Towards Large-scale Ultrasound Simulations in Soft Tissue for Medical Applications

Filip Vaverka
4th year, full-time study
Jiri Jaros

Brno University of Technology, Faculty of Information Technology,
Centre of Excellence IT4Innovations,
Bozetechova 2, 612 00 Brno, Czech Republic
ivaverka@fit.vutbr.cz

***Abstract*—The synergy between advancements in medicine, ultrasound modeling, and high performance computing has lead to emergence of many new applications of biomedical ultrasound. These applications, ranging from HIFU treatment planning to (photo-)acoustic imaging, are requiring accurate large scale ultrasound simulations while putting significant pressure on the cost and time to solution.**

The presented article discusses progress in the development of our k-space pseudo-spectral fullwave non-linear ultrasound simulation code across an assortment of modern accelerated cluster architectures.

***Keywords*—k-Wave toolbox, k-Space method, high performance computing, GPGPU, Intel Xeon Phi, OpenMP, MPI, CUDA, OpenCL, ultrasound simulation, personalized medicine**

I. Introduction

Many emerging applications of ultrasound in medicine require large scale simulations of ultrasound wave propagation in soft tissue. Typical example of an ultrasound based treatment procedure is High Intensity Focused Ultrasound (HIFU) [1] used for noninvasive tumor removal by focusing mechanical energy of the ultrasound to a precise point associated with tissue heating. Another example is photo-acoustic imaging [2] where the tissue structure (image) is reconstructed from a recording of the ultrasound waves produced by rapid light energy absorption in the tissue.

Both of these applications require simulations in domains of up to $25\,\mathrm{cm} \times 25\,\mathrm{cm} \times 25\,\mathrm{cm}$ with frequencies in a MHz range and medium with sound speeds around $1500\,\mathrm{m/s}$. Additionally, in the case of HIFU, the frequencies up to $15\,\mathrm{MHz}$ have to be modeled due to nonlinearity of the wave propagation at high intensities. In both applications, the stress at the time to solution and the cost is significant as multiple simulations are needed for each patient (HIFU is an optimization problem while photo-acoustic reconstruction is an inverse problem).

My Ph.D. work raises to these challenges and sets up a list of following goals:

- Find a suitable mathematical model of the ultrasound wave propagation in soft tissue.
- Identify the most favorable supercomputer architectures.
- Further optimize the code for selected architectures (while accounting for future hardware).
- Generalize the developed techniques to broader a class of problems.

This paper describes the progress made towards the routine employment of large scale ultrasound simulations required for various medical procedures. The paper will first briefly describe challenges associated with rise of accelerated clusters. Second, novel domain decomposition method used for the formulation of pseudo-spectral time-domain methods suitable for these accelerated clusters will be detailed. Next, this technique will be compared with the traditional approach on a CPU-based cluster and results achieved across various architectures will be presented. Finally, practical benefits of our work will be shown and important conclusions drawn.

Challenges of Accelerated Clusters

Most modern HPC clusters heavily rely on acceleration or specialized architectures to attain their performance, and more importantly, power efficiency goals. Putting aside challenges associated with the efficient computation on accelerators (wide SIMD units, memory coherency and threading models, etc.), the shift towards accelerated computing presents new platform wide challenges. The applications architecture have to adapt to changes in communication and data management. Figure 1 shows a typical accelerated cluster. Note that each accelerator has its own memory, which is often represented as a separate address space. The ratio between the local memory bandwidth and inter-node interconnect bandwidth is getting significantly worse [3].

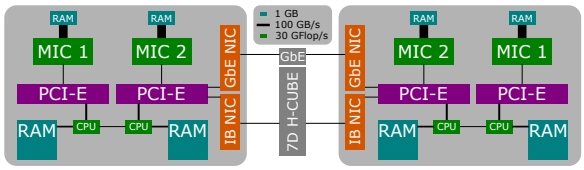


Figure 1: An example of accelerated cluster architecture.

II. Wave Propagation Model

Nonlinear wave propagation in heterogeneous absorption medium can be modeled by the following system of coupled first-order differential equations derived by Treeby [4]:

$$\frac{\partial \mathbf{u}}{\partial t} = -\frac{1}{\rho_0}\nabla p + \mathbf{F}, \qquad \text{(momentum)}$$

$$\frac{\partial \rho}{\partial t} = -\rho_0 \nabla \cdot \mathbf{u} - \mathbf{u} \cdot \nabla \rho_0 - 2\rho \nabla \cdot \mathbf{u} + \mathrm{M}, \qquad \text{(mass)}$$

$$p = c_0^2 \left(\rho + \mathbf{d} \cdot \nabla \rho_0 + \frac{B}{2A}\frac{\rho^2}{\rho_0} - \mathrm{L}\rho \right). \qquad \text{(eq. of state)}$$

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 density, c_0 is the isentropic sound speed, B/A is the nonlinearity parameter and $\mathrm{L}\rho$ is absorption/diffusion operator.

This system of equations can be easily discretized using simple, yet flexible Finite difference methods (FDM). The practical drawback of this approach is the need for a very fine spatial resolution to achieve acceptable accuracy, usually over 10 grid points per minimal wavelength (PPMW) [5]. Alternatively, higher order FDMs can be used, however, that leads to a loss of flexibility. Moreover, a very accurate time-stepping scheme (such as Runge-Kutta [6]) is required.

Fortunately, a fundamentally different approach based on pseudo-spectral time-domain (PSTD) method can be used. The PSTD method is able to accurately represent the solution down to 2-3 PPMW and thus make the simulations of interest tractable. However, the need of a complex time-stepping schemes remains to be solved.

To alleviate the time stepping difficulties, the k-space pseudo-spectral time-domain (KSTD) method [7] can be used. The advantage of KSTD method is its semi-analytical time-stepping scheme, which uses a simple finite-difference time stepping combined with an analytical correction included in the spatial discretization.

However, the correction in KSTD methods has a computational cost lying in the calculation of the spectral spatial derivative operator which is no longer separable and has to be computed using 3D DFTs as in

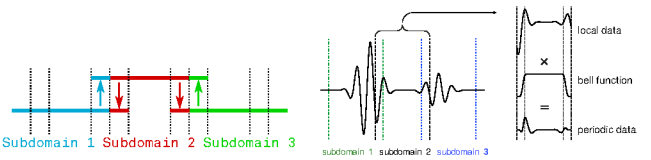
$$\frac{\partial}{\partial \xi} p^n = \mathcal{F}^{-1}\left\{ ik_\xi \kappa e^{ik_\xi \Delta\xi/2} \mathcal{F}\left\{p^n\right\}\right\}, \qquad (1)$$

where $\frac{\partial}{\partial \xi} p^n$ is a corrected spatial derivative of p^n in ξ direction, shifted by half a grid-point, and $\mathcal{F}\{\cdot\}$ is 3D DFT operator.

The 3D DFT (or FFT) is an expensive operation in distributed environment, where it includes global transposition on the simulation domain.

III. Local Fourier Basis Domain Decomposition

In our local decomposition approach, introduced in [8], we reduce the communication complexity of the derivative operator computed by P processing units from $\mathcal{O}(P^2)$ to $\mathcal{O}(26P)$. This reduction is achieved by the uniform decomposition of the simulation domain into subdomains and restricting the global fourier basis [9] used in KSTD method to each subdomain. The subdomains are then coupled with their neighbors (26 in the case of a full 3D decomposition) through the overlaps exchanged before every 3D DFT operation. Note that by restricting the Fourier basis to a local subdomain, the periodic boundary condition has to be enforced on each subdomain. This can be dealt with in many ways [10]. In our approach, we use overlaps together with a suitable taper function which fulfills the boundary conditions while having minimal impact on the solution itself (see Fig. 2).



(a) Overlap exchange (b) Local basis

Figure 2: The principle of the local Fourier basis domain decomposition in 1 dimension. (a) The local subdomain padded with periodically exchanged overlaps. (b) Restoration of local periodicity by multiplication of the data with a bell function.

The accuracy of this approach depends on the chosen decomposition and scales with the size of the overlaps and can be significantly improved by using an optimized taper or bell functions (as compared to typical erf$(\cdot)$ based bell function). In a typical situation where an accuracy of $L_\infty < 10^{-3}$ is required, the overlap depth of 8 to 16 grid points is sufficient. These properties of the approach are further explored in [8].

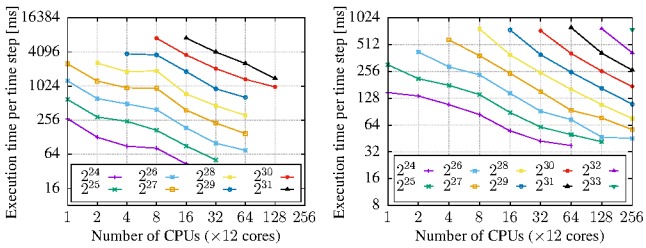


Figure 3: Scaling of global (left) and local (right) decomposition method on the Salomon cluster with domains of $256 \times 256 \times 256$ to $2048 \times 2048 \times 2048$ grid points.

Figure 3 shows that on a typical CPU based cluster speedups between 2 and 5 times can be achieved by switching from the global to the local approach with overlaps of 16 grid points. Both methods achieve similar scaling factor between 1.5 and 2 with the global method scaling slightly better for smaller simulations on less than 8 CPU sockets. These improvements come primarily from a 50% reduction of the execution time spent in the communication routines.

IV. Solver Performance Evaluation

This section shows how our simulation code performs across a wide variety of accelerated and many-core clusters. We

will compare the CoolMUC3 cluster based on Intel Xeon Phi Knight's Landing (KNL) many core architecture, the Salomon cluster accelerated by Intel Xeon Phi Knight's Corner (KNC), Piz Daint accelerated with Nvidia P100 GPUs, and conclude with a few remarks on clusters with dense multi-GPU nodes.

Many-Core Architectures

The main appeal of many-core architectures in the form of the Intel Many Integrated Cores (MIC) architectures is their apparent similarity to traditional multi-core processors. In theory, this means that the code optimized for CPUs should scale nicely on MICs without a need for significant changes. Therefore, a speedup of around 4× when using Salomons KNC accelerators, and 7 to 10× on CoolMUC3s KNLs in one to one comparison to the Salomons 12-core CPU is expected.

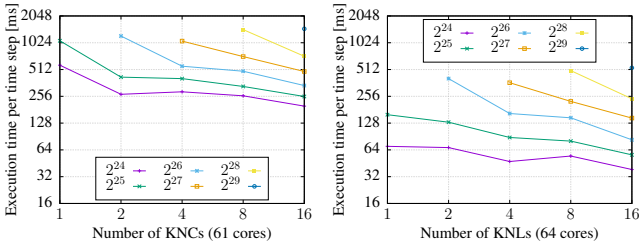


Figure 4: Strong scaling on KNC and KNL clusters.

In practice, we observed that the KNCs are 2.2 to 4.3× slower than the CPU baseline (see Fig. 4). Although the computation on KNCs involves some additional intra-node communication (accelerators are connected via PCI-E 2.0 ×16), this adds only about 50% of the total communication time. Most of the performance is lost in local 3D FFT computations which are known to perform very poorly on this MIC architecture. KNC achieves no more than 50% of 12-code CPUs performance in this task. Our investigation points to problems with the memory hierarchy and the coherency protocols of KNC architecture.

Intel Xeon Phi KNL achieves a speedup of 1.7× on average compared to CPUs. This result can be explained by significant improvements of on-chip interconnect (a grid based topology) and an increase in the memory bandwidth. It should also be noted that the CoolMUC3 cluster uses an OmniPath interconnect which is directly accessible by each KNL.

Despite these results, the scaling factor on both clusters remains above 1.5, which confirms the propsed communication strategy to work as expected. A detailed analysis of the code behavior on Intel MIC platforms was published in [11].

GPU Accelerated Clusters

GPUs are perhaps the most proliferate form of acceleration in modern HPC clusters. Their high parallel efficiency stems from a very different memory coherency and programming models. Although this may make the migration of some algorithms difficult, it is not the case for our simulation code which is data intensive and has a regular workload.

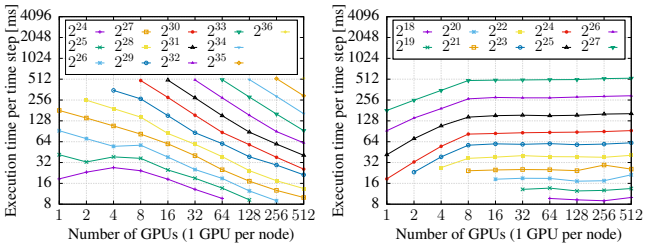


Figure 5: Strong and weak scaling on the Piz Daint cluster.

To evaluate the performance and scaling on a GPU accelerated cluster, the Piz Daint cluster located in Switzerland was employed. Piz Daint comprises of 5704 compute nodes, each of which equipped with a single NVIDIA P100 GPU accelerator. Comparing one accelerator to a single 12-core CPU used in Salomon yields an 8× speedup for our simulation workload. Figure 5 shows the scaling on the Piz Daint cluster reaches a scaling factor of 1.5, and an average speedup of 4.8 over the Salomon CPUs (in comparison on socket/GPU level). The key components enabling these results are NVIDIA P100 GPUs in combination with Piz Daints interconnect based on the Aries ASIC in a Dragonfly network topology [12], which provides excellent connection to neighboring nodes.

Multi-GPU Dense Nodes

Recently, the supercomputer architectures begin a transition to multi-GPU compute nodes with a special high-bandwidth interconnects between GPUs. This brings new challenges to efficient intra-node GPU-to-GPU communication which has to be addressed as well. In practice, this means the elimination of the communication through the CPU (such as message passing) and transition to direct communications between GPUs (such as CUDA Peer-to-Peer transfers).

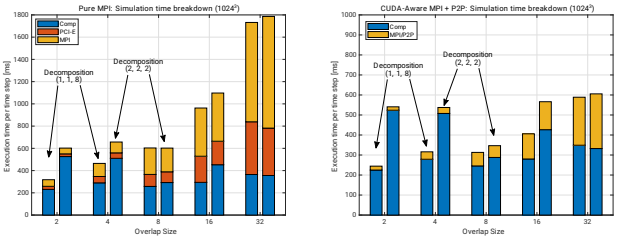


Figure 6: Message passing versus CUDA P2P communication (PCI-E 3.0 based server with 8 Tesla P40 GPUs.)

In this area, we first optimized inter-GPU communications in a traditional dual-socket server with 8 Nvidia Tesla P40 GPUs attached by PCI-E 3.0 ×16. Figure 6 shows up to 3.6× speedup achieved by using direct P2P communication between GPUs without involving CPU.

Finally, to estimate the potential of dense nodes with fast P2P interconnects between accelerators, a series of experiments on an Nvidia DGX-2 server was performed. DGX-2 is a dual socket machine with 24 cores per socket and

1.5 TB of system memory. The machine is accelerated by 16 Nvidia V100 GPUs. These GPUs are interconnected via Nvidia NVlink 2 being able to achieve 300 GB/s bidirectional bandwidth between two GPUs, and 1.98 TB/s bisection bandwidth across all 16 GPUs.

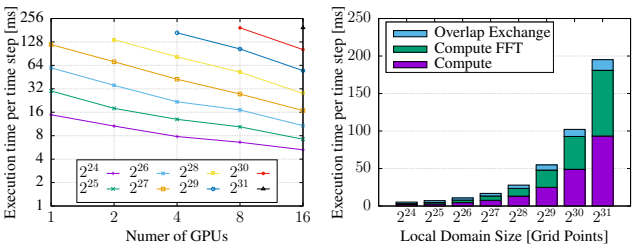


Figure 7: Nvidia DGX-2: Simulation scaling and breakdown.

The results (Fig. 7) indeed show almost 10× faster inter-GPU interconnect yielding a massive reduction in time spent on the overlap exchanges. An average scaling factor is 1.6, which is mostly due to additional computational work and not the communication. The comparison with the results from Piz Daint cluster (Fig. 5) shows that we achieved about 2× speedup when all 16 GPUs are in use. Please note that a single V100 is not much faster than a single P100. The speedup is therefore mostly caused by a superior interconnect in DGX-2.

V. Conclusion

My Ph.D. work has shown that the proposed ultrasound wave propagation model optimized on modern clusters can reach up to 5× speedup on the same hardware, and up to 50× when a GPU accelerated cluster is available. Further, I have shown that the solution is scalable to clusters with hundreds of GPUs, and is ready for recently introduced clusters with multi-GPU compute nodes. Finally, the benefits of ultra-dense GPU accelerated compute servers with high-bandwidth GPU-to-GPU interconnect have been demonstrated.

Let us consider a real-world photo-acoustic tomography image reconstruction on a domain of 20 cm × 20 cm × 20 cm with the maximum frequency of 2 MHz requiring 50 simulations. Each simulation needs over 5500 time-steps on a domain of 1024^3 grid points. My work proves that it is possible to reduce the computation time from 88 hours on 32 dual-socket nodes of Salomon to 35 hours on a single 8 GPU server equipped with PCI-E 3.0. More significantly, the computational cost can be reduced by a factor of 4 at the same time. Our solution can further reduce the reconstruction time down to about 8 hours while maintaining the price when a DGX-2 server is used. This is a stunning breakthrough for photoacoustic imaging since four such machines can cover the computing needs of a medium-size breast cancer medical centre.

Achievement Summary

To achieve these results, a novel approach to the decomposition of PSTD methods has been introduced. The numerical performance of this approach has been improved by optimization of bell functions used to restore the local Fourier basis. Further, the simulation code and communication patterns have been optimized for various accelerated cluster architectures achieving scaling to 6144 CPU cores, 256 Intel Xeon Phi (KNC) accelerators or 1024 GPUs.

Future Work

The presented results can be adapted to other PSTD codes such as the elastic wave propagation simulation. It also allows for novel approaches of model coupling in hybrid fluid-elastic models and models with spatially varying resolution.

Acknowledgment

This work was supported by the FIT-S-17-3994 Advanced parallel and embedded computer systems project. This work was also supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602" and by the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center - LM2015070". This project has received funding from the European Union's Horizon 2020 research and innovation programme H2020 ICT 2016-2017 under grant agreement No 732411 and is an initiative of the Photonics Public Private Partnership.

References

[1] T. J. Dubinsky, C. Cuevas, M. K. Dighe, O. Kolokythas, and H. H. Joo, "High-intensity focused ultrasound: Current potential and oncologic applications," *American Journal of Roentgenology*, vol. 190, no. 1, pp. 191–199, jan 2008.

[2] P. Beard, "Biomedical photoacoustic imaging," *Interface Focus*, vol. 1, no. 4, pp. 602–631, aug 2011.

[3] A. Gholami, J. Hill, D. Malhotra, and G. Biros, "AccFFT: A library for distributed-memory FFT on CPU and GPU architectures," May 2016.

[4] B. E. Treeby, J. Jaros, A. P. Rendell, and B. T. Cox, "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*, vol. 131, no. 6, pp. 4324–36, 2012.

[5] T. D. Mast, L. P. Souriau, D.-D. Liu, M. Tabei, A. I. Nachman, and R. C. Waag, "A k-space method for large-scale models of wave propagation in tissue," vol. 48, no. 2, pp. 341–354.

[6] E. Fehlberg, "Low-order classical runge-kutta formulas with step-size control and their application to some heat-transfer problems," *Computing*, 01 1969.

[7] M. Tabei, T. D. Mast, and R. C. Waag, "A k-space method for coupled first-order acoustic propagation equations." *The Journal of the Acoustical Society of America*, vol. 111, no. 1 Pt 1, pp. 53–63, jan 2002.

[8] J. Jaros, F. Vaverka, and B. E. Treeby, "Spectral domain decomposition using local fourier basis: Application to ultrasound simulation on a cluster of GPUs," vol. 3, no. 3.

[9] M. Israeli, L. Vozovoi, and A. Averbuch, "Spectral multidomain technique with local Fourier basis," *J. Sci. Comput.*, vol. 8, no. 2, pp. 135–149, 1993.

[10] J. P. Boyd, "A Comparison of Numerical Algorithms for Fourier Extension of the First, Second, and Third Kinds," *Journal of Computational Physics*, vol. 178, no. 1, pp. 118–160, may 2002.

[11] F. Vaverka, B. E. Treeby, and J. Jaros, "Evaluation of the suitability of intel xeon phi clusters for the simulation of ultrasound wave propagation using pseudospectral methods," in *Computational Science – ICCS 2019*. Springer International Publishing, vol. 11538, pp. 577–590.

[12] B. Alverson, E. Froese, L. Kaplan, and D. Roweth, "Cray XC Series Network," 2012. [Online]. Available: https://www.cray.com/sites/default/files/resources/CrayXCNetwork.pdf