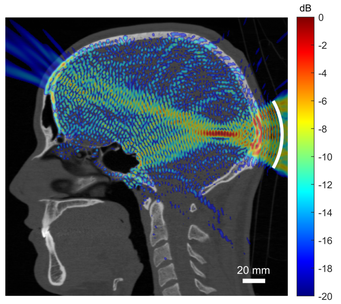
Advancements in Ultrasound Simulations Enabled by High-bandwidth GPU Interconnects

Filip Vaverka[1], Bradley E. Treeby[2] and Jiri Jaros[1]

[1]Faculty of Information Technology, Brno University of Technology, Centre of Excellence IT4Innovations, CZ
[2]Department of Medial Physics and Biomedical Engineering, University College London, UK

Overview

Transcranial ultrasound therapy is a rapidly emerging technology used to treat major brain disorders. The key challenge is to ensure the ultrasound energy is delivered to a precise location identified by a clinician. This is difficult because the skull is very rigid and causes reflections and distortions of the ultrasound waves. These effects may be predicted and corrected for by the use of complex numerical models of the ultrasound waves propagation in the body (see figure). Unfortunately, these models can take many hours or days to run even on large supercomputers. The reduction of the compute time is thus critical for clinical workflows.



Ultrasound Wave Propagation in Tissue

The governing equations modeling the ultrasound wave propagation in heterogeneous absorbing tissues can be written as follows:

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

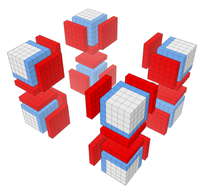
$$\frac{\partial \rho}{\partial t} = -(2\rho + \rho_0)\nabla \cdot \mathbf{u} - \mathbf{u} \cdot \nabla \rho_0 + S_M \qquad \text{(mass conservation)}$$

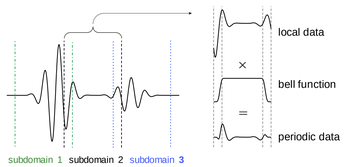
$$p = c_0^2\left(\rho + \mathbf{d} \cdot \nabla \rho_0 + \frac{B}{2A}\frac{\rho^2}{\rho_0} - L\rho\right) \qquad \text{(pressure-density relation)}$$

These equations are discretized using the k-space pseudospectral approach which achieves excellent convergence and low dispersion, but requires multiple evaluations of 3D Fourier transforms (3D FFTs) per time step.

Global and Local Gradient Operators

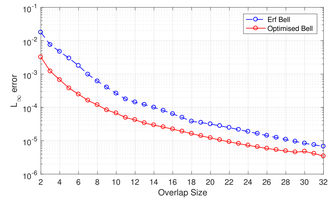
The k-space gradient operator is traditionally implemented globally by means of distributed 3D FFTs. The local version partitions the simulation domain into overlapping subdomains and computes the 3D FFTs locally. These two approaches offer a trade-off between the amount of communication and local computation. This poster investigates how Nvidia's NVlink high-bandwidth GPU interconnect influences this trade-off compared to conventional PCI-Express based architectures.

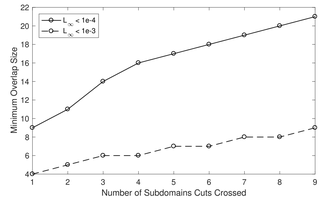




Simulation Accuracy

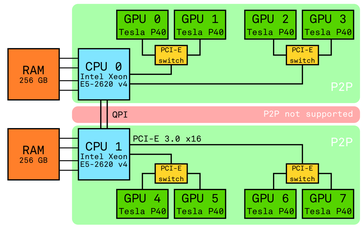
When the gradient is calculated by the local operator, numerical error is introduced. The error level can be controlled by the shape of the bell function and the size of the overlap region.

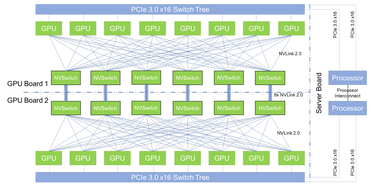




Architectures of Dense Multi-GPU Systems

Fast interconnections such as NVlink 2.0 in Nvidia DGX-2 (right) make communication intensive multi-GPU algorithms including distributed 3D FFTs feasible. These algorithms have been very limited by the PCI-Express 3.0 interconnection in multi-GPU servers such as PNY (left).

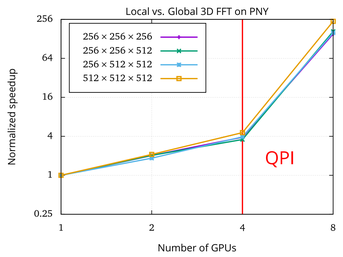


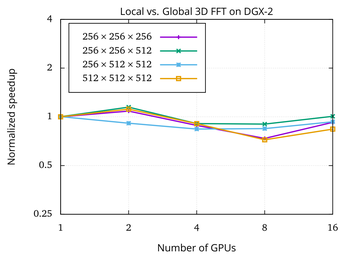


The all-to-all NVlink network does also not suffer from the communication bottleneck introduced by the QPI links between CPU sockets.

Performance of the Gradient Operator

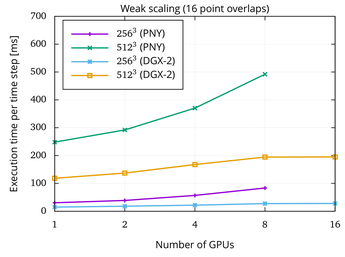
The local gradient operator shows a massive speedup on the PCI-Express machine (left), while being often overcome by the global variant on the NVlink machine (right). This points towards hybrid decompositions with the global operator within multi-GPU nodes and the local one among them.

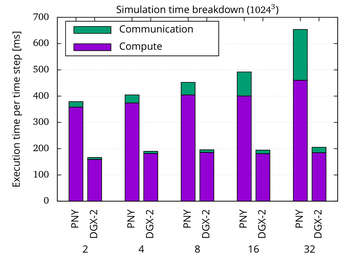




Performance and Scaling

The transition from P40 in the PNY machine to V100 GPUs in the DGX-2 machine results in a speedup of 2. Since the maximum attained speedup by DGX-2 reaches almost 4 when 8 GPUs and 32 grid point overlaps are used, the NVlink must bring another factor of two.

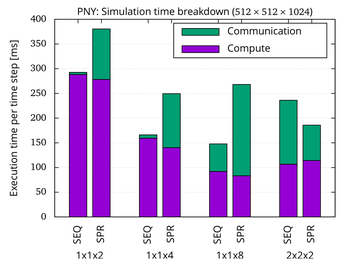


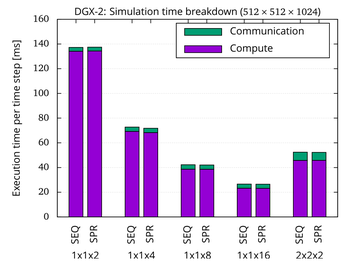


The left figure shows weak scaling for two fixed subdomain sizes and 16 point overlaps. Note the desired flattening between 8 and 16 subdomains when the full 3D decomposition rank is achieved. On the right, the execution time breakdown for a 1024^3 grid point simulation across 8 GPUs shows up to 10 times reduction in the communication overhead due to NVlink.

Domain Decomposition and Mapping

Multi-socket PCI-Express based machines are very sensitive to the mapping between subdomains and GPUs, which is caused by the QPI links between sockets hindering GPU-to-GPU communication. In contrast, DGX-2 is not affected by the changes in mapping at all.





The sequential mapping (SEQ) maps neighboring subdomains to the closest GPUs while the spread mapping (SPR) maps them to GPUs across QPI or the GPU board boundary.

Impact and Outlook

A typical simulation needed in transcranial ultrasound therapy covers $30\,\mathrm{cm} \times 30\,\mathrm{cm} \times 30\,\mathrm{cm}$ with the maximum frequency of $1\,\mathrm{MHz}$. This translates into a simulation over 1200^3 grid points and 7200 time steps. Such a simulation can be computed within $30\,\mathrm{min}$ using either 128 dual-socket 12-core Haswell CPU nodes or a single DGX-2 GPU server.

This work was 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. This work was also supported by the Engineering and Physical Sciences Research Council, UK, grant numbers EP/L020262/1 and EP/P008860/1.