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

Nicholas D.M. Hine¹, Arash A. Mostofi², Peter D. Haynes³

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

Abstract

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 $H_{\text{loc}}+\mu$, and $K^{\text{loc}}$ can be highly structured, allowing considerable optimisation of the communication and computation patterns. Recent improvements include:

- Hexagonal sparsity: division of matrix into node/segment and atom-atom ‘blocks’, dense storage used when filling exceeds threshold η.
- Efficient dense matrix algebra used on the level of segments or submatrices as appropriate.
- Reduced total volume sums by communicating only those blocks or segments of multithreads contributing to matrix product.

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

NGWF Operations

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 involves a given $\phi_i$ matrix. Examples include kinetic and local potential matrices and energies, the electron density and the NGWF gradients.

The sparsity pattern of the overlap matrix can be used to ‘plan in advance which pairs of NGWFs contribute to those expressions. This plan allows each node to request the NGWFs it requires from other nodes in advance of when they are needed, minimising communications latency.

Figure 8: Electrastatic Potential of a pair of GaAs Nanorods

Recent Applications

GaAs / ZnO Nanorods

The properties of III-V and II-VI semiconductor nanorods in the Wurtzite structure, including polar bonding and lack of inversion symmetry, are of interest to a variety of fields, including spintronics and photovoltaics. Linear-scaling DFT with ONETEP enables simulation of realistic systems of large systems of atoms, elucidating the complex interplay between bonding and long-range electrostatics effects required to model these systems.

Figure 9: Formation energies of Aluminium and Oxygen vacancies in $\text{Al}_2\text{O}_3$ as a function of system size, showing slow convergence and requiring large system sizes for accurate results.

Strongly Correlated Systems

We have recently completed an implementation of the DFT+$U$ method within ONETEP, which allows treatment of systems where so-called “strong correlation” effects are important. This is the case wherever there are strongly localized states in tight-binding $d$- or $f$-subshells which are partially filled.

Figure 8: Scaling of DFT+$U$ calculations on NIO nanoclusters. The method meets minimal overhead compared to a standard calculation.

Protein-Ligand Binding Energies

ONETEP is also ideal for quantum-mechanical investigations of biological systems. Recent investigations have included binding of protein fragments with nanoparticles.

Figure 8: 3D backbone binding with peptides from a protein involved in formation of amyloid fibrils. (C) Molecular Pilot 2016 (molecularpilot.co.uk)

Abstract

ONETEP Theory

Traditional Kohn-Sham DFT finds single-electron states $\phi_i$ with energies $e_i$ to solve the Schrödinger Equation for an effective potential $V(r)$:

$$H_{\text{KSh}}(\phi_i) = -\frac{1}{2}\nabla^2 \phi_i - V(\phi_i) - e_i \phi_i \tag{1}$$

In linear-scaling DFT, we use the density matrix $\rho(r,r')$ rather than the eigenstates. In terms of $\rho(r,r')$ and occupation numbers $f_i$ this is:

$$\rho(r,r') = \sum_i f_i \phi_i(r) \phi_i(r') \tag{2}$$

or in terms of a set of localized nonorthogonal functions $\phi_i(r)$:

$$\rho(r,r') = \sum_{i,j} \phi_i(r) \phi_j(r') \tag{3}$$

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

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

Approaches that use eigenstates inevitably scale as $N^3$ (left) scaling by using $N_\text{SIZE}$ with the number of atoms $N$ the system has $O(N)$ eigenstates, each size $O(N^2)$, and each needing to stay orthogonal to $O(N)$ others. Localised-orbital approaches, however, can scale as $O(N)$. In an insulator, the kernel $K^{\text{loc}}$ can be truncated beyond some cutoff radius $R_{\text{gs}}$, so the matrix is sparse. The overlap matrix $S_{\alpha\beta} = \langle \phi_\alpha | \phi_\beta \rangle$ is also sparse for localized $\phi_i$ as are elements of the Hamiltonian matrix $H_{\text{loc}} = \langle \phi_\alpha | H | \phi_\beta \rangle$. With $H_{\text{loc}}$ and $K^{\text{loc}}$ we can find the total energy $E$ with $O(N)$ scaling by using:

$$E[\{\phi_i\}] = \sum_{i,j} H_{\text{loc}} \phi_i \phi_j + \frac{1}{2} \sum_i K^{\text{loc}} \phi_i \phi_i \tag{4}$$

and simultaneously minimising $E$ with respect to both $\phi_i$ and the coefficients denoting the NGWFs, subject to the constraint that the density kernel remains idempotent and that $\int \rho(r) dr$ equals the number of electrons.

Figure 2: (left) A plane function (middle) FFT box containing overlapping NGWFs (right) Example of NGWF optimisation of a $p$-related ONETEP combines $O(N)$ scaling with $\text{plane-wave}$ accuracy, in that the convergence of the total energy is systematically improvable by increasing cutoffs. The localized basis in ONETEP combines Nonorthogonal Generalised Wannier Functions (NGWFs) expressed in terms of a basis of periodic band-limited delta functions, or plane waves, (see Fig 2) strictly localized to spherical regions of radius $R_{\text{gs}}$. The plane-wave functions, with coefficients $C_{\alpha\beta}$ are centered on the grid points $r_i$ of a regular grid specified by the plane-wave cutoff $R_{\text{gs}}$. The minimisation of the energy occurs via nested loops: the outer loop minimizes the energy with respect to the coefficients $C_{\alpha\beta}$.

Parallel Optimisation

Figure 3: Sparsity patterns of $(K^{\text{loc}})_{\alpha\beta}$ for ~4000-atom systems $C$ Nanotube, DNA Strand, GaAs Nanorod, Crystalline Silicon

NGWF Operations

Another major contribution to computational work is from the ‘row sums’ operations, for calculating all the contributions to a matrix or other quantity involving a given $\phi_i$ matrix. Examples include kinetic and local potential matrices and energies, the electron density and the NGWF gradient.

The sparsity pattern of the overlap matrix can be used to ‘plan in advance which pairs of NGWFs contribute to those expressions. This plan allows each node to request the NGWFs it requires from other nodes in advance of when they are needed, minimising communications latency.

Figure 6: Timings for row sums operations on 16: melange of 3000-atom range of systems C Nanotube, Organic Hyg- toxin, Al2O3 Crystal, GaAs Nanorod, Si Crystal

Figure 7: (left) Scaling with system size for DNA strand, clear linear scaling of the total time is observed (right). Scaling with number of cores on which the calculation is run — efficient speedups are obtained up to at least 256 cores.

Parallel Scaling

Figure 8: Scaling of DFT+$U$ calculations on NIO nanoclusters. The method meets minimal overhead compared to a standard calculation.

Protein-Ligand Binding Energies

ONETEP is also ideal for quantum-mechanical investigations of biological systems. Recent investigations have included binding of protein fragments with nanoparticles.