GPU acceleration of the Random Phase Approximation within the separable resolution-of-the-identity approach

TECHNISCHE UNIVERSITÄT DRESDEN

Moritz Leucke, Francisco A. Delesma-Diaz, Dorothea Golze

Chair of Theoretical Chemistry, Dresden University of Technology, Bergstr. 66c, Dresden, Germany
Email: moritz.leucke@tu-dresden.de

Introduction

Density functional theory (DFT) is a widely used method in electronic structure theory but struggles at describing non-local correlation effects like dispersion interaction. The Random Phase Approximation (RPA) can be applied as perturbative correction to DFT, offering a reliable treatment of electron correlation. However, the high computational cost of RPA has led to the development of low-scaling algorithms. Our low-scaling RPA implementation uses the space-time method[5] in combination with the separable resolution-of-the-identity (RI) approach[1,2], also known as real-space RI (RI-RS), to reduce the scaling from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^3)$. The key quantity in RPA, the polarizability, is computed on an optimized real-space grid in imaginary time and efficiently transformed into the imaginary frequency domain. Here, we present our work on GPU offloading for our low-scaling RPA implementation to reduce the computational pre-factor.

RPA Workflow

DFT

RI-V coefficients

$$C^P_{\mu\nu} = \sum_Q (\mu\nu\,|\,Q)_m\, M^{-1}_{QP}$$

RI-RS coefficients fit

$$\min_{M_{Pk}} \sum_{\mu\nu P} \left(\left[C^P_{\mu\nu}\right]_{\text{RI-RS}} - \left[C^P_{\mu\nu}\right]_{\text{RI-V}} \right)^2$$

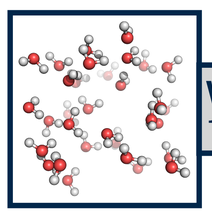
Polarizability

$$\chi^0_{PQ}(i\tau) = \sum_{kk'} M_{Pk}\, \chi^0(\mathbf{r}_k, \mathbf{r}'_k, i\tau)\, M_{k',Q}$$

E_c^{RPA}

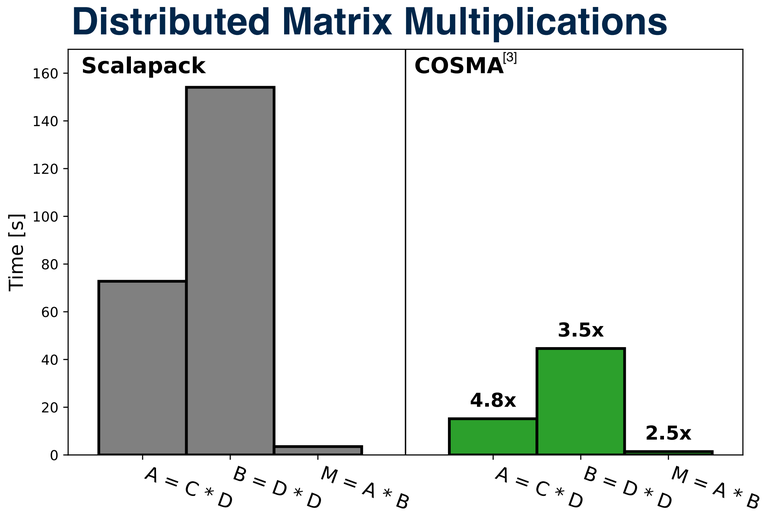
RI-RS Coefficients Fit

Linear least-squares estimator: $\mathbf{M} = \mathbf{A}\mathbf{B}^{-1} = \mathbf{C}\mathbf{D}\,(\mathbf{D}^T\mathbf{D})^{-1}$

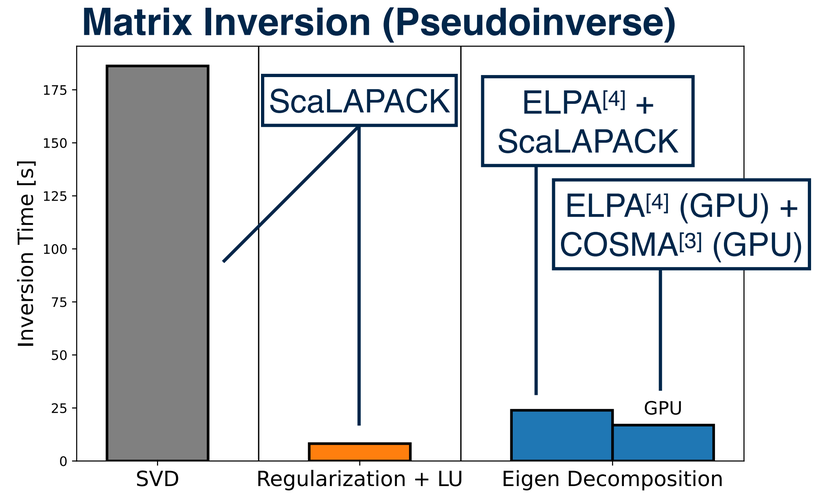


Water cluster (96 atoms)
144 MPI tasks + 12 GPUs

Distributed Matrix Multiplications



Matrix Inversion (Pseudoinverse)



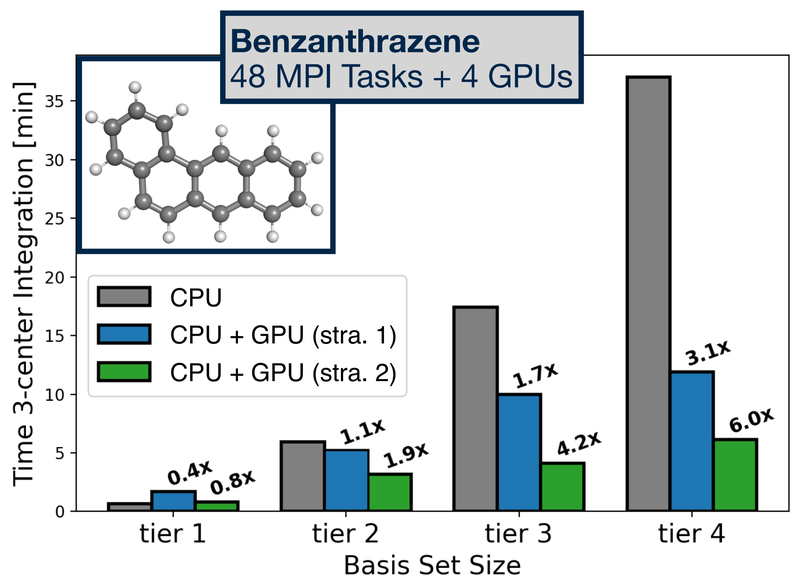
B	$\mathbf{U}\Sigma\mathbf{V}^*$	$\mathbf{D}^T\mathbf{D}$	$\mathbf{Q}\Lambda\mathbf{Q}^T$
$\mathbf{B}^+$	$\mathbf{V}\Sigma^{-1}\mathbf{U}^*$	$\mathbf{d}\,(\tilde{\mathbf{D}}^T\tilde{\mathbf{D}} + \epsilon\mathbb{I})^{-1}\,\mathbf{d}$	$\mathbf{Q}\Lambda^{-1}\mathbf{Q}^T$
Accuracy			

RI-V Coefficients

$$(\mu\nu Q) = \int \phi_\mu(\mathbf{r})\phi_\nu(\mathbf{r})\Omega_Q[m](\mathbf{r})\, d\mathbf{r}$$

$$= \sum_{\substack{\text{atoms}\\ a}} \sum_{\substack{\text{radial}\\ \text{shells}\\ s}} \sum_{\substack{\text{angular}\\ \text{points}\\ t}} w(\mathbf{r}_{\mathbf{st}})\phi_\mu(\mathbf{r}_{\mathbf{ast}})\phi_\nu(\mathbf{r}_{\mathbf{ast}})\Omega_P(\mathbf{r}_{\mathbf{ast}})$$

GPU



Implementation strategy 1

```
for each atom
  for each point batch
    evaluate_functions_on_points
    for each basis function
      compute_local_indices
      send indices_to_GPU
      prepare_arrays (CUDA kernels)
      multiply_functions (cublas)
      add_to_resultarray (CUDA kernels)
```

Implementation strategy 2

```
for each atom
  for each point batch
    evaluate_functions_on_points
    compute_local_indices
    send indices_to_GPU
    for each basis function
      prepare_arrays (CUDA kernels)
      multiply_functions (cublas)
      add_to_resultarray (CUDA kernels)
```

memory copies (CPU to GPU)
work on GPU

Polarizability

Matrix multiplications of large distributed Matrices:

$$\chi^0_{PQ}(i\tau) = \sum_{kk'} M_{Pk}\, \chi^0(\mathbf{r}_k, \mathbf{r}'_k, i\tau)\, M_{k',Q}$$

ScaLAPACK 2D distribution

COSMA
Communication-Optimal Matrix-Multiplication

COSMA Library[3]**:**

- Near communication-optimal distributed MM multiplication
- Same API as ScaLAPACK pdgemm
- GPU acceleration
- But higher memory footprint

Polarizability calculation **~ 2x faster** when using COSMA instead of ScaLAPACK

References

[1] **Delesma, F. A.**, et al., *J. Chem. Phys.* 160, (**2024**).

[2] Duchemin, I., et al., *J. Chem. Phys.* 150, (**2019**).

[3] Kwasniewski, G., et al., *Proc. - Int. Conf. High Perform. Comput. Netw. Storage Anal.,* (**2019**).

[4] Yu, V. W., et al., *Comp. Phys. Comm.* 262, (**2021**).

[5] Rojas, H. N., et al., *Phys. Rev. Lett.* 74, (**1995**).

Funded by: Emmy Noether-Programm, DFG Deutsche Forschungsgemeinschaft

Computing time: GCS Gauss Centre for Supercomputing

Research stay: Argonne NATIONAL LABORATORY

Software package: FHI-aims The ab initio materials simulation package