitly into intra-node and inter-node. In this scheme the PEs on each node are split into sub-groups, and one PE in each sub-group is designated the master PE. This master PE gathers the data from all PEs within the same sub-group. The inter-node phase then handles the global data exchange, but this now involves only the master PEs, rather than all PEs. Finally the master PEs broadcast the exchanged data to the other PEs in their sub-groups.

For small numbers of nodes the overhead of performing multiple communication phases is prohibitive, but as the all-to-all communications start to dominate the improved scaling of this method more than makes up for the additional cost. For nodes with several PEs the flexibility in choice of the size of the sub-groups allows for maximum performance taking account of both fast internal communication and optimal inter-node message size. All of the work presented here was carried out on the HECToR XT4, which has only two PEs per node at present.

2 CASTEP Performance on the XT4

All program development and benchmarking was performed with the CASTEP 4.2 codebase, which was the most recent release of CASTEP at the commencement of this work. The code was compiled with version 3.0 of Pathscale's fortran compiler, and linked with Cray's LibSci 10.2.1 and FFTW 3.1.1. CASTEP uses MPI for its parallel communications. All development and testing was performed on the HECToR XT4.

The performance of CASTEP 4.2 was measured using two benchmarks from the standard CASTEP benchmark suite: a 33-atom titanium nitride simulation (which also contains a single hydrogen atom); and a 270-atom aluminium oxide slab calculation. Neither benchmark exploits the task-farming parallelism.

Figure 2 shows the performance of CASTEP for the TiN benchmark on the HECToR XT4 using both the one-phase (SMP=1) and multi-phase (SMP=2) communication modes. This benchmark has 8 k-points, and so scales near-perfectly from 1 to 8 PEs. Beyond 8 PEs CASTEP has to distribute the data and workload by G-vector in addition to k-point, which means each 3D Fourier transform now requires 3 1D FFTs with 2 transpositions. This change in algorithm accounts for the sharp loss in performance from 8 to 16 PEs, but if we allow for this drop then the calculation scales well up to 128-256 PEs. There are approximately 11,000 plane-waves for this calculation, so allowing for the 8-way k-point parallelism, 128-256 PEs corresponds to about 350-700 plane-waves per PE.

Figure 3 shows the performance for the aluminium oxide slab benchmark. This benchmark has only 2 k-points, but 88,000 plane-waves and the memory limitations were such that it could not be run on fewer than 16 PEs of HECToR.

3 Band Parallelism

The main data object in CASTEP is the wavefunction, which contains the data for all of the bands at all k-points. The array of coefficients

```
wvfn%coeffs(1:max_waves,1:nbands,1:nkpts)
```

is already distributed over the plane-waves and the k-points; the aim of this work is to distribute the data and workload by bands as well. The PEs will now be arranged in a 4D logical grid, and the $N_G \times N_b \times N_k$ coefficients array for the bands will be completely distributed.

3.1 Implementation

CASTEP is written in modular Fortran 95, and the details of the communications are abstracted by the `comms` module. A new communicator was defined in this module (along with the basic operations involving this communicator), and support for BLACS contexts was also added to enable the use of ScaLAPACK elsewhere in the code.

Aside from the `comms` module, the modular structure of CASTEP meant that the non-trivial code changes were restricted to just 2 of the 54 modules: the `algor` module, which handles matrix inversion and diagonalisation; and the `wave` module, which

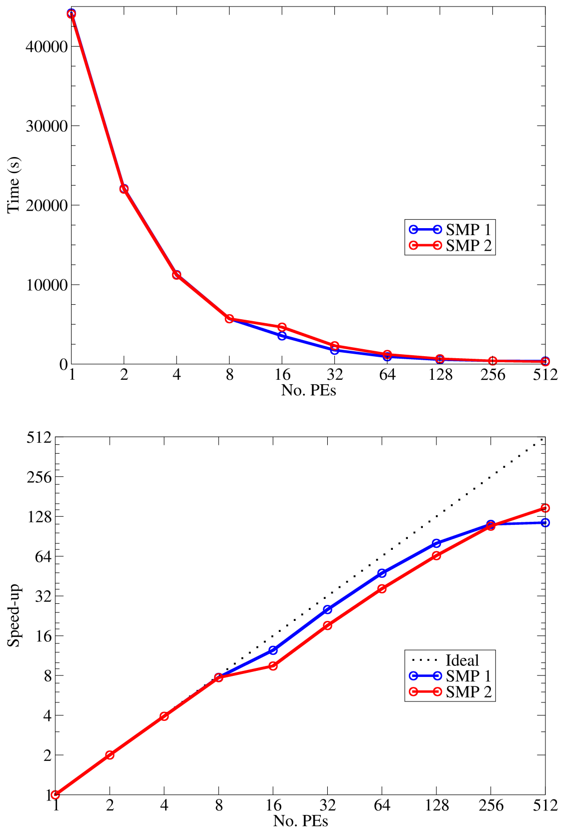

Figure 2: Execution time and corresponding speed-up for the 33 atom TiN benchmark on the HECToR XT4, using the default communication pattern (SMP 1) and the multi-phase communications (SMP 2) in the FFT. This calculation is performed at 8 k-points, and shows near-linear scaling in this regime.

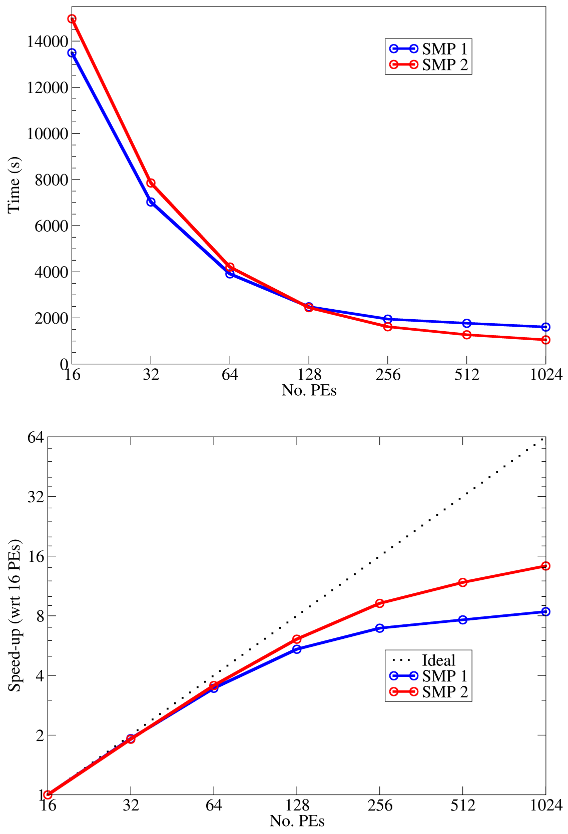

Figure 3: Execution time and corresponding speed-up for the 270 atom Al_2O_3 slab benchmark on the HECToR XT4, using the default communication pattern (SMP 1) and the multi-phase communications (SMP 2) in the FFT. This calculation is performed at 2 k-points.

handles the band data itself and basic operations on it.

The changes to the `algor` module were straightforward. Using the BLACS contexts provided by the comms module, each LAPACK call was replaced by its ScaLAPACK counterpart.

The band data in the `wave` module was distributed in a simple round-robin fashion to facilitate load balancing. If the bands are distributed over B PEs, each PE i holds all data for band i, and every Bth band from there onwards. Most of the data associated with the bands, from the eigenvalues to the Fourier coefficients themselves, are distributed naturally in this fashion.

Constructing the band-overlap matrix (required for orthogonalisation) involves computing dot-products between every pair of bands. With the bands distributed amongst the PEs, this now requires all-to-all communications amongst those PEs with the same G-vector and k-point data. Once the band-overlap matrix is constructed, the appropriate transformation is computed using the ScaLAPACK subroutines in `algor`. These overlap and transformation matrices are always distributed over the fastest index, such that each PE contains all of the transformation data required to transform its own band data.

3.2 Testing

The changes to many of the CASTEP modules were trivial, and needed little testing. The changes to the `wave` module were rather more extensive, and difficult to test in isolation. In order to test and debug these modules, a short CASTEP calculation was run in serial to create a wavefunction and density in a CASTEP `.check` file, and then the calculation was restarted from that checkpoint file in serial and band-parallel modes. Because both jobs started from the same known point the calculations could be compared in detail, right down to individual wavefunction coefficients where necessary. The serial and band-parallel calculations gave computed eigenvalues which only differed by $O(10^{-14})$eV; as this is the same order as machine precision (ϵ) for double-precision arithmetic on the XT4, this may be attributed to rounding errors.

4 Results

4.1 Aluminium oxide benchmark

With the new distribution over bands as well as G-vectors and k-points, the performance and scaling of CASTEP was improved significantly. Using 8-way band parallelism, CASTEP calculations now scale effectively to between four and eight times more PEs compared to CASTEP 4.2 (see figure 4) on the standard aluminium oxide slab benchmark. Indeed this benchmark can now be run on 1024 PEs with almost 50% efficiency, which equates to over three PEs per atom.

4.2 Immidazolium chloride

Immidazolium chloride is a room-temperature ionic liquid, and a 408 atom simulation cell was constructed in order to simulate its behaviour over time using CASTEP's molecular dynamics capabilities. Unfortunately CASTEP 4.2 was too slow for this to be practical, which made this system an ideal test case for the new parallel scheme.

The immidazolium chloride simulation has over 130,000 plane-waves and only a single k-point, which restricted CASTEP 4.2 to G-vector parallelism. The results for CASTEP 4.2 and the new band-parallel version are shown in figure 5; this system could only be run on 64 PEs or more of HECToR due to memory limitations. It can be seen clearly that although performance of the band-parallel CASTEP was significantly worse for 64 and 128 PEs, its greatly improved scaling means substantial speed-ups even up to 1024 PEs, despite the lack of k-point parallelism.

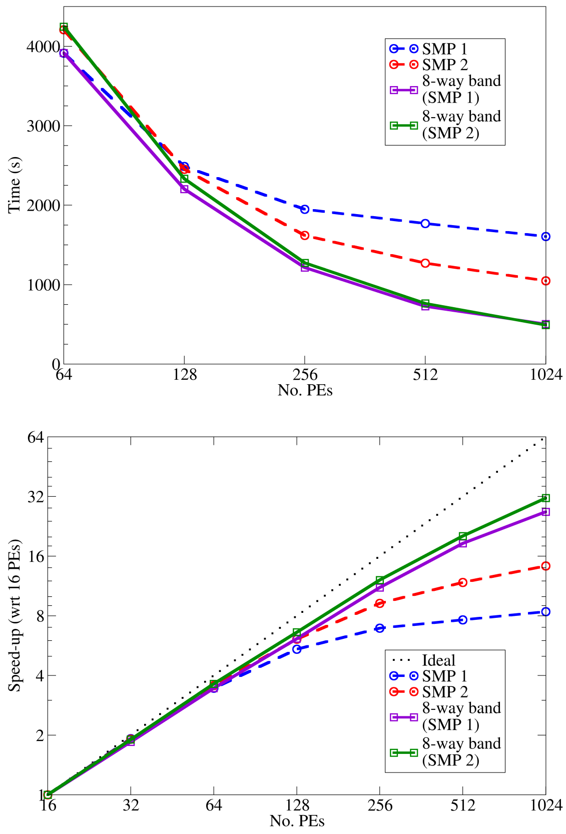

Figure 4: Comparison of CASTEP scaling for the new band-parallel CASTEP 4.2 (using 8-way band parallelism) and the original CASTEP 4.2, for 10 SCF cycles of the Al_2O_3 slab benchmark. Parallel speed-up was measured relative to the 16 PE calculations.

In particular the reduction in simulation time to 16 minutes on 1024 PEs means that long molecular dynamics simulations may now be possible on this system.

4.3 Performance analysis

In the immidazolium chloride test, the performance of the band-parallel CASTEP is substantially poorer than that of CASTEP 4.2 for smaller numbers of PEs. This is attributable to two main factors: limitations ScaLAPACK; and limitations in the band-transformation code.

Almost all of the overlap matrices computed by CASTEP are either Hermitian, or positive definite, and when inverting or diagonalising these matrices

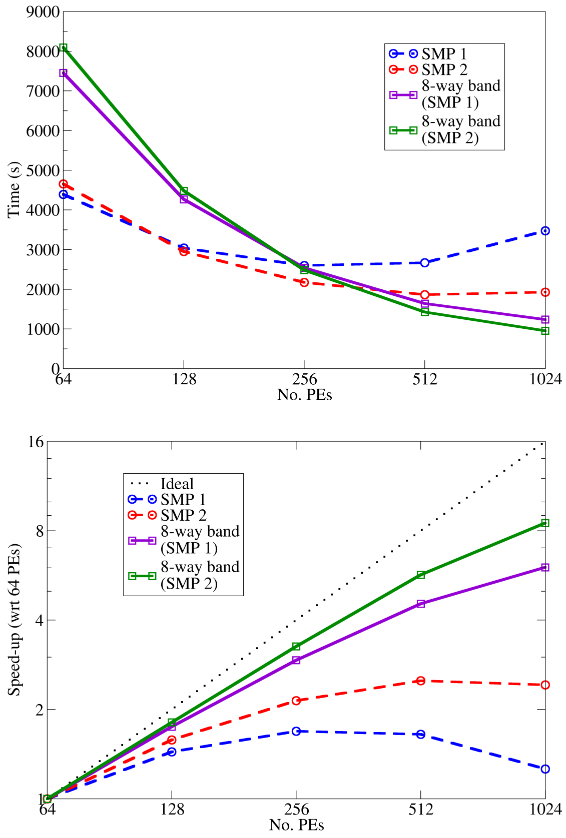

Figure 5: Comparison of CASTEP scaling for the new band-parallel CASTEP 4.2 (using 8-way band parallelism) and the original CASTEP 4.2, for 20 SCF cycles of the immidazolium chloride system. Parallel speed-up was measured relative to the 64 PE calculations.

these properties are exploited by the LAPACK subroutines to improve performance greatly. Unfortunately ScaLAPACK lacks analogues of many of these subroutines, and when running in band-parallel mode the general matrix subroutines must be used instead.

A similar limitation is found in CASTEP's own band-transformation subroutines. The inverting transformations computed by LAPACK from the positive-definite overlap matrices during orthogonalisation are always triangular, so CASTEP 4.2 can exploit this to save a factor of 2 when applying this unitary rotation. When the bands are distributed amongst the PEs however, exploiting this property of the transformation becomes more difficult.

The band-parallel scaling of CASTEP is limited by two factors: all-to-all communications; and load balancing. The band-transformations required for orthogonalisation mix all of the bands, so that following the transformation of each PEs local data, each PE has contributions for every other PE. This requires all-to-all communications and, as expected, this becomes costly as the bands are distributed over more and more PEs.

The load balancing issue is a subtle problem. As CASTEP improves the bands iteratively, some bands converge faster than others. When a band is deemed to have converged, CASTEP 4.2 stops calculating preconditioned gradients for them so that at any point CASTEP is only expending effort on the unconverged bands. However it is not possible to predict *a priori* which bands will converge slowest, so when CASTEP is run in band-parallel mode, the performance is limited by the PE with the largest number of unconverged bands. When each PE holds a large number of bands the convergence properties tend to average out, but as the bands are distributed over more and more PEs this load imbalance can become a severe problem.

5 Conclusions

The original goal of the work was to improve the scaling of CASTEP with PEs on the XT4 to enable ordinary groundstate calculations to be run on O(1000) PEs. This has been achieved by implementing band-parallelism on top of the existing CASTEP parallel strategies, coupled with efficient use of ScaLAPACK for linear algebra.

The considerable computational resources of XT4 machines such as HECToR, coupled with these developments in CASTEP, should enable accurate plane-wave DFT calculations on a far greater scale than has been possible to this point.

6 Acknowledgements

This work was supported by a distributed Computational Science and Engineering (dCSE) project grant from EPSRC (UK). PJH wishes to thank Christof Vömel and Edward Smyth for providing access to, and assistance with, a new parallel matrix diagonalisation subroutine, and all at NAG (Oxford, UK) for several useful discussions.

7 About the authors

Dr Phil Hasnip (pjh503@york.ac.uk) is a theoretical condensed matter physicist in the Condensed Matter Dynamics research group at the University of York. He is one of the authors of the CASTEP program, with a particular interest in the fundamental algorithms employed. Contact telephone +44-1904-434624.

Dr Matt Probert (mijp1@york.ac.uk) leads the Condensed Matter Dynamics research group in the Physics Department of the University of York. He is one of the authors of the CASTEP program, with a special interest in genetic algorithms and molecular dynamics simulations. Contact telephone +44-1904-432239.

Correspondance for Dr Hasnip or Dr Probert should be addressed to Physics Dept, University of York, York, YO10 5DD, United Kingdom.

Dr Keith Refson (keith.refson@stfc.ac.uk) is a theoretical physicist in the Computational Materials Science group at the STFC Rutherford Appleton Laboratory. He is one of the authors of the CASTEP program, with particular responsibility for the phonon and linear response code, as well as an interest in parallel optimisations. Contact telephone +44-1235-778023. Correspondance for Dr Refson should be addressed to STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire, OX11 0QX, United Kingdom.

Dr Martin Plummer (martin.plummer@stfc.ac.uk) is a computational scientist and group leader for Theoretical Atomic and Molecular Physics in the Advanced Research Computing Group, Computational Science and Engineering, at STFC Daresbury Laboratory. He also optimises and manages various codes on the UK HPC service HPCx, with particular responsibility for CASTEP. Contact telephone +44-1925-603244.

Dr Mike Ashworth (mike.ashworth@stfc.ac.uk) leads the Advanced Research Computing Group, Computational Science and Engineering, at STFC Daresbury Laboratory. His own work focuses on the development and optimisation of environmental modelling and CFD codes, including performance engineering and application of Grid technologies. Contact telephone number +44-1925-603178.

Correspondance for Dr Plummer or Dr Ashworth should be addressed to STFC Daresbury Laboratory, Warrington, Cheshire, WA4 4AD, United Kingdom.

References

[1] S.J. Clark, M.D. Segall, C.J. Pickard, P.J. Hasnip, M.J. Probert, K. Refson and M.C. Payne, Zeit. für Kryst. **220(5-6)** (2004) 567–570

[2] P. Hohenberg and W. Kohn, Phys. Rev. **136** (1964) 864B

[3] W. Kohn and L. Sham, Phys. Rev. **140** (1965) 1133A

[4] P.J. Hasnip and C.J. Pickard, Comp. Phys. Comm. **174** (2006) 24–29

[5] P. Pulay, Chem. Phys. Lett. **73** (1980) 393–398

[6] M. Plummer, J. Hein, M.F. Guest, K.J. d'Mellow, I.J. Bush, K. Refson, G.J. Pringle, L. Smith and A. Trew, J. Mater. Chem. **16** (2006) 1885–1893