Linear-Scaling Density Functional Theory with Tens of Thousands of Atoms: ONETEP

Nicholas D.M. Hine[1], **Peter D. Haynes**[1]

1. Blackett Laboratory and Thomas Young Centre, Imperial College London, Exhibition Road, London SW7 2AZ, U.K.

Abstract

We present recent improvements to the ONETEP code. ONETEP is an ab initio electronic structure package for total energy calculations within density-functional theory. Its main distinguishing features are true 'linear scaling', in that the total computational effort scales only linearly with system size, and 'plane-wave' accuracy, in that the convergence of the total energy is systematically improvable with increasing cutoffs. We present recent improvements to the parallel performance of the code, and thus in effect considerable increases in the scope and scale of feasible calculations with ONETEP, especially in solids. On parallel computers comprising large clusters of commodity servers, our recent improvements make calculations of tens of thousands of atoms in a solid feasible even for small numbers of cores (10-100). Efficient scaling with number of atoms is demonstrated up to 32,768 atoms on 64 cores, and efficient scaling with number of cores is demonstrated up to 512 cores for 32,768 atoms.

ONETEP Theory

Traditional Kohn-Sham DFT finds extended eigenstates $\psi_i(\mathbf{r})$ with eigenvalues ϵ_i to solve $\hat{H}$ for some effective potential $V[n](\mathbf{r})$:

$$\hat{H}\psi_i(\mathbf{r}) = \left[-\frac{\hbar}{2m}\nabla^2 + V[n](\mathbf{r})\right]\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}) \ , \qquad (1)$$

The density matrix $\rho(\mathbf{r},\mathbf{r}')$ can then be written either in terms of the eigenstates $\psi_i(\mathbf{r})$ and occupation numbers f_i as

$$\rho(\mathbf{r},\mathbf{r}') = \sum_i f_i\psi_i(\mathbf{r})\psi_i^*(\mathbf{r}') \ . \qquad (2)$$

or in terms of a set of *localised* nonorthogonal functions $\phi_\alpha(\mathbf{r})$

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta}\phi_\alpha(\mathbf{r})K^{\alpha\beta}\phi_\beta(\mathbf{r}') \ , \qquad (3)$$

where the matrix $K^{\alpha\beta}$, the density kernel, is a generalisation of occupation numbers to a nonorthogonal basis.

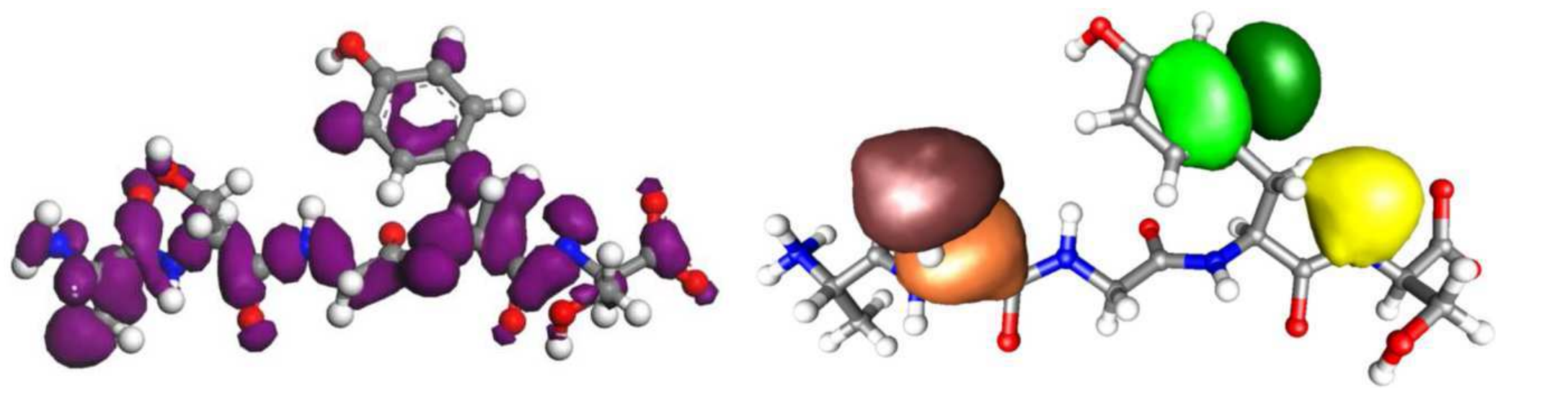

Figure 1: (left) An extended eigenstate for an oligopeptide molecule (right) Example localized NGWFs in the same molecule

Eigenstate-based approaches inevitably scale as $O(N^3)$ with the number of atoms N: the system has $O(N)$ eigenstates, each of size $O(N)$, and each needing to stay orthogonal to $O(N)$ others. Localised-orbital approaches, however, can scale as $O(N)$. In an insulator, the kernel $K^{\alpha\beta}$ can be truncated beyond some cutoff radius R_K, so the matrix is *sparse*. The overlap matrix $S_{\alpha\beta} = \langle\phi_\alpha|\phi_\beta\rangle$ is also sparse for localised $\phi_\alpha(\mathbf{r})$, as are elements of the Hamiltonian matrix $H_{\alpha\beta} = \langle\phi_\alpha|\hat{H}|\phi_\beta\rangle$. With $H_{\alpha\beta}$ and the density $n(\mathbf{r}) = \sum_{\alpha\beta}\phi_\alpha(\mathbf{r})K^{\alpha\beta}\phi_\beta(\mathbf{r})$ we can find the total energy E with $O(N)$ scaling by using

$$E[\{K^{\alpha\beta}\},\{\phi_\alpha\}] = \sum_{\alpha\beta}K^{\alpha\beta}H_{\beta\alpha} + E_{\mathrm{DC}}[n(\mathbf{r})] \ , \qquad (4)$$

and simulatenously minimising E with respect to the kernel and the coefficients describing the NGWFs, subject to the constraint that the density kernel remains idempotent and that its trace equals the number of electrons.

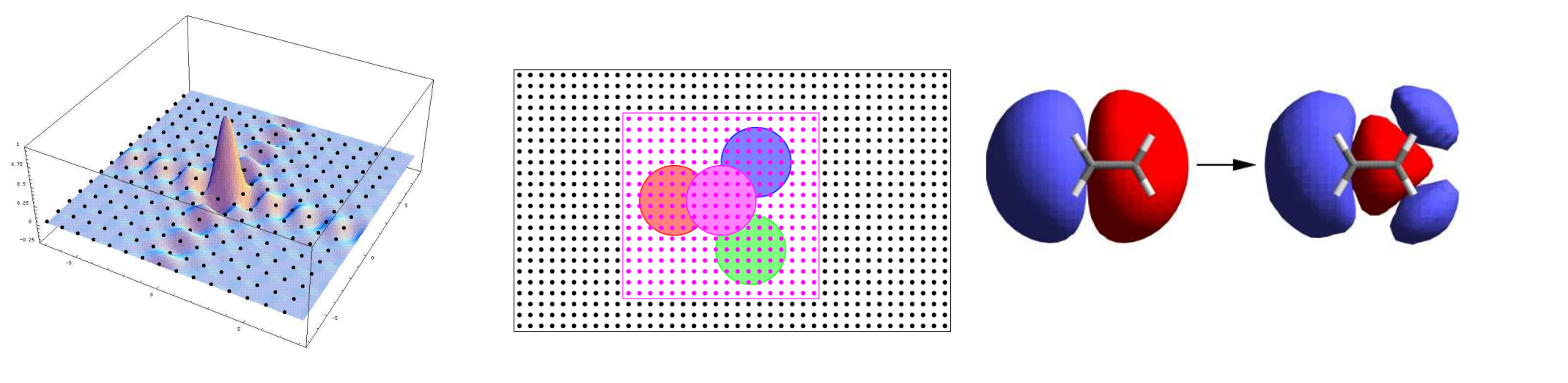

Figure 2: (left) A psinc function (middle) FFT box containing overlapping NGWFs (right) Example of NGWF optimisation of a *p*-orbital.

ONETEP combines $O(N)$ scaling with 'plane-wave' accuracy, in that the convergence of the total energy is systematically improvable by increasing cutoffs. The localised basis in ONETEP comprises 'Nonorthogonal Generalised Wannier Functions' (NGWFs) expressed in terms of a basis of periodic bandwidth-limited delta functions, or psinc functions, (see Fig 2) strictly localized to spherical regions of radius R_{ϕ_α}. These psinc functions, with coefficients $C_{i,\alpha}$, are centered on the grid points $\mathbf{r}_i$ of a regular grid specified by a plane-wave cutoff energy E_{cut}.

The minimisation of the energy occurs via nested loops: The outer loop minimises the energy with respect to the coefficients $C_{i,\alpha}$

$$E_{\mathrm{min}} = \min_{\{C_{i,\alpha}\}} L(\{C_{i,\alpha}\}) \ , \qquad (5)$$

while inner loop, performed at fixed $C_{i,\alpha}$, minimizes the energy with respect to the kernel elements $K^{\alpha\beta}$

$$L(\{C_{i,\alpha}\}) = \min_{\{K^{\alpha\beta}\}} E(\{K^{\alpha\beta}\};\{C_{i,\alpha}\}) \ . \qquad (6)$$

Parallel Optimisation

ONETEP was developed from the beginning as a parallel code and its efficient scaling and performance on isolated molecules, nanotubes and similar systems with a high degree of sparsity has been well-documented. Recent work has focused on improving performance in solids, where communications bottlenecks in large systems has previously limited the useful applicability of the code.

Matrix Algebra

One time-limiting component of ONETEP calculations is sparse matrix algebra, especially during kernel optimisation. As Fig 3 shows, the pattern of filling of the sparse matrices representing $S_{\alpha\beta}$, $H_{\alpha\beta}$, and $K^{\alpha\beta}$ can be highly structured, allowing considerable optimisation of the communication and computation patterns. Recent improvements include:

- Dense matrix algebra to replace sparse algebra in small systems or systems such as metals where kernel truncation is not possible. See Fig 4.
- Reordered, non-blocking asynchronous comms to allow different node-node pairings to take different lengths of time.
- Reduced total comms volume by communicating only blocks of multiplicands contributing to matrix products.
- Loop-unrolled block multiplication hardcoded for common block sizes (1,4,9).

Combined, these developments have dramatically improved both the speed and scaling (with system size and number of parallel processors) of matrix algebra.

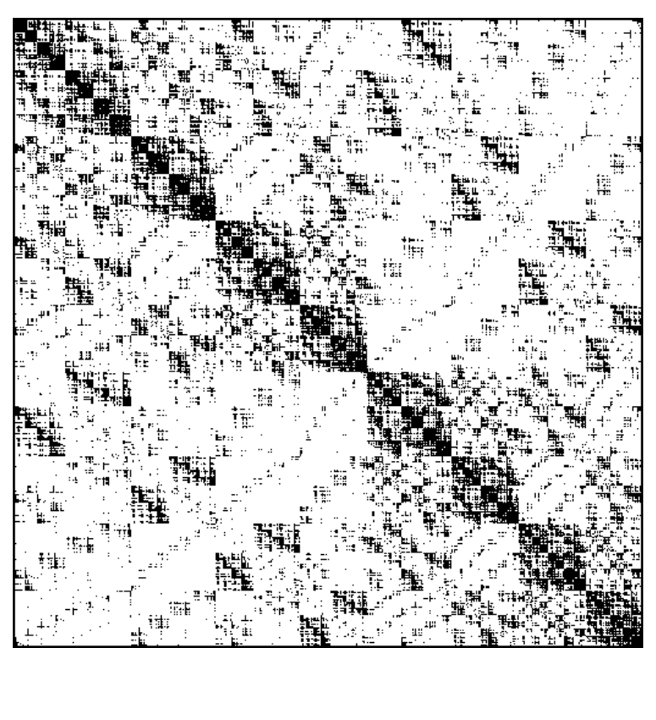

Figure 3: Sparsity pattern of $S_{\alpha\beta}$ for 512-atom fcc silicon.

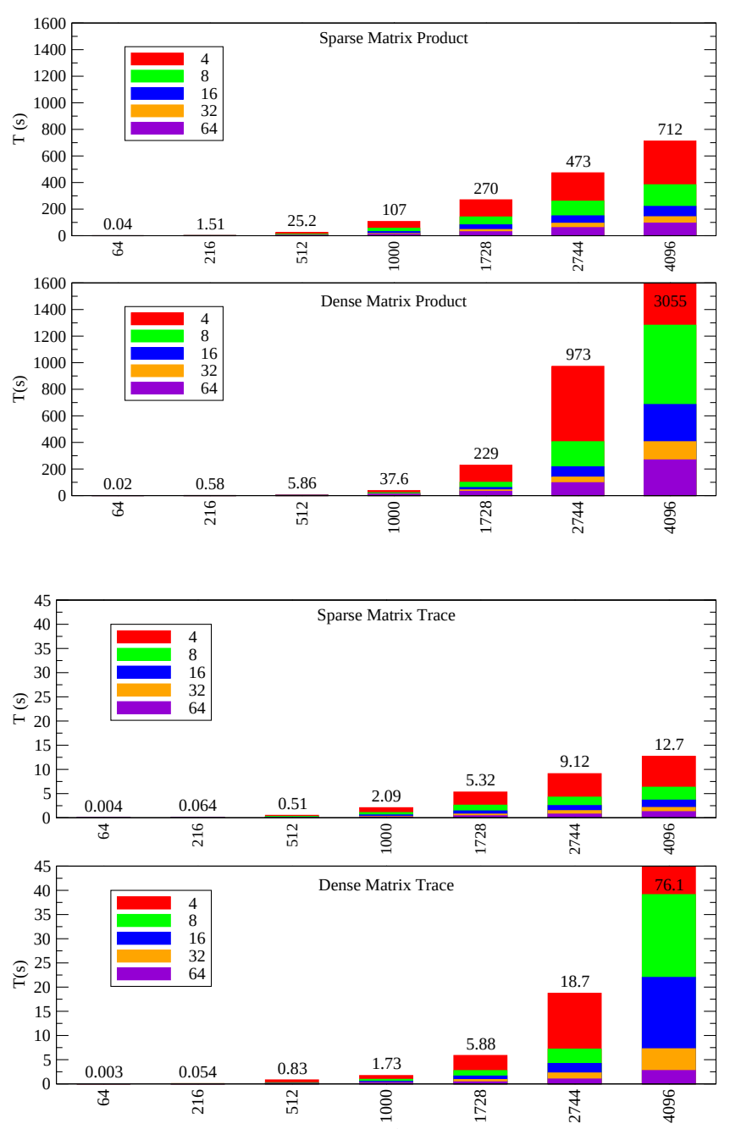

Figure 4: Timings for sparse and dense matrix products in fcc silicon with 4 NGWFs per atom, with supercells of 64 to 4096 atoms on 4 to 64 cores.

Row Sums

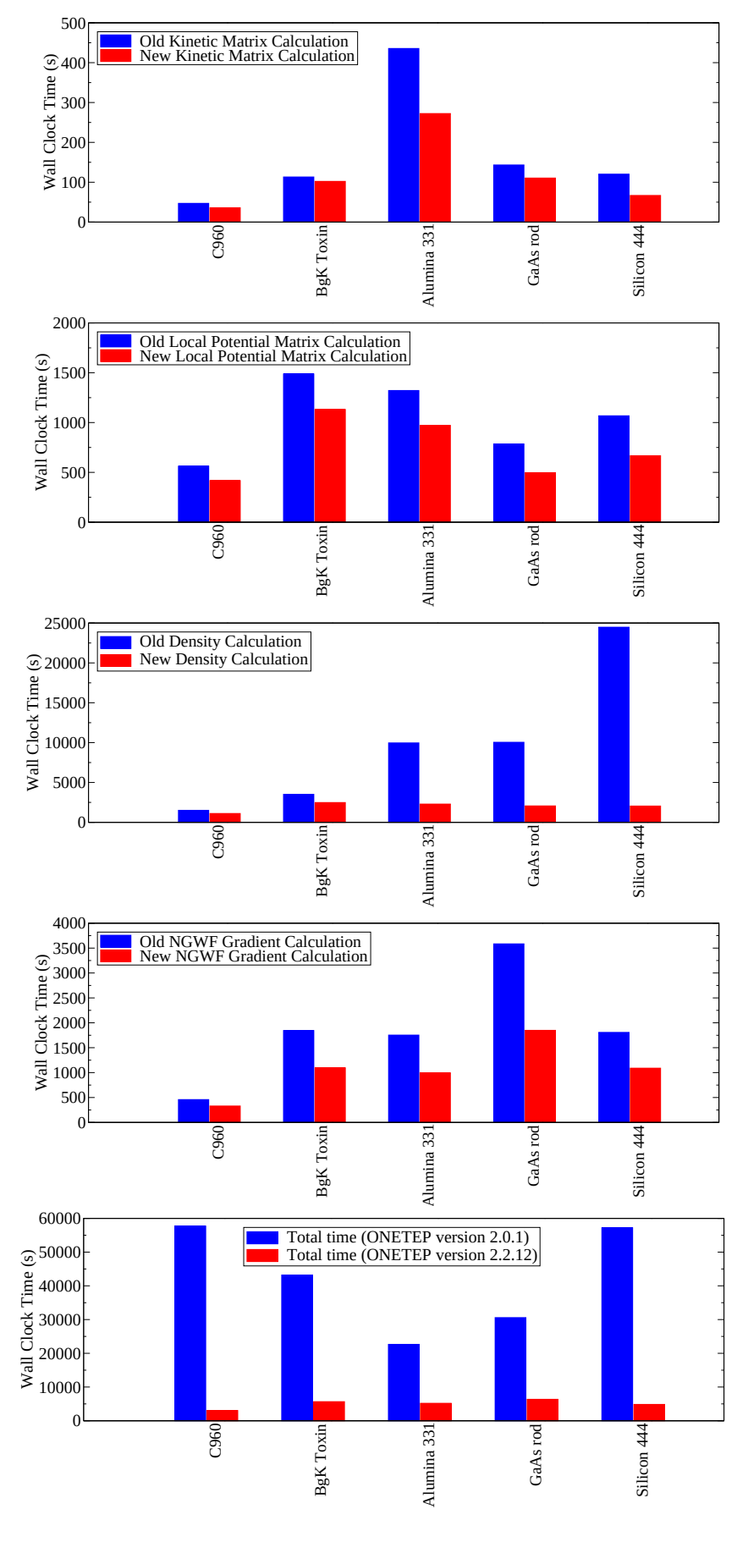

Figure 5: Timings for row sums operations on 16 nodes, for a range of systems (C Nanotube, Organic BgK toxin, Al_2O_3 Crystal, GaAs Nanorod, Si Crystal).

Another major contributor to the computational work is from the 'row sums' operations, for calculating all the contributions to a matrix or other quantity that involve a given $\phi_\alpha(\mathbf{r})$. Examples include kinetic and local potential matrices and energies, the electron density and the NGWF gradient:

$$\begin{array}{ll} \text{i)} & E_{\mathrm{kin}} = K^{\alpha\beta}\langle\phi_\alpha|\hat{T}|\phi_\beta\rangle \\ \text{ii)} & E_{\mathrm{loc}} = K^{\alpha\beta}\langle\phi_\alpha|V_{\mathrm{loc}}|\phi_\beta\rangle \ , \\ \text{iii)} & n(\mathbf{r}) = K^{\alpha\beta}\phi_\alpha(\mathbf{r})\phi_\beta(\mathbf{r}) \ , \\ \text{iv)} & \partial E/\partial\phi_\alpha(\mathbf{r}) = Q^{\alpha\beta}\phi_\beta(\mathbf{r}) + \ldots \end{array}$$

The sparsity pattern of the overlap matrix can be used to 'plan' in advance which pairs of NGWFs contribute to these expressions. Sharing this plan with all other nodes creates an efficient communication system, since each node is able to send NGWFs to other nodes exactly as they are needed.

Parallel Scaling

Combined with other parallel optimisations, these improvements have resulted in very considerable decreases in the total computation time:

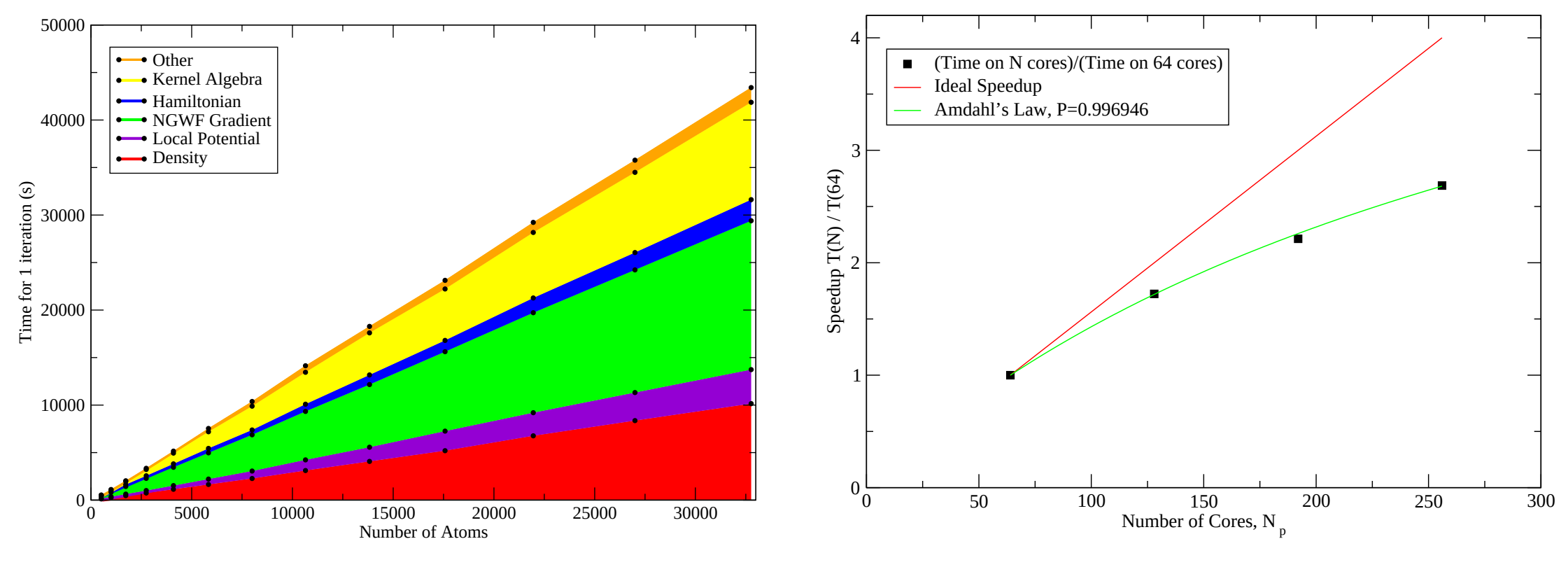

Figure 6: (left) Scaling with system size for fcc silicon — clear linear scaling of the total time for 1 iteration is observed up to 32768 atoms (right) Scaling with number of cores on which the calculation is run — efficient speedups are obtained up to at least 256 cores for 27000 atoms.

Recent Applications

GaAs Nanorods

'Self-assembly' is a promising route to constructing working nanotech devices. Wurtzite structure GaAs nanorods, which display spontaneous polarisation due to the polar bonding and lack of inversion symmetry, have been observed forming a variety of self-assembled structures. Linear-scaling DFT with ONETEP allows a window on the complex interplay between bonding and long-range electrostatic effects (which can be treated with cutoff coulomb interactions) required to model these systems.

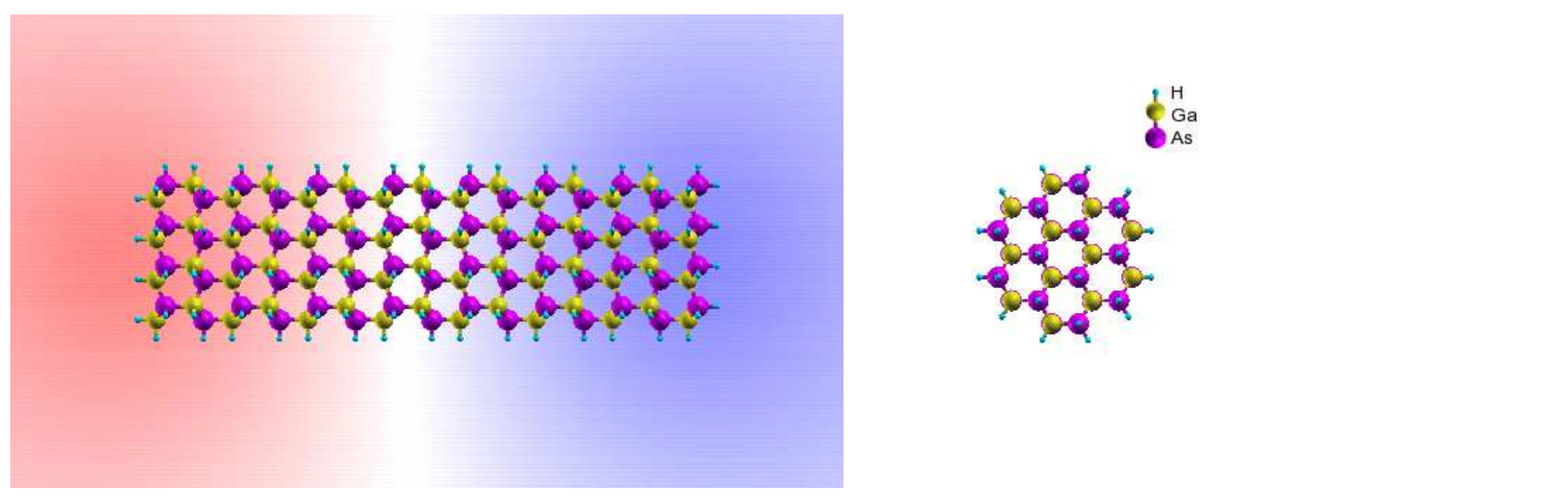

Figure 7: (left) Effective potential for an isolated H-terminated GaAs nanorod (564 atoms) on a plane 6Å behind the rod. End-to-end charge-separation results in the dipole field seen above. (right) Cross section of the rod.

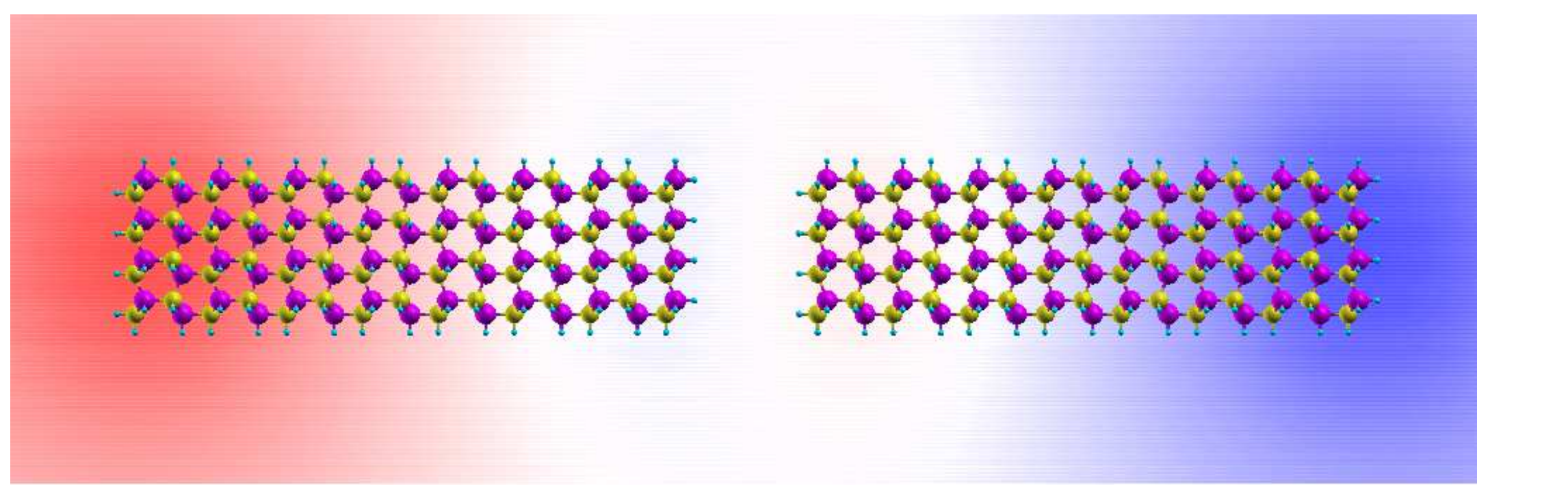

Figure 8: Two rods (1128 atoms) placed near each other to investigate binding energies.

Defect Formation Energies

Understanding defects and defect clusters in crystalline materials is a tough challenge for electronic structure methods due to the requirement of embedding (often large) localised systems in periodic hosts. As a very simple example, even the formation energy of the comparatively simple neutral vacancy in Germanium V_{Ge}^0 tests the robustness of DFT total energy methods, since the defect state in Ge is doubly-occupied but quadruply-degenerate. Therefore, while the bulk crystal is an insulator, there is a degeneracy at the Fermi level in an unrelaxed defect supercell. In small cells with a traditional DFT approach, this manifests itself as a band whose occupation varies as function of $\mathbf{k}$ and may require a fudging of occupation numbers to provide results converged with respect to supercell size.

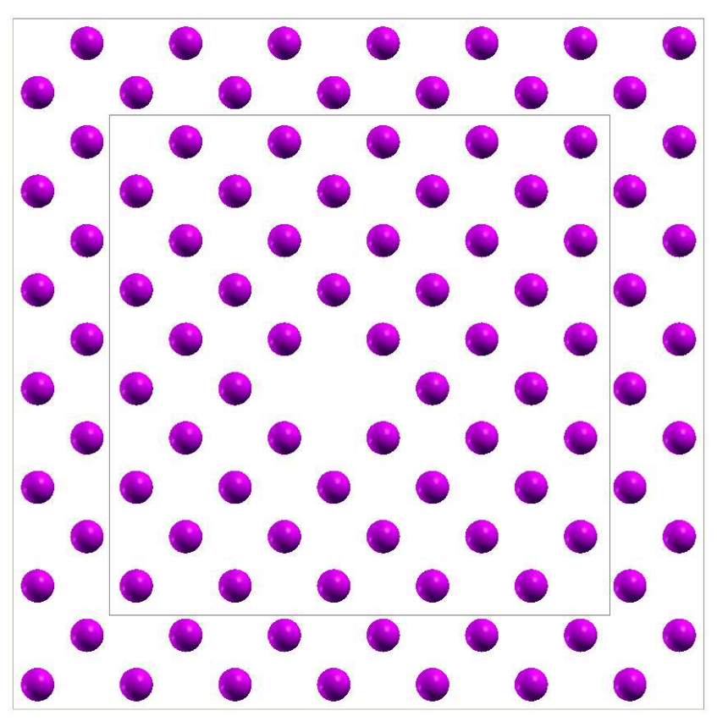

Figure 8: Cross section through an unrelaxed vacancy in Ge, showing (a) the largest supercell realistically feasible with a traditional plane-wave DFT approach on 64 cores: 1000 atoms (inner box) and (b) a typical considerably bigger cell (2744 atoms) that is quite feasible in a realistic time with ONETEP (outer box).

In ONETEP cells large enough to require only Γ-point calculations for accurate results can be used, giving results much less affected by finite-size errors. However, a localised degeneracy can still lead to problems with initial density kernel occupation numbers not summing exactly to N_e. There are two solutions to this — either the initial degeneracy can be broken by minor random adjustments to the initial $H_{\alpha\beta}$, or the initial $H_{\alpha\beta}$ can be explicitly diagonalised (a single $O(N^3)$ operation) to provide an perfectly idempotent kernel as a starting-point.

Conclusions

Improvements to the ONETEP code have lowered the 'crossover point', the number of atoms beyond which a linear scaling algorithm becomes requires less calculations than a comparable traditional DFT method. This was already ($\sim$ 100 atoms) for a non-dense system such a nanotube, but it has now been brought down to the regime of feasible calculations for solids as well. For example, in covalent semiconductors such as Si, Ge and III-V's, we estimate the crossover point relative to CASTEP to be around $600 - 1000$ atoms.

We have also demonstrated the efficiency and accuracy of the code to two important applications: self-assembly in GaAs nanorods, and defect formation energies in semiconductors.