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

Oliver Kunik
ikunik@fit.vutbr.cz
Faculty of Information Technology, Brno University of Technology
Brno, Czech Republic

Jiri Jaros
jarosjir@fit.vutbr.cz
Faculty of Information Technology, Brno University of Technology
Brno, Czech Republic

Abstract

Simulating wave propagation with the Fourier collocation method is computationally intensive due to its reliance on discrete Fourier transforms (DFTs). While DFTs enable near-minimal spatial discretization, they scale poorly on modern high-performance computing systems. This work evaluates two multi-GPU strategies for three-dimensional simulations: a Global FFT approach using distributed transforms and a Local FFT approach based on domain decomposition with halo exchanges. Experiments were performed on system with 8 NVIDIA A100 GPUs connected via NVSwitch. Precision tests show that the Local FFT approach maintains errors around 0.1% when the halo covers the local PML region. Performance results demonstrate that the Local FFT approach achieves lower runtimes and significantly reduced communication overhead compared to the Global FFT approach, particularly for larger domains. These findings indicate that Local FFT decomposition is a promising strategy for scalable, large-scale multi-node ultrasound simulations.

CCS Concepts

• **Computing methodologies** → **Massively parallel algorithms**; Simulation evaluation; • **Applied computing** → *Physics*.

Keywords

k-Wave, wave propagation, Fourier collocation method, FFT, multi-GPU, domain decomposition, CUDA, high-performance computing, MPI

ACM Reference Format:

Oliver Kunik and Jiri Jaros. 2025. Local vs. Global FFT Approaches for High-Performance Ultrasound Simulation on Multi-GPU Systems. In *Proceedings of The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '25)*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Wave propagation simulation is computationally demanding in both spatial and temporal dimensions. The Fourier collocation method offers a key advantage: it reduces the spatial discretization requirement close to the theoretical minimum of two points per wavelength. However, it relies on discrete Fourier transforms (DFTs) for spatial derivatives, which generally scale poorly on modern HPC systems.

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

2 Global FFT Approach

The **Global FFT** approach distributes both computation and data across multiple GPUs. In the spatial domain, data are partitioned along the Z-axis, while in the frequency domain they are partitioned along the Y-axis, reducing communication during FFT execution. Multi-GPU FFTs are performed using NVIDIA's `cufftXt` API, which supports distributed transforms and overlaps communication with computation. Although relatively straightforward to implement, this method can suffer from scaling bottlenecks caused by the global data transpositions required for all 14 FFTs in each time step.

3 Local FFT Approach

To further reduce execution time, a domain decomposition strategy is employed. The computational domain is divided into multiple subdomains, each processed independently on a separate GPU. Communication between subdomains is maintained through periodic halo exchanges, ensuring accurate wave propagation across subdomain boundaries.

When using local FFTs, each subdomain requires its own PML to maintain the periodicity condition necessary for FFT-based differentiation. The PML thickness defines the minimum halo size required to capture the full wavefield before attenuation; a smaller halo would only capture attenuated waves, leading to inaccuracies in inter-domain propagation.

In each time step, halo zones from six matrices must be exchanged between neighboring subdomains. CUDA-aware MPI is employed to transfer these halo zones directly between GPU memories, avoiding unnecessary staging through the host. Since the exchanged halo zones overwrite the attenuated regions of the PML, a bell-shaped blending function is applied to smoothly merge the incoming data with the local domain.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SC '25, St. Louis, Mo, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4 Experimental Setup

The FFTs represent the most computationally demanding part of the simulation. cuFFT performance depends strongly 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 selected with a maximal prime factor of 7. In the **Local FFT** approach, the domain was divided in half along each axis to create 8 subdomains. The halo zone size was chosen to cover at least 80% of the PML while also ensuring subdomain dimensions with maximal prime factors of 7.

All experiments were conducted on an accelerated compute node of the IT4Innovations supercomputer *Karolina*. Each node consists of two AMD EPYC CPUs, 1 TB of DDR4 RAM, and eight NVIDIA A100 GPUs with 40 GB of HBM2 memory each. The GPUs are interconnected via NVSwitch using NVLink, providing a high-bandwidth communication fabric with an aggregate bandwidth of up to 600 GB/s. The simulation core was implemented entirely in CUDA for both approaches.

5 Simulation Precision

Simulation precision was assessed using an initial spherical pressure source placed at the center of the domain. The wavefield was propagated until just before reaching the outer PML layer, at which point the pressure distribution was recorded. The results obtained with the two approaches were compared using the error metric

$$L_\infty = \frac{\max \left| P_{\text{glob}} - P_{\text{loc}} \right|}{\max \left| P_{\text{glob}} \right|} \tag{1}$$

where P denotes the final acoustic pressure. The corresponding results are summarized in Table 1.

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}

Table 1: Error comparison between Global FFT and Local FFT approaches with varying domain and halo zone sizes.

The results demonstrate that larger halo zones consistently reduce the error, as the halo must fully cover the local PML region to avoid attenuating transmitted waves. For sufficiently large halo zones, the error decreases to approximately 0.1%, which is acceptable for practical simulations. With single-precision arithmetic, the theoretical accuracy limit is about 10^{-6}, so errors below this limit cannot be expected.

6 Speed Comparison

Both approaches show reduced execution time when additional GPUs are used, although the scaling efficiency is not ideal. For small domains in the **Global FFT** approach, the limited amount of work per GPU restricts scalability.

With larger domains, execution time does not increase monotonically, as certain dimensions remain more favorable for FFT performance under the prime-factor constraint. On 8 GPUs, the **Local FFT** approach achieves lower runtimes in most cases. Further performance improvements are expected from overlapping halo exchanges with computation and from employing the NCCL communication library to optimize GPU-to-GPU transfers.

The **Local FFT** approach also benefits from significantly reduced communication costs. Instead of 14 global matrix transpositions per iteration, it requires only six halo zone exchanges, making it more communication-efficient and better suited for scaling to larger multi-node simulations. For example, for a domain of size 1024^3 distributed across 8 GPUs, the communication volume is reduced by approximately 96%.

7 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 the potential for further optimizations, makes it a strong candidate for large-scale, multi-node, multi-GPU ultrasound simulations. Future work will focus on systematically exploring domain sizes, subdomain partitioning strategies, and halo configurations to maximize performance across diverse HPC platforms.

Acknowledgments

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. Text polishing was assisted by OpenAI's GPT-5 language model. All scientific content and interpretations remain the sole responsibility of the authors.