Acceleration of Ultrasound Neurostimulation Using Mixed-Precision Arithmetic

Jiri Jaros
jarosjir@fit.vutbr.cz
Brno University of Technology
Brno, Czech republic

Radek Duchon
radek.dchn@gmail.com
Brno University of Technology
Brno, Czech republic

ABSTRACT

Ultrasound neurostimulation, a technique that modulates the brain's electrical activity, has emerged as a significant secondary treatment option for cases resistant to pharmacological interventions. The therapy is achievable through the application of a three-dimensional steerable ultrasound, directed by patient-specific stimulation plans. These plans are meticulously crafted through full-wave ultrasound propagation simulations. Nonetheless, the computational intensity required for calculating these plans poses a significant challenge, often reaching the memory capacities of contemporary graphics processing units (GPUs). By representing material properties and k-space operators more efficiently, we achieved up to 22% reduction in GPU memory usage, while accelerating calculations by 8.5% on an Nvidia Volta V100. This optimization introduced an error that reduced focal pressure by 0.5% without any focus movement, values that are clinically acceptable.

CCS CONCEPTS

• **Software and its engineering** → *Software performance*; • **Mathematics of computing** → *Solvers*; *Partial differential equations.*

KEYWORDS

GPU, Nvidia, CUDA, k-Wave, Acceleration, Ultrasound, Acoustic waves, Neurostimulation, Mixed precision

ACM Reference Format:
Jiri Jaros and Radek Duchon. 2024. Acceleration of Ultrasound Neurostimulation Using Mixed-Precision Arithmetic. In *The 33rd International Symposium on High-Performance Parallel and Distributed Computing (HPDC '24), June 3–7, 2024, Pisa, Italy.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3625549.3658823

1 INTRODUCTION

Disorders of the brain, including neurological and psychiatric diseases, affect one in four people[1]. The personal impact can be devastating and societal costs are enormous (5.9% GDP EU). New treatment options are needed with enhanced efficacy and reduced side-effects, costs, and invasiveness. Yet, brain disorders are among the medical conditions most difficult to treat. This is caused not only by the complexity of human brain anatomy and function, but also by the intricate challenge of targeting specific subregions and networks in an anatomically precise manner to modulate dysfunctional neural activity. Pharmacological interventions, the first-line treatment for most brain disorders, act not only on the entire brain but also the remaining organism, and are therefore often associated with considerable systemic side-effects.

Neurostimulation techniques that modulate the electrical activity of the brain have thus evolved as an important class of second-line treatments for pharmacoresistant cases. What is needed is a non-invasive brain stimulation technique that can stimulate brain targets with high anatomical precision, unlimited penetration depth, full reversibility, and low risk-profile. This can be achieved using the newly emerging technique of low-intensity focused transcranial ultrasonic stimulation for neuromodulation.

This objective is achievable through the application of a three-dimensional steerable ultrasound, directed by patient-specific stimulation plans. These plans are meticulously crafted through full-wave ultrasound propagation simulations. Nonetheless, the computational intensity required for calculating these plans poses a significant challenge, often reaching the memory capacities of contemporary graphics processing units (GPUs).

[1]World Health Organization (2001)https://apps.who.int/iris/handle/10665/42390

2 IMPLEMENTATION

The tissue realistic models of ultrasound wave propagation in the human body have to take into account many specific aspects. A defactor standard in this area is represented by the k-space corrected pseudospectral model implemented the k-Wave toolbox [2]. The required governing equations can be written as three-coupled first-order partial differential equations derived from the conservation laws and a Taylor series expansion for the pressure about the density and entropy:

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} &= -\frac{1}{\rho_0}\nabla p + \mathbf{F} \\ \frac{\partial \rho}{\partial t} &= -\rho_0 \nabla \cdot \mathbf{u} - \mathbf{u} \cdot \nabla \rho_0 - 2\rho \nabla \cdot \mathbf{u} + \mathrm{M} \\ p &= c_0^2 \left(\rho + \mathbf{d} \cdot \nabla \rho_0 + \frac{B}{2A}\frac{\rho^2}{\rho_0} - \mathrm{L}\rho \right) \end{aligned} \tag{1}$$

Here $\mathbf{u}$ is the acoustic particle velocity, $\mathbf{d}$ is the acoustic particle displacement, p is the acoustic pressure, ρ is the acoustic density, ρ_0 is the ambient (or equilibrium) density, c_0 is the isentropic sound speed, and B/A is the nonlinearity parameter which characterises the relative contribution of finite-amplitude effects to the sound speed. All the material parameters are allowed to be heterogeneous. Two linear source terms are also included, where $\mathbf{F}$ is a force source term, and M is a mass source term.

HPDC '24, June 3–7, 2024, Pisa, Italy

ACM ISBN 979-8-4007-0413-0/24/06…$15.00
https://doi.org/10.1145/3625549.3658823

Table 1: The summary of implemented reduction levels.

Quantity	Data type	Reduction Level		
		Low	Med	High
Nonlinearity coefficient	Half	1	1	1
Sound Speed	Half	1	1	1
Density	Half	1	4	4
Reference Density	Half	0	0	3
Initial Pressure	Brain float	1	1	1
Pressure Source	Brain float	1	1	1
Absorption Tau and Eta	Brain float	2	2	2
Absorption Nabla	Brain float	1	1	1
Kappa Operator	Half	0	0	0.5
Kappa Source Operator	Half	0.5	0.5	0.5
Reduced matrices		8.5	11.5	15
Memory reduction		10.5%	15.7%	21.8%

2.1 Simulation Code Description

The simulation code was implemented in the C++ and CUDA languages using the standard 32b float data type, with the help of HDF5 IO library and the cuFFT library for Fourier transform calculation.

To improve accuracy and decrease the spatial resolution needed, k-space pseudospectral methods utilize a 3D fast Fourier transform (FFT) for gradient calculation across the entire domain, accounting for 50-60% of the execution time. Apart from FFTs, the simulations conduct straightforward element-wise matrix operations through approximately 20 CUDA compute kernels on over 30 real or complex matrices containing acoustic quantities and medium parameters.

2.2 Mixed-Precision Arithmetic

An innovative strategy showcased in this approach is the reduction of the simulation's memory requirements by utilizing reduced precision data types for storing specific quantities. This not only diminishes the memory footprint but may also accelerate execution by enabling the simultaneous performance of two operations in half precision.

Due to the high dynamic range needed for Fourier transforms and the disparate units of measure used for acoustic pressure and acoustic particle velocity, these quantities and their gradients are excluded from reduction. However, material properties, with their 2-3 digit precision, are suitable for reduced data types.

Table 1 outlines the quantities whose accuracy was decreased and specifies the employed data type, either the 16-bit half floating point or the Google Brain 16-bit floating point format. The half data type was utilized for basic material properties which have a low dynamic range. Conversely, the Brain float data type was applied to pressure sources and absorption coefficients, whose dynamic range can span up to six orders of magnitude.

This study examines three levels of memory reduction, aiming to harmonize computational efficiency with accuracy. The low level targets basic material properties that remain constant throughout the computation. The medium level additionally encompasses acoustic density, recalculated at each time step. The high level further incorporates the kappa derivation operator, directly influencing gradient computation.

Table 2: Simulation benchmarks. Nx, Ny, and Nz denote spatial grid sizes, and Nt indicates the total number of time steps.

Dataset	Nx	Ny	Nz	Nt	GPU Memory
PH1-BM7-SC1	324	192	192	3600	1.78 GB
PH1-BM8-SC1	512	384	432	12000	8.60 GB
PH1-BM9-SC1	512	512	432	12000	11.3 GB

Table 3: Execution time reached on an Nvidia Volta V100.

Dataset	Full	Low	Mid	High
PH1-BM7-SC1	20.09s	19.39s (3.61%)	18.79s (6.92%)	18.51s (8.54%)
PH1-BM8-SC1	449s	435s (3.27%)	423s (6.31%)	415s (8.25%)
PH1-BM9-SC1	597s	577s (3.33%)	560s (6.39%)	550s (8.42%)

The CUDA compute kernels were refactored to operate in mixed precision, with quantities stored in reduced precision being read from memory in pairs. When feasible, calculations in reduced precision allowed for simultaneous execution of two operations. If not, data was extended to higher precision (float data type) for sequential operation. Only CUDA cores were used for calculations. While this approach did not lessen the computational load, it conserved memory bandwidth, a critical bottleneck in k-Wave simulations.

3 EXPERIMENTAL RESULTS

The experimental assessment of the proposed technique was conducted using three realistic simulation benchmarks released by the iTRUSST consortium [1]. These benchmarks assess the maximum acoustic pressure distribution within the brain induced by the bowl transducer. Different spatial and temporal resolutions were selected, as detailed in Table 2. The GPU memory consumption in the original simulation ranged from 1.8 GB to 11.3 GB.

3.1 Execution Acceleration

Given the high dynamic range of acoustic quantities and the inherent limitations of performing calculations in reduced precision, the overall reduction in execution time ranges between 3.27% and 8.54%, as detailed in Table 3. Notably, since Fourier transforms account for 50-60% of the total computation time, the observed acceleration is substantial. This performance improvement is consistent across various simulation sizes. Furthermore, detailed CUDA kernel profiling, outlined in Table 4, indicates that several kernels markedly benefit from operating in lower precision.

3.2 Accuracy Investigation

The impact of mixed precision calculations on accuracy was evaluated using four metrics. The first two, L^2 norm and L^∞ norm, assessed noise levels across the entire simulation domain. The latter two metrics focused on analyzing the pressure field within the focal region, see Table 5.

The low memory reduction level resulted in negligible errors across all three benchmarks, aligning with the 0.1-1% uncertainty

Table 4: Acceleration for particular CUDA kernels.

CUDA Kernel	Reduction Level		
	Low	Med	High
AddPressureSource	5.96%	102.37%	103.28%
ComputeAbsorptionTerm	10.20%	10.12%	10.14%
ComputeDensityLinear	7.34%	53.38%	52.99%
ComputePressureGradient	0.08%	0.21%	3.86%
ComputePressureTerms	9.05%	31.53%	31.51%
ComputeVelocityGradient	0.21%	0.22%	3.47%
ComputeVelocityUniform	0.28%	0.00%	16.16%
SumPressureTerms	28.51%	28.46%	28.47%

typically associated with material properties. The medium memory reduction yielded a maximum absolute error below 0.5% throughout the simulation domain. The high memory reduction exceeded the 1% threshold, reaching up to 3% error in the largest benchmark. As depicted in the comparison of pressure fields, where Fig. 1 illustrates the original code's pressure distribution and Fig. 2 shows the absolute error, the error manifests as random noise without altering the focus's shape or position. Notably, no movement of the focal point was detected in benchmarks, and the pressure amplitude difference remained under a 0.5% margin. These findings affirm the method's clinical applicability and robustness.

Table 5: Accuracy under various levels of memory reduction.

Dataset	Metric	Reduction Level		
		Low	Med	High
PH1-BM7-SC1	L^2 error	0.071%	0.415%	1.166%
	L^∞ error	0.081%	0.313%	1.004%
	Amplitude diff	0.072%	0.045%	0.065%
	Focus movement	0.0 mm	0.0 mm	0.0 mm
PH1-BM8-SC1	L^2 error	0.091%	0.519%	2.765%
	L^∞ error	0.080%	0.301%	1.014%
	Amplitude diff	0.071%	0.197%	0.264%
	Focus movement	0.0 mm	0.0 mm	0.0 mm
PH1-BM9-SC1	L^2 error	0.092%	0.731%	6.98%
	L^∞ error	0.106%	0.542%	2.941%
	Amplitude diff	0.070%	0.132%	0.530%
	Focus movement	0.0 mm	0.0 mm	0.0 mm

4 CONCLUSIONS

This study explored the feasibility of applying mixed precision calculations within the CUDA-accelerated k-Wave toolbox. Key quantities were identified for conversion into reduced precision data types, leading to the adaptation of CUDA computing kernels for the utilization of these data types and the execution of multiple operations concurrently where feasible.

Efficient representation of material properties and k-space operators resulted in up to 22% reduction in GPU memory usage for precision tasks and accelerated computations by a factor of 8.5% on an Nvidia Volta V100 graphic card. This reduction in memory usage allows for an increase in simulation resolution by 7.3%, enabling more precise focus targeting. The introduced optimization slightly decreased the focal pressure by 0.5%, without causing any shift in the focal point, an outcome within clinically acceptable margins.

Despite the performance improvement being under 9%, this enhancement significantly reduces the cost of simulations. Considering that a single neurostimulation procedure may target multiple locations, this reduction is deemed substantial.

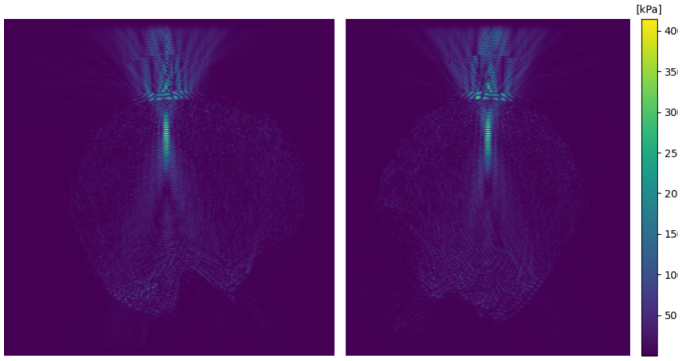

Figure 1: Maximum acoustic pressure distribution across the entire simulation domain, computed with full precision.

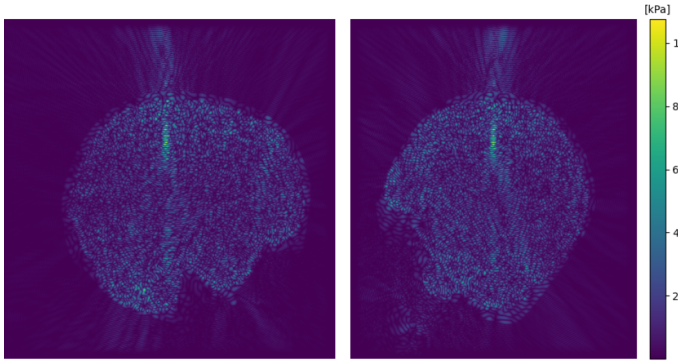

Figure 2: The absolute difference in acoustic pressure across the entire domain, quantified by the L^∞ norm, resulting from the high memory reduction.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Jean-Francois Aubry, Oscar Bates, Christian Boehm, Butts Kim Pauly, Douglas Christensen, et al. 2022. Benchmark problems for transcranial ultrasound simulation: Intercomparison of compressional wave models. *Journal of the Acoustical Society of America* 152, 2 (2022), 1003–1019. https://doi.org/10.1121/10.0013426

[2] Bradley E Treeby, Jiri Jaros, Alistair P Rendell, and B T Cox. 2012. Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method. *The Journal of the Acoustical Society of America* 131, 6 (2012), 4324–36. https://doi.org/10.1121/1.4712021