A GPU-accelerated Fast Multipole Method for GROMACS: performance and accuracy

Bartosz Kohnke, **Carsten Kutzner**, and Helmut Grubmüller

Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

PLEASE CHECK POSTER B304 FOR CUDA FMM IMPLEMENTATION DETAILS!

Introduction

- Task: Compute Coulomb forces $\mathbf{f}_i$ acting on N atoms at positions $\mathbf{x}_i$ with partial charges q_i for $N \approx 1{,}000 - 10$ M atoms in periodic boundary conditions (PBC) in molecular dynamics (MD) simulation software such as GROMACS[1]

$$\mathbf{f}_i = q_i \sum_{j=1(j\neq i)}^{N} q_j \frac{\mathbf{x}_i - \mathbf{x}_j}{||\mathbf{x}_i - \mathbf{x}_j||^3} \quad \text{for } i = 1, ..., N \qquad (1)$$

- The demand to study increasingly large MD systems is growing, $N = 10^8 - 10^9$ could become routine soon
- Direct calculation has $O(N^2)$ complexity and is impossible with PBC
- For numerical stability resons, the time step in MD can be no longer than a few femtoseconds
- Long trajectories are needed to collect enough sampling and to reach biologically relevant time scales
- Thus, the time required to finish one MD step needs to be shortened to a millisecond or less so that long enough trajectories can be produced in reasonable time
- The prevalent method to compute Coulomb forces in MD is Particle Mesh Ewald

Particle Mesh Ewald (PME)

- PME uses Ewald summation to split up the calculation in a short-range part, for which all interactions up to a cutoff are computed directly, and a long-range part, which is solved in reciprocal space
- To take advantage of fast Fourier transforms (FFTs) for the conversions to and from reciprocal space, the charges are interpolated onto a uniform grid
- PME computation scales with $O(N \log(N))$, but in parallel the communication for the FFTs grows quadratically with the number of involved processes. While PME is extremely fast on a single processor, parallel PME suffers from a scaling bottleneck[1]
- Additionally, due to the underlying uniform grid, PME becomes memory limited if high accuracies are needed or if the atoms are spread inhomogeneously across large volumes

The Fast Multipole Method (FMM)

- FMM[2] is an alternative way for the rapid evaluations of Coulomb forces, which does not suffer the aforementioned limitations and even scales with $O(N)$
- While FMM won't be able to beat PME performance for small MD systems, it is likely competitive for large N at high parallelization, and for inhomogenous atom distributions. FMM additionally allows for open boundaries.
- FMM splits the calculation into a directly calculated near field, and a far field. For the far field, groups of sufficiently separated point charges are combined and described as spherical harmonics-based multipole expansions truncated at order p
- Grouping is accomplished by recursively subdividing the simulation box into sub-boxes in an octree fashion, i.e. each parent box is subdivided into 8 child boxes when the tree depth d is increased
- On the lowest tree level, interactions within a box or between directly neighboring boxes are directly calculated (near field), whereas al other interactions are approximated via multipoles (far field)

CUDA implementation

- Our GPU-FMM[3,4] is based on the ScaFaCos[5] FMM, which we parallelized using CUDA and optimized for GROMACS as a drop-in PME replacement
- The implementation is described on poster B304 by Bartosz Kohnke

References

1. Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E.: Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In: Solving Software Challenges for Exascale: International Conference on Exascale Applications and Software, EASC 2014, Stockholm, Sweden, April 2–3, 2014, pp. 3–27 (Eds. Markidis, S.; Laure, E.). Springer (2015)
2. Greengard, L., Rokhlin, V. A new version of the Fast Multipole Method for the Laplace equation in three dimensions. Act. Num., 6, 229–269 (1997)
3. Kohnke, B.; Kutzner, C.; Beckmann, A.; Lube, G.; Kabadshow, I.; Dachsel, H.; Grubmüller, H.: A CUDA fast multipole method with highly efficient M2L farfield evaluationfield evaluation. IJHPCA, 35(1), pp. 97–117 (2020)
4. Kohnke, B.; Kutzner, C.; Grubmüller, H.: A GPU-accelerated fast multipole method for GROMACS: Performance and accuracy. JCTC 16(11), pp. 6938–6949 (2020)
5. Arnold, A., Fahrenberger, F., Holm, C., Lenz, O., Bolten, M., Dachsel, H., Halver, R., Kabadshow, I., Gähler, F., Heber, F. and Iseringhausen, J. Comparison of scalable fast methods for long-range interactions. Phys. Rev. E, 88(6), (2013)

Methods

- Here, we assess the performance of our CUDA FMM implementation and compare its accuracy to GROMACS' PME implementation
- To that end, we first verify that our implementation yields accurate energies and forces by comparing to known reference solutions
- We then determine which FMM parameters reproduce the accuracy with PME concerning energies, forces and energy drift
- Finally we use typical MD systems to compare FMM and PME performances in GROMACS 2019

Benchmark Systems

- ***Infinite ideal crystal:*** A (32 nm)^3 PBC box with alternating +1e positive and –1e negative charges at 0.5, 1.5, 2.5, ... 30.5, 31.5 nm in each dimension, in total 32,768 charges
- ***Periodic salt water:*** A (8 nm)^3 PBC box with 16,861 TIP3P waters + 46 Na^+ + 46 Cl^- ions, in total 50k atoms, 300 K, 1 bar
- ***Water droplet:*** As above, but in a (14 nm)^3 box with open boundaries
- ***Aerosol / multi-droplet system:*** 75 small water droplets in a (135.6 nm)^3 box + 63 Na^+ + 63 Cl^- ions (distributed in the droplets), in total 109 k atoms (Fig. 1)
- ***Water boxes of increasing size:*** A series of cubic boxes of side lengths 3.13 – 67.4 nm containing 1,000 – 10,000,000 TIP3P waters, i.e. 3 k – 30 M atoms
- ***Random charges:*** 1,000 – 286,000,000 randomly distributed charges in a (100 nm)^3 box (FMM standalone test without GROMACS)
- All GROMACS benchmarks use a 4 fs time step.

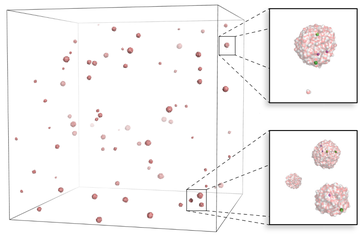

Fig. 1: Sketch of the multi-droplet aerosol system. Close-ups show individual water droplets (red/white) with Na^+ ions in magenta and Cl^- ions in green.

Conclusions

- We have assessed the accuracy and perfomance of our GPU FMM implementation described in Kohnke et al. (2020)[3]
- We demonstrated that our implementation provides correct electrostatic energies and forces for single and double precision by comparison to reference solutions for open and periodic systems
- For a representative MD system of 50 k atoms in size, a multipole order $p = 8$ at depth $d = 3$ yields similar accuracy in energy and forces as well as in energy drift as with default PME parameters
- For typical biomolecular simulation systems of up to 30 M atoms in size, GROMACS 2019 performance with CUDA FMM is about a third of the performance with PME
- For large systems with nonuniform particle distributions, such as our 100 k atom aerosol benchmark, FMM easily outperforms PME

Results and discussion

Results I. FMM convergence and correctness

- First we show that our implementation converges to the correct solution with increasing multipole order p by comparing FMM computed energies and forces to a reference solution

How accurate is FMM for open boundaries?

- For open boundaries, the reference solution is gained by directly evaluating all Coulomb interactions in double precision
- For the water droplet, we compare the reference solution to FMM for different d and p (Figs. 2–3)
- As can be seen, force errors decrease exponentially with growing p and begin to saturate at $p = 40$.
- For open boundaries, FMM forces are as accurate as with direct summation for high multipole orders p. In single precision, $p = 12$ reaches the numerical limits.

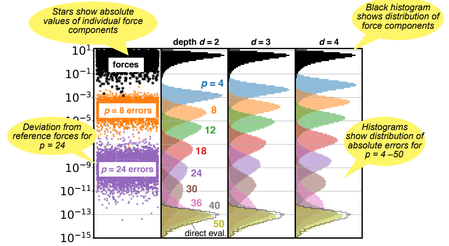

Fig. 2: FMM errors for the water droplet in double precision.

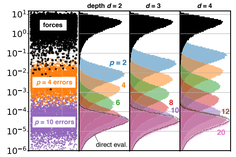

Fig. 3: FMM errors for the water droplet in single precision.

Comparison to analytical solution for periodic boundaries

- Next, we checked the FMM for periodic boundaries
- For simple periodic charge distributions like the infinite ideal crystal, the reference can be obtained analytically
- With growing p, our implementation converges to the correct periodic solution (Fig. 4). Reaching relative accuracies at the numerical limit at $p = 40$ verifies that the treatment of PBC in our implementation is correct

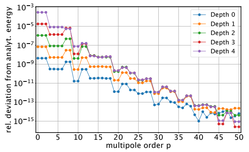

Fig. 4: FMM energy error for the ideal crystal (double precision). Dots show relative deviation of the FMM energy from the correct value for different p and d.

Results II. Comparison to PME

- After establishing the correctness of our FMM impementation, we compared it to PME
- Which FMM parameters p and d yield accuracies similar to representative PME settings?
- For the periodic salt water system, we compute a reference solution using our FMM with $p = 50$ at depth $d = 0$. Fig. 5 shows the errors in the Colomb forces for various FMM and PME parameters
- In single precision $p = 7$ and $d = 3$ yields as accurate Coulomb forces as the default PME parameters

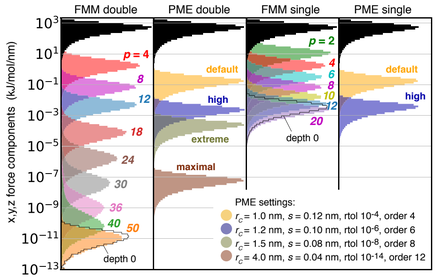

Fig. 5: Accuracy of FMM and PME Coulomb forces for a snapshot of the 50 k atom periodic salt water system for double precision (left two panels) and single precision (right two panels). FMM depth $d = 3$ throughout here.

Energy conservation

- We need to make sure that FMM does not increase the overall energy drift when it is used as a PME replacement
- Therefore, we monitored the drift of the total energy for the periodic salt water system when run in the NVE ensemble with various FMM settings in comparison to default PME settings (Fig. 6)
- With FMM, at $d = 3$, the PME default drift level is met for multipole orders of 8 or larger

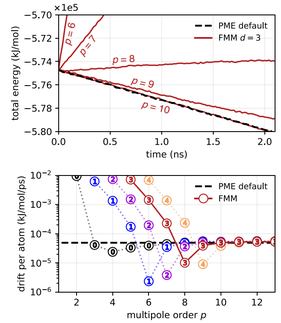

Fig. 6: Drift of the total energy at typical single precision settings for the periodic salt water system. Upper panel shows drift versus time, lower panel the absolute drift derived from a linear fit of the curves from the upper panel.

Results III. Performance of GROMACS w/ FMM

- We have established that $p = 8$ at $d = 3$ achieves the same approximation quality as PME with default parameters in terms of force and energy accuracy and energy drift
- Therefore, we compare FMM to PME at these parameters (Figs. 7 and 8)
- For the homogeneous salt water system, GROMACS with GPU FMM reaches about a third of the GPU PME performance
- For the strongly inhomogeneous multidroplet aerosol the situation reverses: FMM outperforms PME by more than a factor of two

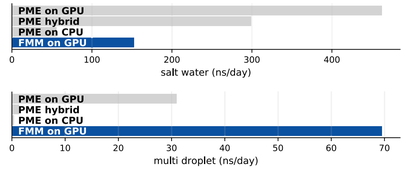

Fig. 7: FMM vs. PME performance for two benchmark systems run on a 10-core Xeon E5-2630v4 with RTX 2080Ti GPU.

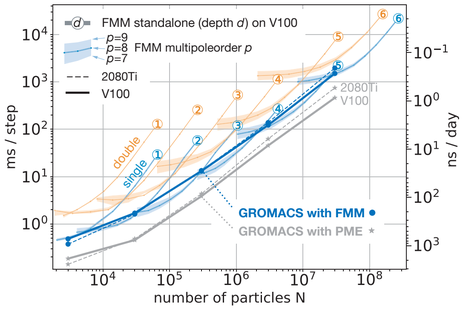

Fig. 8: FMM and PME scaling with respect to system size N for up to 268 million charges. Benchmarks were run on a 10-core Xeon E5-2630v4 with RTX 2080Ti GPU (dashed) and 20-core Xeon Gold 6148F with V100 GPU (solid) with all nonbonded interactions offloaded to the GPU.

Acknowledgments

- This study was supported by the DFG priority program „Software for Exascale Computing" (SPP 1648). Frank Wiederschein provided the multidroplet system. We thank Ivo Kabadshow and Andreas Beckmann (Jülich Supercomputing Centre) for many discussions about FMM electrostatics and for providing the ScaFaCos FMM, which served as a starting point for this CUDA implementation. The benchmarks on the V100 GPUs were done at the Max Planck Computing and Data Facility in Garching.