Local vs. Global FFT Approaches for High-Performance Ultrasound Simulation on Multi-GPU Systems

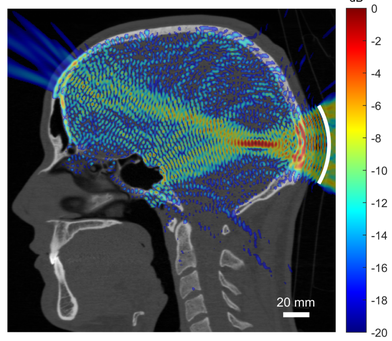
Oliver Kunik and Jiri Jaros

Faculty of Information Technology, Brno University of Technology, CZ

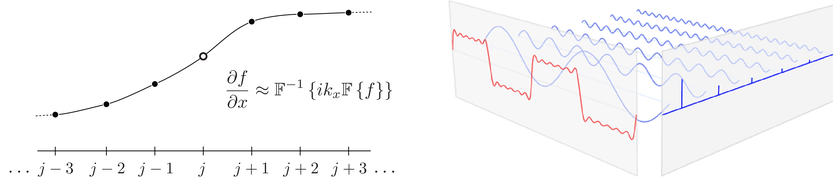
Overview

Transcranial ultrasound therapy is an emerging technology for treating brain disorders. Delivering energy to precise targets is challenging due to skull-induced reflections and distortions. Numerical models can predict and correct these effects, but they are computationally expensive.

Wave propagation simulation is demanding in both space and time. The Fourier collocation method reduces spatial discretization requirements to nearly the theoretical minimum of two points per wavelength. However, it relies on discrete Fourier transforms (DFTs) for spatial derivatives, which generally scale poorly on HPC systems.

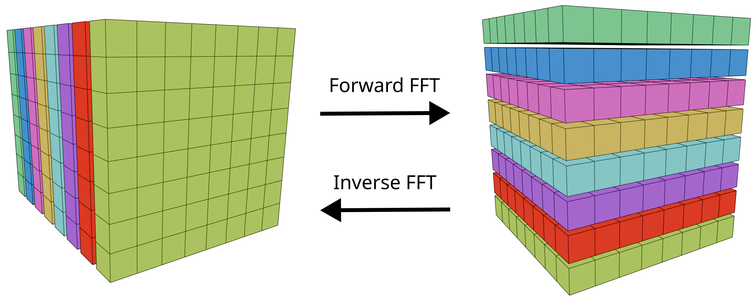


Using DFT for Simulating Ultrasound



DFTs require periodic boundary conditions, enforced with a perfectly matched layer (PML). The PML attenuates wave amplitudes to near zero at domain boundaries, preventing reflections. In three-dimensional simulations, up to 14 fast Fourier transforms (FFTs) are performed per time step. This high computational cost, combined with communication overhead, motivates the exploration of efficient multi-GPU strategies.

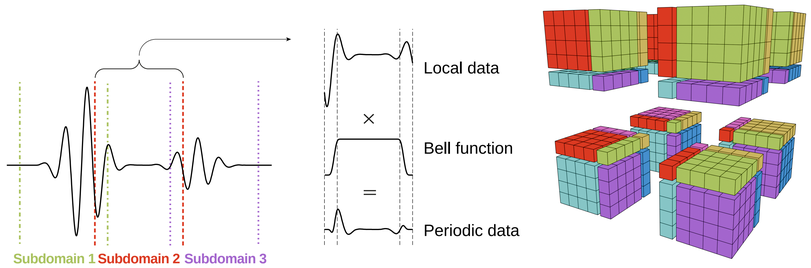
Global FFT Approach



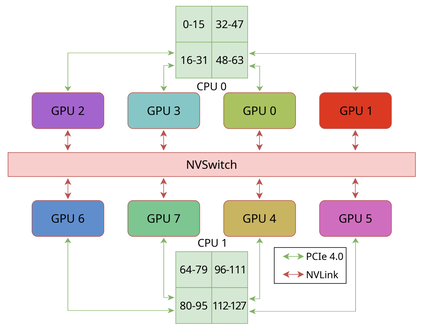
The **Global FFT** approach distributes computation and data across multiple GPUs. In the spatial domain, data are partitioned along the Z-axis; in the frequency domain, they are partitioned along the Y-axis. Multi-GPU FFTs are executed using NVIDIA's cuFFTXt API, which supports distributed FFTs and employs blocking to overlap communication with computation. While straightforward to implement, this method faces scaling bottlenecks due to the global data transpositions required for all 14 FFTs per time step.

Local FFT Approach

In the **Local FFT** approach, the computational domain is divided into subdomains, each processed on a separate GPU, with periodic halo exchanges ensuring accurate wave propagation. Each subdomain requires its own PML, whose thickness defines the minimum halo size. At every time step, halo zones from six matrices are exchanged directly between GPU memories using CUDA-aware MPI. A bell-shaped function blends incoming data with the local PML to maintain periodicity and accuracy.



FFT Performance and Multi-GPU System Used



The FFTs are the most time-consuming part of the simulation. cuFFT performance is highly dependent on domain size and is optimized for dimensions with prime factors ≤ 7. To ensure a fair comparison between the two approaches, the global domain size was chosen with maximal prime factor 7, and local domain sizes were padded (halo zones enlarged) to meet the same condition.

All tests were performed on a multi-GPU machine with 8 NVIDIA A100 40 GB GPUs, interconnected via NVLink through NVSwitch, providing a high-speed interconnect bandwidth of 600 GB/s, as shown in the figure.

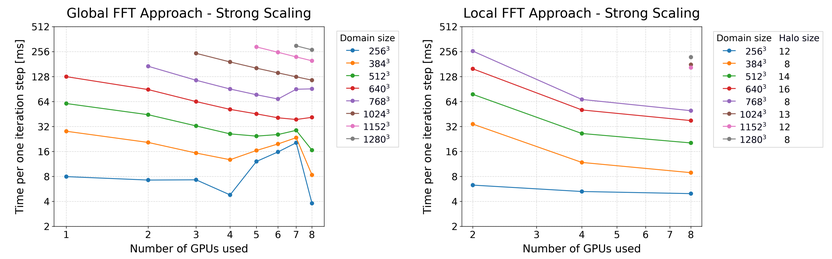
Simulation Precision

Simulation precision was compared between the two approaches using different domain and halo sizes. The error metric was $L_\infty = \frac{\max|P_{glob} - P_{loc}|}{\max|P_{glob}|}$, where P is the final acoustic pressure.

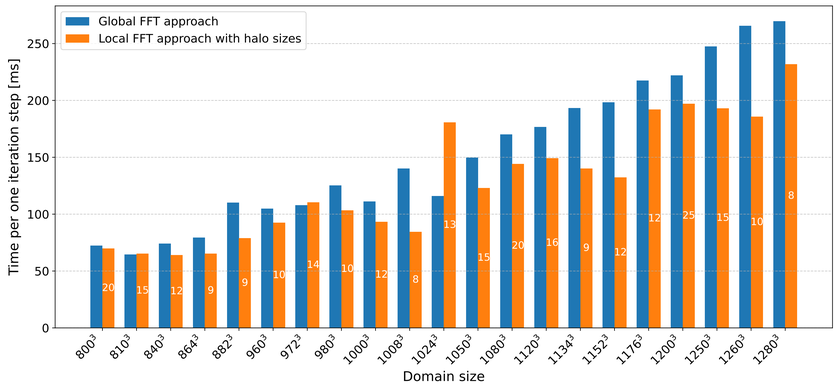
Domain size	Halo zone size	L_∞
800^3	20	9.43×10^{-4}
1008^3	8	1.22×10^{-2}
1050^3	15	1.81×10^{-3}
1200^3	25	1.23×10^{-3}
1260^3	10	6.13×10^{-3}

The table shows that larger halo zones reduce the error, partly due to the local PMLs of size 10 in the **Local FFT** approach. The halo zone must be at least as wide as the local PML to avoid attenuating transmitted waves.

Speed Comparison

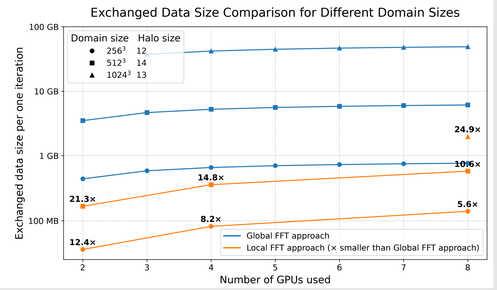


Both approaches show speedup with additional GPUs, except for small domains with the **Global FFT** approach, where scaling is limited. For the **Local FFT** approach, the 1152^3 domain outperforms the smaller 1024^3 case because the subdomains have more favorable prime factors, thus faster FFTs.



On 8 GPUs, the **Local FFT** approach is consistently faster in most scenarios. Further optimizations, such as overlapping halo exchanges with computation and using the NCCL communication library, are expected to improve performance even more.

The **Local FFT** approach also exhibits significantly lower communication overhead. It only needs to exchange the halo zones 6 times per iteration instead of 14 whole matrix exchanges, making it particularly well-suited for scaling to larger multi-node simulations.



Conclusion

The **Local FFT** approach demonstrates superior performance and scalability compared to the **Global FFT** approach on multi-GPU systems, while maintaining acceptable accuracy. Its lower communication overhead, combined with potential optimizations and ongoing research into optimal domain, subdomain, and halo sizes for FFTs, makes it a promising strategy for large-scale, multi-node, multi-GPU ultrasound simulations.

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).
This project has received funding from the European Unions Horizon Europe research and innovation programme under grant agreement No 101071008.
This work was supported by Brno University of Technology under project number FIT-S-23-8141.
This work was supported by Brno University of Technology under project number FIT/FSI-J-25-8755.