Enabling Large-Scale GPU-Accelerated Correlated Quantum Chemistry Calculations: the Divide-Expand-Consolidate Scheme

Dmytro Bykov[1], Ashleigh Barnes[1], Dmitry Lyakh[1]

The DEC scheme can be applied to evaluate molecular energy and properties in a linear-scaling and embarrassingly parallel manner using a set of local Hartree-Fock molecular orbitals. The essence of the method lies in the fact that all manipulations with the intermediate four-dimensional quantities are carried out independently within small local orbital fragment spaces. The sizes of the orbital fragment spaces are determined in a black-box manner to ensure that the error in the DEC implementation is proportional to a single input threshold, denoted as the fragment optimization threshold (FOT). The scaling behavior, performance and benchmark studies as well as series of showcase calculations prove the DEC method to be a highly effective tool for approaching large molecular systems.

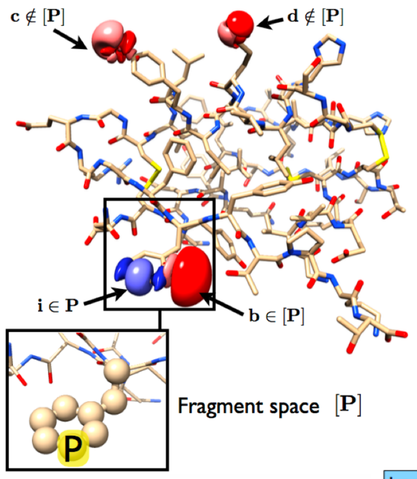
The Approximation

The correlation energy is partitioned as a sum of fragment and pair fragment energies. [P] denotes a set of orbitals assigned to atomic site P:

$$E_{corr} = \sum_P E_P + \sum_{P>Q} \Delta E_{PQ}$$

$$E_P = \sum_{ij \in P} \sum_{ab \in [P]} E_{ij}^{ab}$$

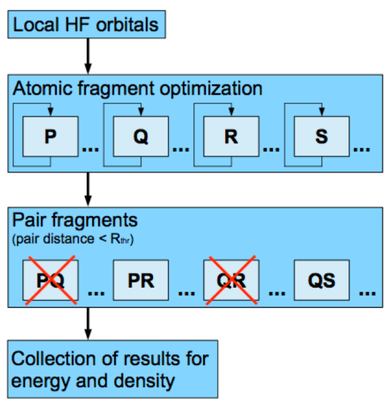
$$\Delta E_{PQ} = \sum_{i \in P} \sum_{j \in Q} \sum_{ab \in [P] \cup [Q]} E_{ij}^{ab} + P \leftrightarrow Q$$



Examples of localized orbitals – blue for occupied and red for virtual orbitals. For a given fragment P, the local amplitude equations are solved using orbitals assigned to atoms in the [P] space. Subsequently, fragment energy is evaluated using fragment amplitudes.

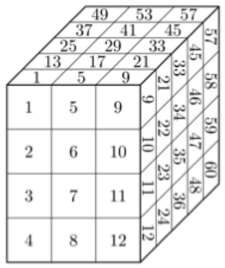
The accuracy is ensured by reaching self-consistency in the atomic fragment optimization.

The resulting scheme is linear-scaling (distant pair fragments can be neglected) and embarrassingly parallel (the fragment calculations are totally independent of each other).



The Machinery

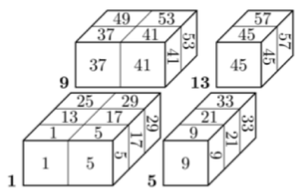
The ScaTeLib library; $C_{iljm} = \alpha \sum_k A_{ikj} B_{mkl} + \beta C_{iljm}$,



- *call sort* $\mathbf{A}[i_n, k_n, j_n]$ with $[1,3,2]$ to obtain $\mathbf{A}'[i_n, j_n, k_n]$
- *call sort* $\mathbf{B}[m_n, k_n, l_n]$ with $[2,1,3]$ to obtain $\mathbf{B}'[k_n, m_n, l_n]$
- *call dgemm* with α, $\mathbf{A}'$, $\mathbf{B}'$, β and obtain $\mathbf{C}'[i_n, j_n, m_n, l_n]$
- *call sort* $\mathbf{C}'[i_n, j_n, m_n, l_n]$ with $[1,4,2,3]$ to add to $\beta \cdot \mathbf{C}[i_n, l_n, j_n, m_n]$

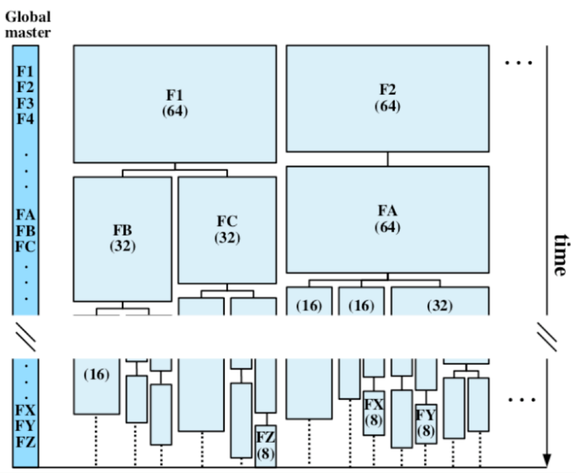
$$\mathbf{C}_{ILJM} = \alpha \sum_K \mathbf{A}_{IKJ} \mathbf{B}_{MKL} + \beta \mathbf{C}_{ILJM},$$

- load asynchronously tile A_{IKJ}
- load asynchronously tile B_{MKL}
- do the contractions as for the dense tensor with restricted indicies.



Parallelism:

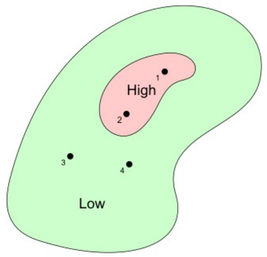
- Coarse grained: different fragments are calculated independently.
- Medium grained: Each fragment calculation is distributed over a group of nodes.
- Fine grained: Use of OpenMP/OpenACC.



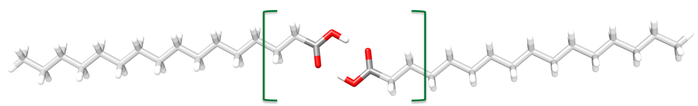
The TAL-SH Backend

TAL-SH: Tensor Algebra Library for Shared Memory Computers has been integrated with ScaTeLib to offload tensor contractions to GPUs. The execution of tensor contractions on GPU is pipelined, overlapping computation and data transfers.

The Multilayer Applications



The independent nature of the fragments allows for treatment of fragments at different levels of theory. For local effects, this can reduce the computational cost while preserving accuracy relative to the higher level of theory. This method is well suited to treat adsorption sites, defects, or catalytically active sites.



HF/RI-MP2

Model	Time (s)	ΔE (E_h)
DEC-RI-MP2	418	-0.027
ML-DEC	206	-0.033
ML-DEC [ext.]	239	-0.027

RI-MP2/CCSD

Model	Time (h)	ΔE (E_h)
DEC-CCSD	8.40	-0.027
ML-DEC	1.85	-0.026

Results:

- Significant reduction in time to solution, particularly for RI-MP2/CCSD.
- Convergence to high level reference with extension of the high level layer.
- Smaller high level layer required when low level ≠ HF.

References/Acknowledges

K. Aidas et al., *WIREs Comput. Mol. Sci.*, 4, 269 (2014)
T. Kjærgaard et al., WIREs Comput Mol Sci., 7, e1319. doi:10.1002/wcms.1319, (2017)
T. Kjærgaard et al., *Wires.* Comp. Phys. Comm. (**2017**), 212:152

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

[1]*National Center for Computational Sciences, Oak Ridge National Laboratory (ORNL), 1 Bethel Valley Rd, Oak Ridge, TN 37831, USA*