A tensor based approach for accelerating exact exchange computations

Vishal Subramanian[1], Sambit Das[2], Vikram Gavini[1,2]

[1]Department of Materials Science and Engineering; [2]Department of Mechanical Engineering, University of Michigan, Ann Arbor, USA

Overview

Density Functional Theory (DFT) is a widely used electronic structure theory to predict material properties. In DFT, the many-electron problem is mapped onto an effective single-electron problem. This reduction is achieved by introducing the exchange-correlation (XC) potential, which encapsulates the quantum many-body effects. However, the exact nature of the XC potential is unknown and is approximated in practice. Different approximations to XC potentials with increasing accuracy have been proposed, ranging from Local Density Approximation (LDA - depends on the electron density), Generalised Density Approximation (GGA - depends on the electron density and its gradient) to hybrid XC potentials (depends on electron density, gradient, and the orbitals). The higher accuracy of hybrid XC functionals is attributed to the inclusion of the exact exchange. However, this inclusion of exact exchange results in a tremendous increase in the calculation cost.

Challenge: Develop an accurate, robust, and scalable algorithm to perform DFT simulations involving hybrid XC functionals that can handle large system sizes.

Methodology and Goals: Develop an algorithm based on systematically convergent Tucker-Tensors to accelerate the computation of the exact exchange

Progress: We have implemented a Tucker-Tensor algorithm to accelerate the exact exchange computations in DFT-FE. DFT-FE is a massively parallel, open-sourced C++ code that uses finite element discretization to solve the DFT equations. The algorithm's accuracy was validated by comparing it with the Quantum Espresso (QE) results - a widely used plane waves code. We have exploited various HPC strategies and devised innovative communication strategies to ensure the efficiency and scalability of the algorithm. We demonstrate the efficiency of the code by comparing the wall times with QE. Further, a strong scaling study was performed to exhibit the scalability of the code.

Real-space formulation of GKS-DFT

- **Mathematical formulation** The problem of finding ground-state properties under the Generalised Kohn-Sham Density Functional theory (GKS-DFT) is equivalent to minimizing the following energy functional:

$$E(\boldsymbol{\Psi},\boldsymbol{R}) = T_s(\boldsymbol{\Psi}) + \alpha E_{xx}[\boldsymbol{\Psi}] + E_{xc,sl}(\rho) + J(\rho,\boldsymbol{R}) \quad \text{subject to} \quad \int \psi_i^*(\mathbf{x})\psi_j(\mathbf{x})\,d\mathbf{x} = \delta_{ij}\,.$$

The electron-density ρ and the kinetic energy $T_s(\boldsymbol{\Psi})$ are given by

$$\rho(\mathbf{x}) = \sum_i f_i|\psi_i(\mathbf{x})|^2\,, \qquad T_s(\boldsymbol{\Psi}) = \sum_i \int f_i\,\psi_i^*(\mathbf{x})\left(-\frac{1}{2}\nabla^2\right)\psi_i(\mathbf{x})\,d\mathbf{x}\,.$$

The exact exchange (E_{xx}) and the semi-local exchange-correlation functional are given by

$$E_{xx}[\boldsymbol{\Psi}] = -\frac{1}{2}\sum_i\sum_j f_i f_j \frac{\psi_i(\boldsymbol{r})\psi_j(\boldsymbol{r})\psi_j(\boldsymbol{r}')\psi_i(\boldsymbol{r}')}{|\boldsymbol{r}'-\boldsymbol{r}|}d\boldsymbol{r}'d\boldsymbol{r}\,, \qquad E_{xc,sl}(\rho) = \int \epsilon_{xc,sl}(\rho(\mathbf{x}))\rho(\mathbf{x})\,d\mathbf{x} \quad .$$

Electrostatic interactions are computed using the following local variational form

$$J(\rho,\boldsymbol{R}) = -\min_{\phi}\left\{\frac{1}{8\pi}\int|\nabla\phi(\mathbf{x},\boldsymbol{R})|^2\,d\mathbf{x} - \int(\rho(\mathbf{x})+b(\mathbf{x},\boldsymbol{R}))\phi(\mathbf{x},\boldsymbol{R})\,d\mathbf{x}\right\}\,.$$

Euler-Lagrange equation of the above Generalised Kohn-Sham energy functional is the following nonlinear eigenvalue problem which has to be solved for $N(>= N_e/2)$ smallest eigenvalues and its corresponding eigenfunctions. (N_e denotes the number of electrons.)

$$\left(-\frac{1}{2}\nabla^2 + V_{\text{eff}}(\boldsymbol{\Psi},\boldsymbol{R})\right)\psi_i = \epsilon_i\psi_i \quad \text{where} \quad V_{\text{eff}}(\boldsymbol{\Psi},\boldsymbol{R}) = \alpha V_F + \frac{\delta E_{xc,sl}}{\delta\rho} + \frac{\delta J}{\delta\rho}\,.$$

The action of the Fock operator on an arbitrary orbital is given by,

$$V_F[\boldsymbol{\Psi}]\phi = -\sum_j f_j \int \frac{\psi_j(\boldsymbol{r})\psi_j(\boldsymbol{r}')\phi(\boldsymbol{r}')}{|\boldsymbol{r}'-\boldsymbol{r}|}d\boldsymbol{r}'$$

- **Nature of Fock operator**
  - The convolution integral $\int \frac{\psi_j(\boldsymbol{r}')\phi(\boldsymbol{r}')}{|\boldsymbol{r}'-\boldsymbol{r}|}d\boldsymbol{r}'$ can be computed by solving a Poisson equation.
  - The action of Fock operator on one orbital requires $O(N_e)$ Poisson solves.
  - The action of Fock operator on N_e orbitals are required. Hence $O({N_e}^2)$ Poisson equations have to be solved.

Previous attempts

- **Reduce the number of times the action of Fock operator is needed** Adaptively Compressed Exchange (ACE) operator.
  - Constructs a low rank approximation that is exact in the space spanned by occupied orbitals.
  - **Drawback** The action of Fock operator still forms the bottleneck.
- **Accelerate the computation of action of Fock operator** Linear Scaling approaches.
  - Exploits locality of the orbitals.
  - **Drawback** Can not be generically applied to metallic systems.
  - **Drawback** Might require the explicit construction of the Fock operator (which is very costly).

Algorithmic Implementation : ideas and details

- **Tucker-Tensor decomposition of Functions** Functions can be approximated as a sum of rank-1 matrices.

$$f(\boldsymbol{r}') \approx \sum_a^{R_x}\sum_b^{R_y}\sum_c^{R_z} g_{abc}\,A_a(x')\,B_b(y')\,C_c(z')$$

  - g is called core tensor and the side matrices (A,B,C) depend on only one coordinate.
  - The number of terms R_x, R_y, R_z is called the rank of the decomposition in each direction.
  - By increasing the rank the error in the approximation can be systematically reduced.
  - The error in the approximation reduces exponentially with the rank of decomposition.
  - A rank of $[20, 30]$ is sufficient to achieve chemical accuracy. This is much less than the size of the mesh.

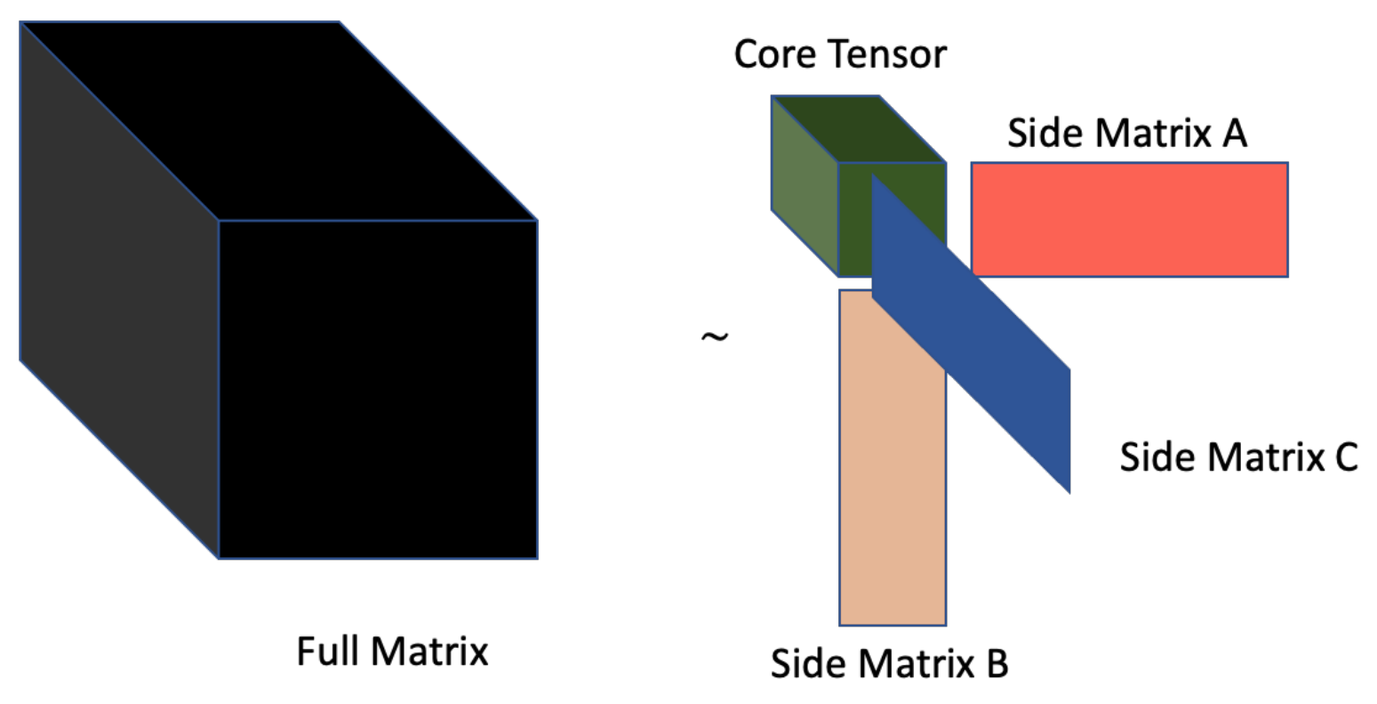


Figure: **Schematic of Tucker-Tensor decomposition**

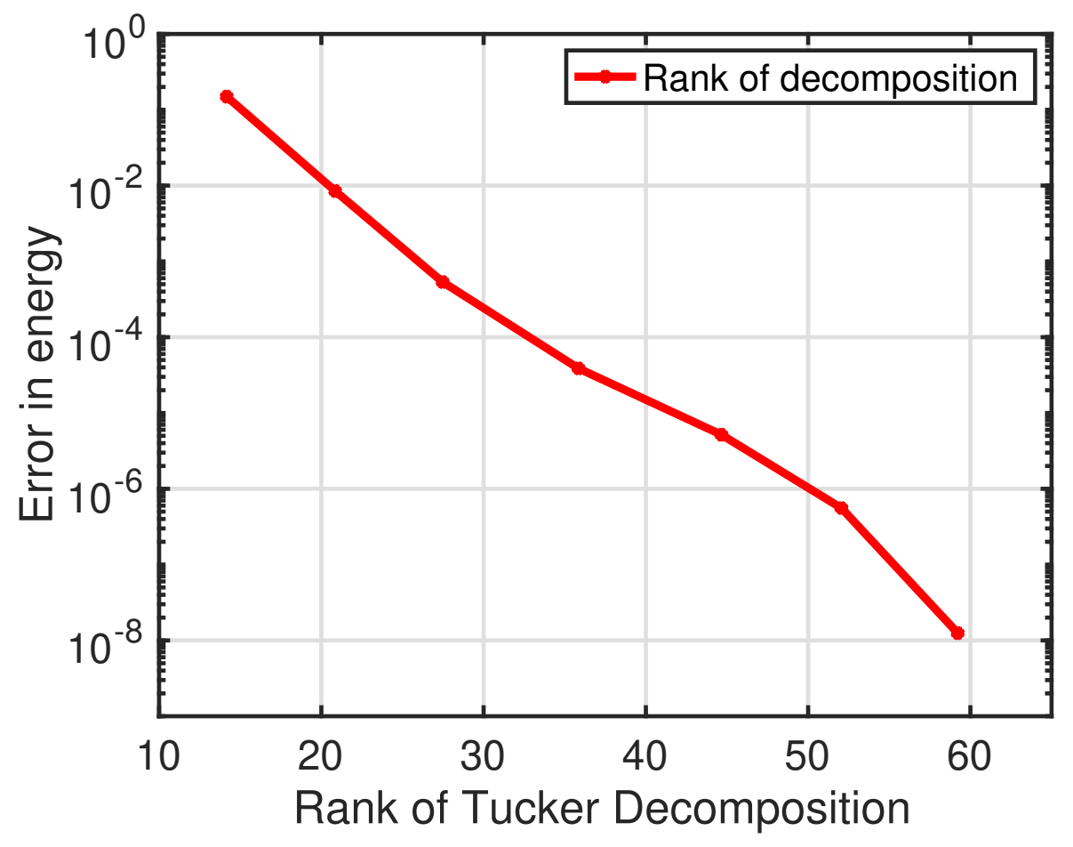


Figure: **Dependence of the exchange energy on the rank of decomposition**

- **Gaussian approximation for 1/r kernel**

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = \sum_t^T \alpha_t\; e^{-\beta_t(x-x')^2}\; e^{-\beta_t(y-y')^2}\; e^{-\beta_t(z-z')^2}$$

  - Point-wise error is bounded in a region [a,b] which can be decided based on the mesh and domain sizes.
  - The number of terms (T) can be increased to reduce the error in the convolution.
- **Accelerating the convolution integral** 3-D convolution is converted into 3 1-D integrals!!!!

$$\int\frac{f(\mathbf{r}')}{|\mathbf{r}'-\mathbf{r}|}d\mathbf{r}' = \int\sum_{abct} g_{abc}A_a(x')B_b(y')C_c(z')\;\alpha_t e^{-\beta_t(x-x')^2}e^{-\beta_t(y-y')^2}e^{-\beta_t(z-z')^2}dx'dy'dz'$$

$$= \sum_{abct} g_{abc}\,\alpha_t \int A_a(x')\; e^{-\beta_t(x-x')^2}dx' \int B_b(y')e^{-\beta_t(y-y')^2}dy' \int C_c(z')\; e^{-\beta_t(z-z')^2}dz'$$

- **MPI parallelisation strategies**
  - A blocked approach is pursued where in a bunch of operators and input orbitals are assigned to a processor. The processor computes the convolution for all the pairs of operator-input pairs assigned to it.
  - Such an approach allows extreme task parallelisation while minimising communication.
  - A round robin approach is pursued to achieve memory parallelisation over the operators.

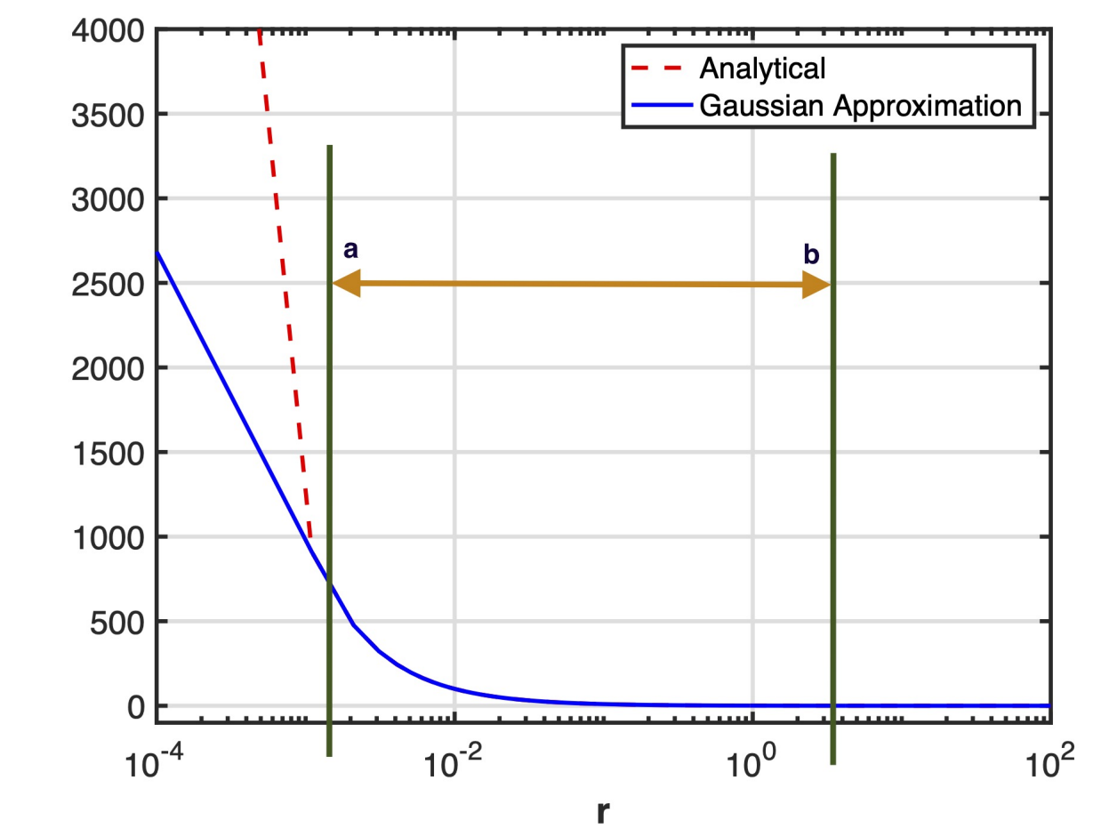


Figure: **Gaussian approximation for the 1/r kernel**

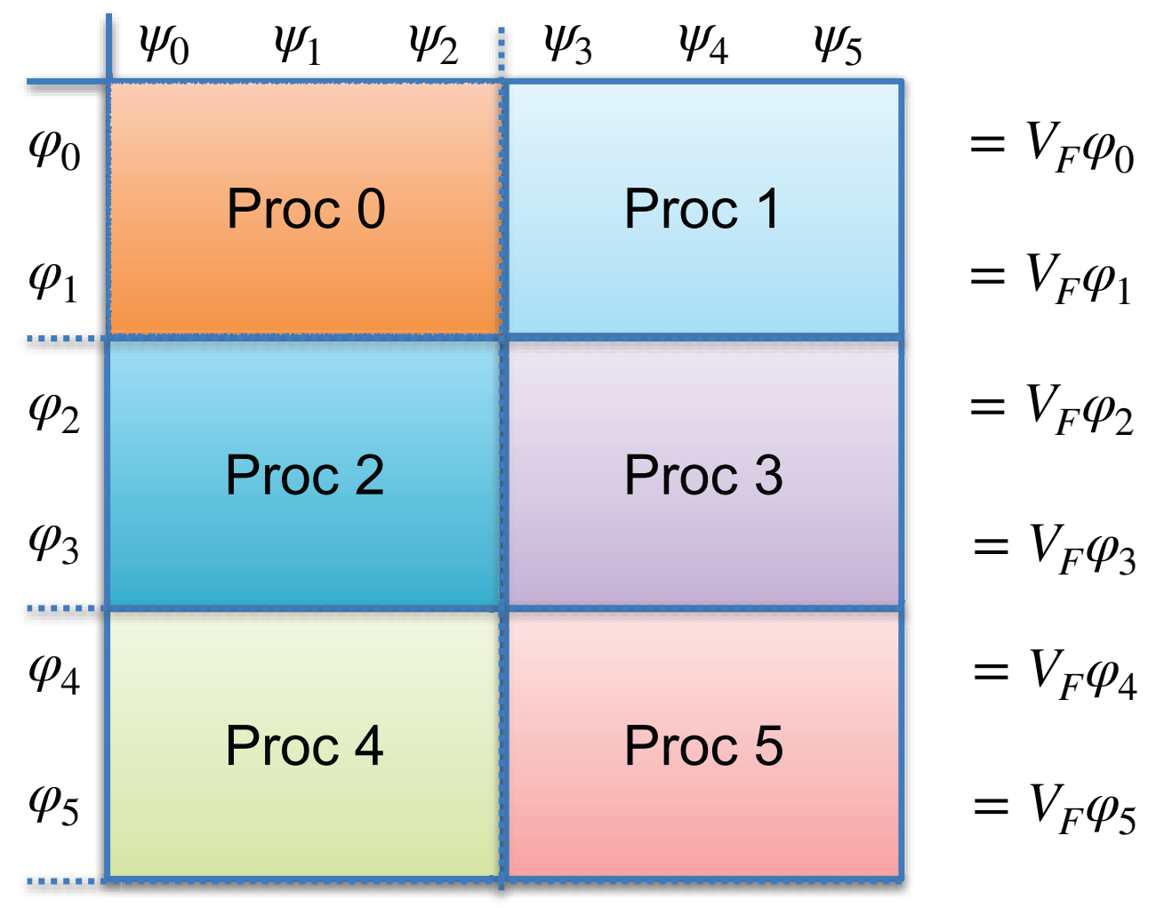
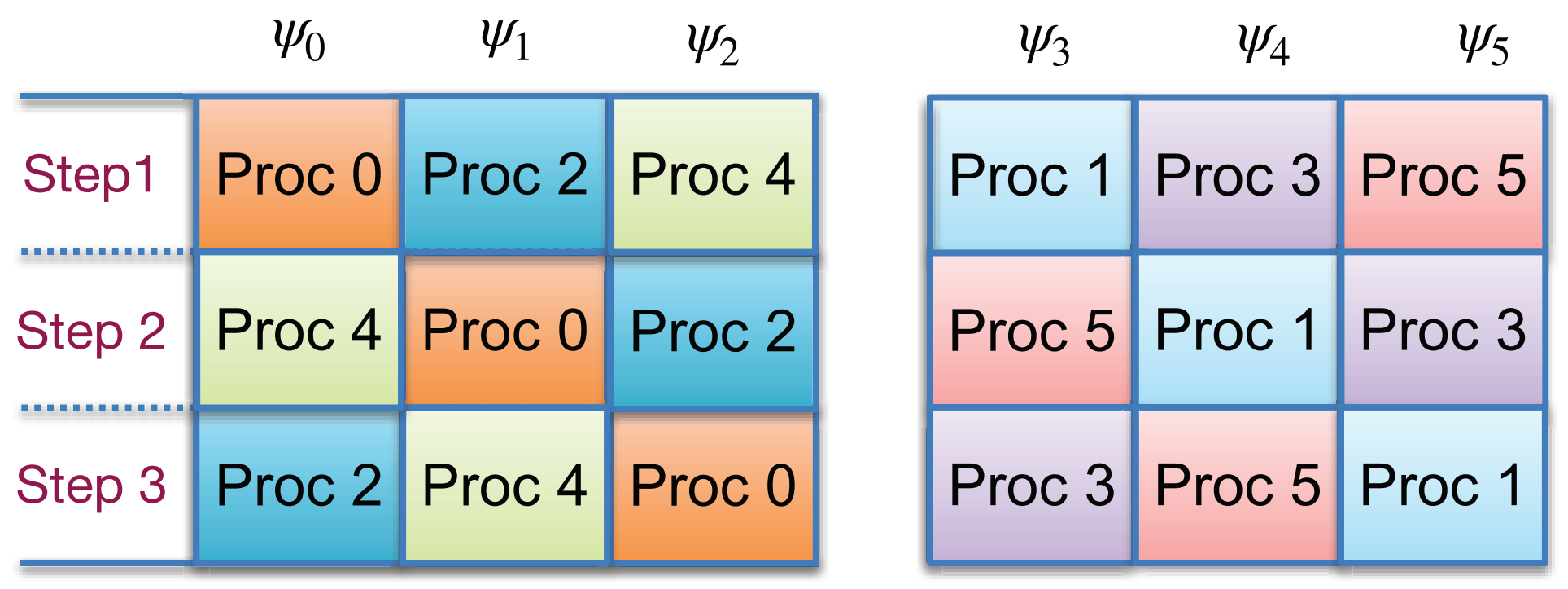


Figure: **Schematic of MPI parallelisation**



	ψ_0	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
Step1	Proc 0	Proc 2	Proc 4	Proc 1	Proc 3	Proc 5
Step 2	Proc 4	Proc 0	Proc 2	Proc 5	Proc 1	Proc 3
Step 3	Proc 2	Proc 4	Proc 0	Proc 3	Proc 5	Proc 1

Figure: **Schematic of round-robin algorithm for memory parallelisation**

Acknowledgements

I would like to thank Dr. Bikash Kanungo for the discussions and the financial support from TRI.

Accuracy benchmark

Can accurately compute the ground state energies by achieving the target accuracy of 1E-4 Ha/atom

- Implemented PBE0 hybrid functional in DFT-FE, in which the exact exchange energy is computed using Tucker -Tensor algorithm.
- Quantum Espresso computed the PBE0 ground state by solving the Poisson equations in exact exchange using plane waves.

Table: **Comparison of ground state energies**

System (No. of electrons N_e)	**Q-Espresso**	**DFT-FE**	Error (in Ha/atom)
Pt-Au dimer (N_e = 37)	-258.3823	-258.382	**1.72E-04**
Benzamide (N_e = 46)	-69.9225	-69.9213	**7.55E-05**
Pt 19 atom cluster (N_e = 342)	-2303.1149	-2303.1163	**7.37E-05**
Pt 38 atoms cluster (N_e = 684)	-4606.2690	-4606.2655	**9.22E-05**

Performance Benchmark

- The speed ups obtained increases with increasing system size and this observation is consistent with our complexity analysis.
- Obtained a **11x** speed up over Quantum Espresso for a single updateFock() for a large TiO2 system.

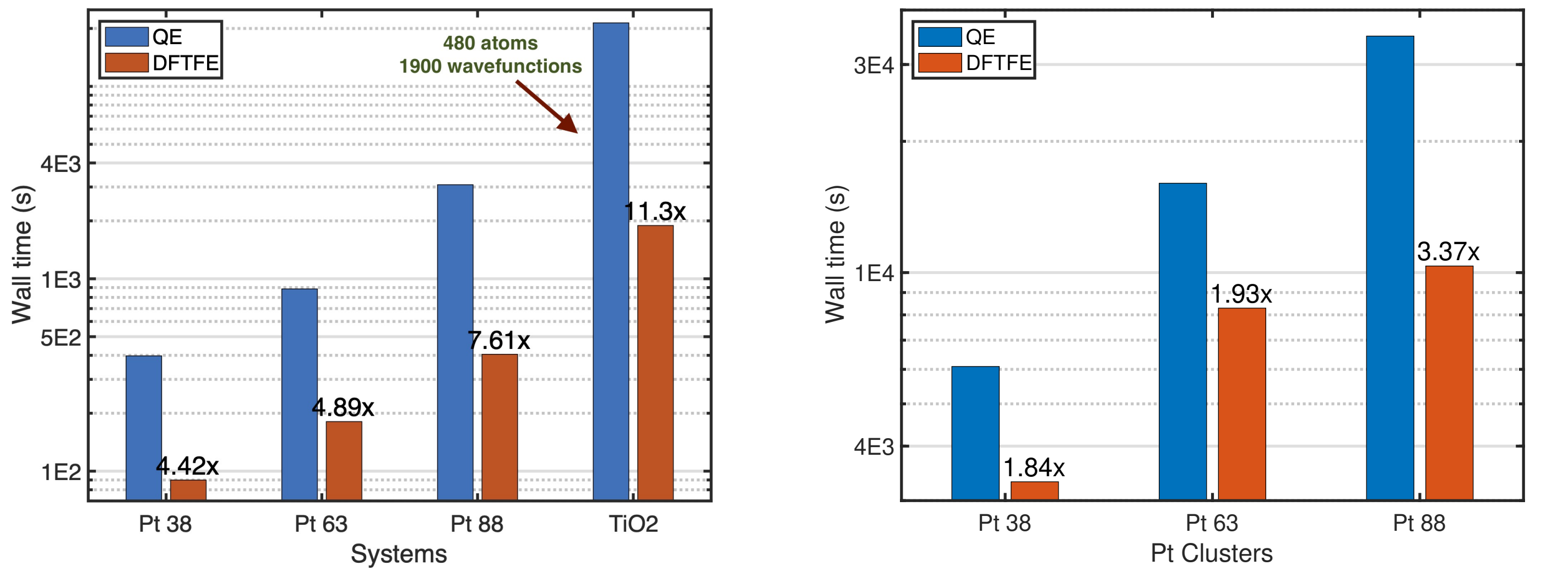


Figure: **One updateFock()**

Figure: **Total ground state**

Figure: **Comparison of wall time**

- Exhibits good scaling even in the extreme scaling regimes.
- The relative fraction of communication remains almost constant with scaling establishing the scalability of the algorithm.

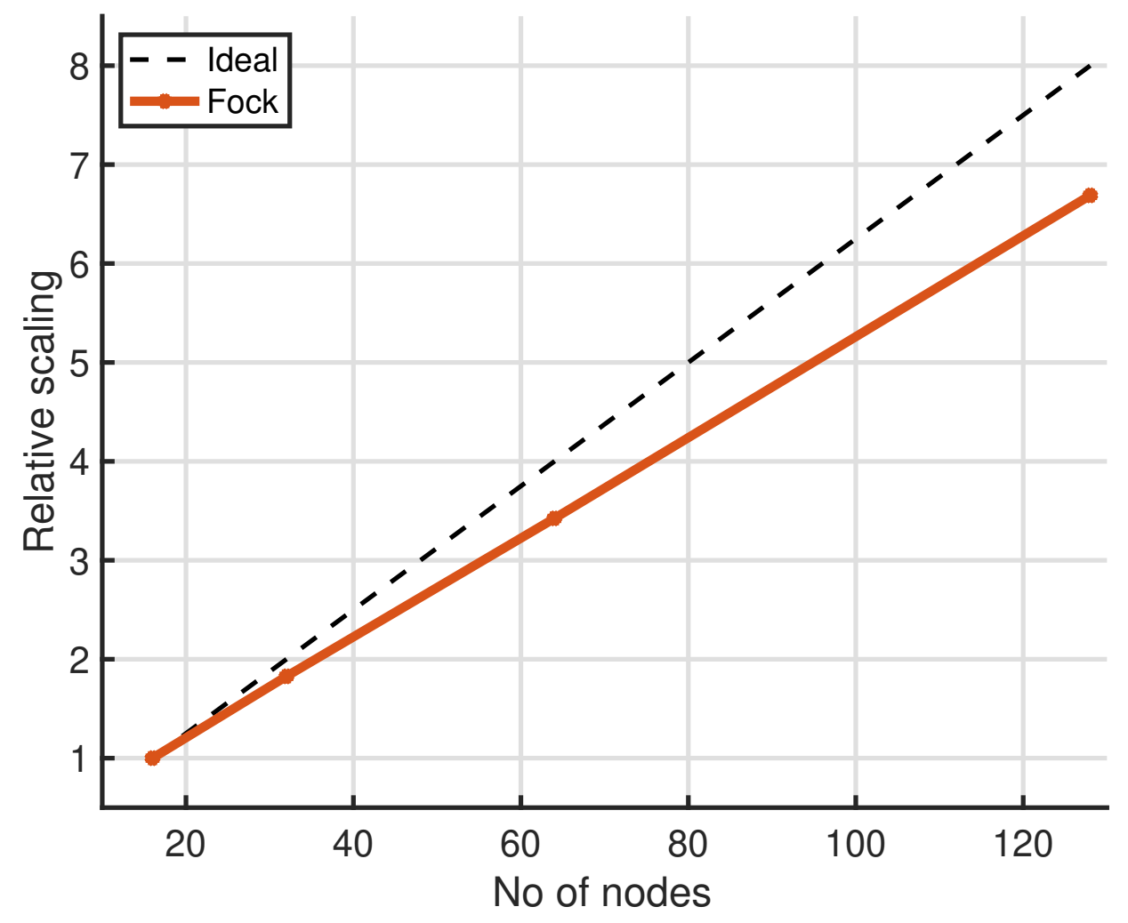


Figure: **Strong scaling of updateFock()**

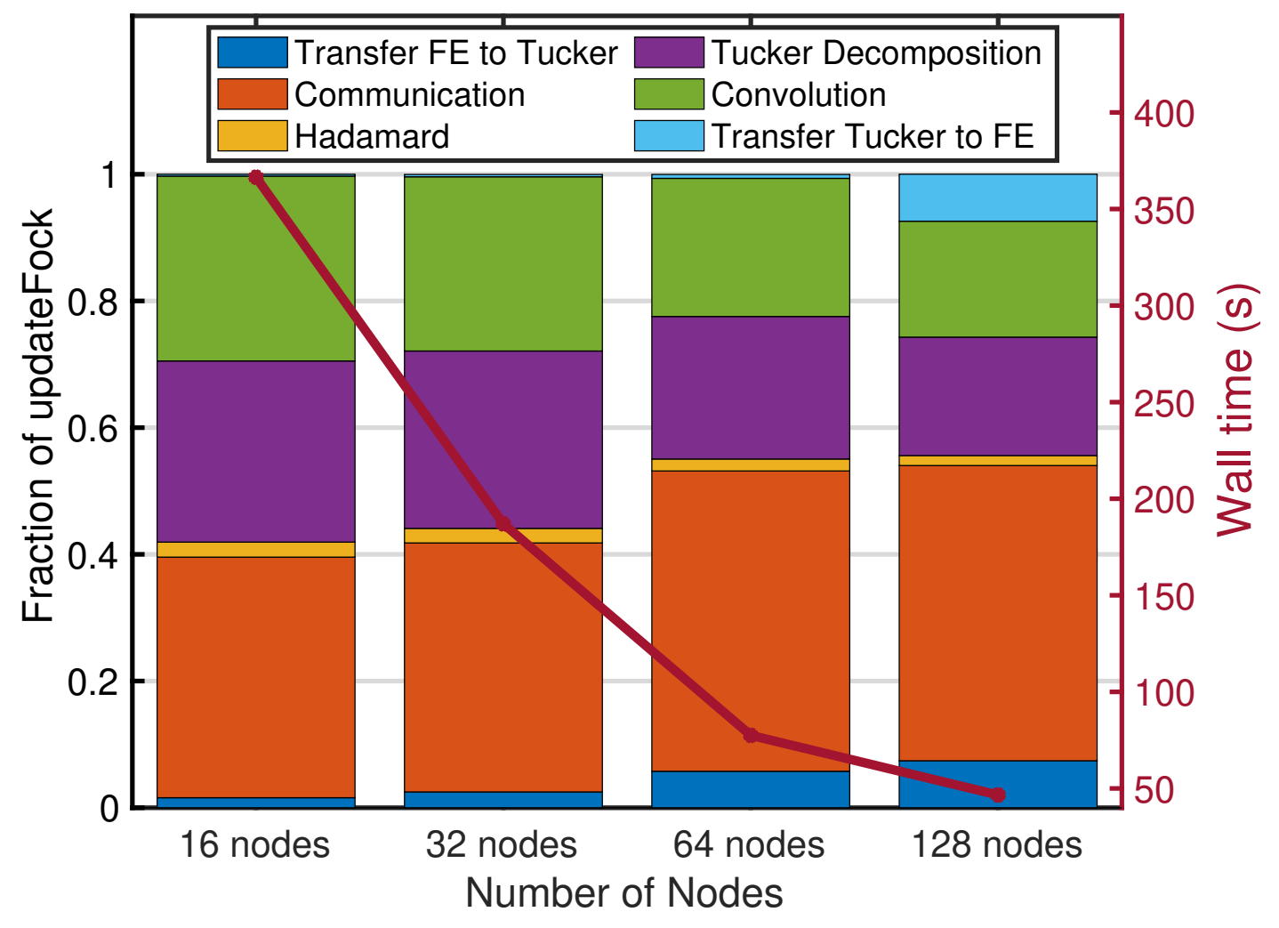


Figure: **Timings breakdown of different parts of the algorithm**

Ongoing/Future Work

1 GPU porting of the algorithm.
2 Extending to periodic and spin polarized systems.
3 Developing mixing strategies to improve the rate of convergence.

References

1 P. Motamarri, et al., *Comput. Phys. Commun.*, **246**, 106853, 2020.
2 Ballard, Grey, et al., *ACM Transactions on Mathematical Software*, 2020.
3 DeVore, et al., *Springer Berlin Heidelberg*, 2009.
4 Khoromskij, Boris N. *Chemometrics and Intelligent Laboratory Systems*, 2012.
5 Lin Lin. *Journal of chemical theory and computation*, 2016.
6 DFT-FE Open-source repo: `https://github.com/dftfeDevelopers/dftfe`