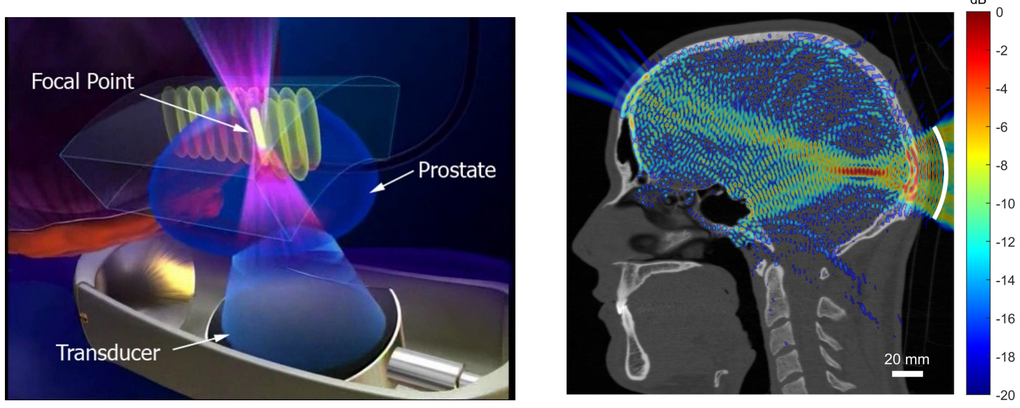
Performance of k-Wave Toolbox on Multi-GPU Accelerated Clusters

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

[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

The k-Wave toolbox is among the most widely used ultrasound propagation simulation tools in medical applications. These applications include both diagnostics methods such as photoacoutic imaging (PAI) and non-invasive treatment methods such as high-intensity focused ultrasound (HIFU) or acoustic pressure brain stimulation.



All these applications translate into inverse problems (transducer array configuration for HIFU and initial pressure distributions in PAI), whose solutions typically require many evaluations of ultrasound wave propagation through the tissue. The HIFU methods pose additional challenge in the form of non-linearity which has to be captured by the simulation, which leads to high resolution simulations. In clinical settings the need for fast and cheap simulations in these methods is amplified as part of personalized medicine.

Ultrasound Wave Propagation in Tissue

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

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

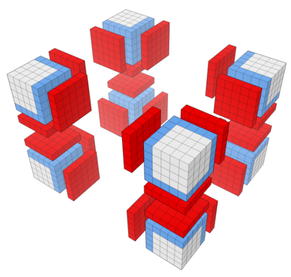
$$\frac{\partial \rho}{\partial t} = -(2\rho + \rho_0)\nabla \cdot \mathbf{u} - \mathbf{u} \cdot \nabla \rho_0 + S_M \qquad \text{(mass conservation)}$$

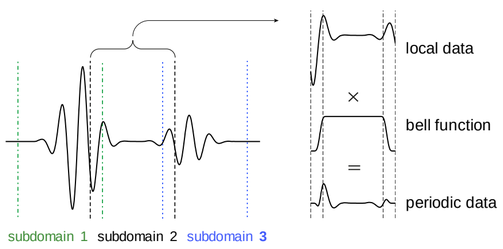
$$p = c_0^2\left(\rho + \mathbf{d} \cdot \nabla \rho_0 + \frac{B}{2A}\frac{\rho^2}{\rho_0} - L\rho\right) \qquad \text{(pressure-density relation)}$$

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

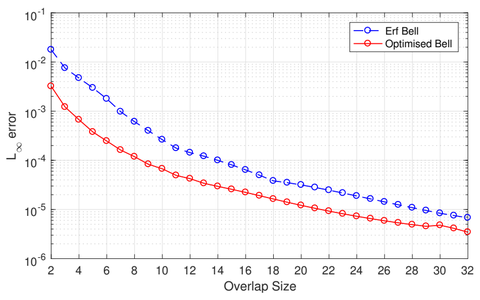
The k-space gradient operator is traditionally implemented globally by means of distributed 3D FFTs. The local version partitions the simulation domain into overlapping subdomains and computes the 3D FFTs locally. These two approaches offer a trade-off between the amount of communication and local computation. In this poster we compare performance of both approaches across range of cluster architectures.

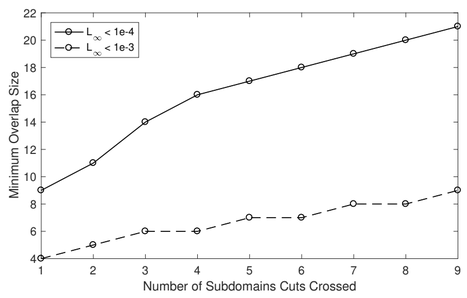




Simulation Accuracy

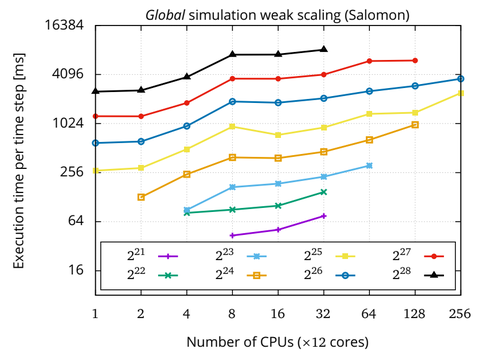
When the gradient is calculated by the local operator, numerical error is introduced. The error level can be controlled by the shape of the bell function and the size of the overlap region.

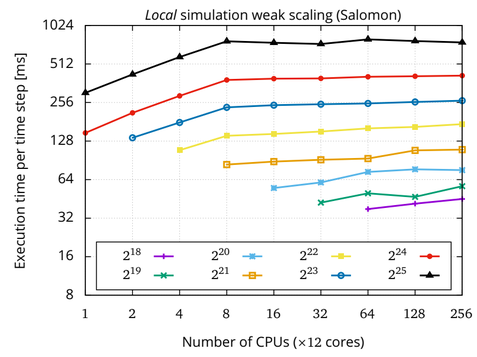




Global and Local Method on Salomon Cluster (MPI + OpenMP)

Let us consider a traditional CPU based cluster with compute nodes integrating two Intel Xeon E5-2680v3 CPUs connected in a 7D hypercube topology with FDR Infiniband interconnect.

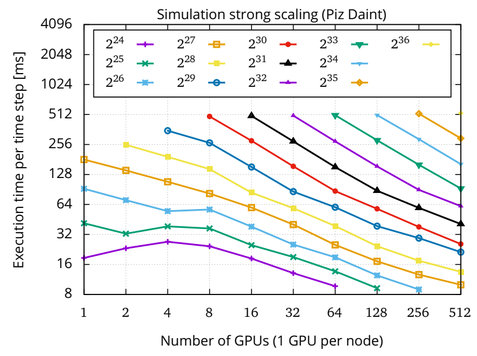


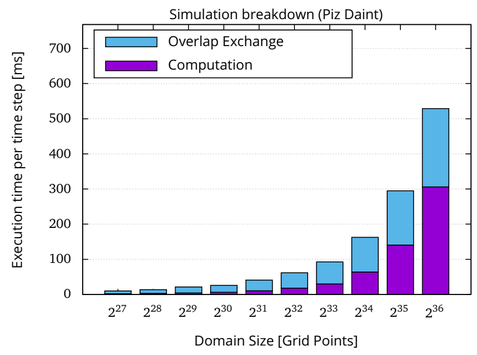


The transition from the global (left) to the local (right) KSTD method offers up to 5× speedup for the same amount of CPUs. This is, in significant part, caused by the reduction in the order of the communication complexity because all-to-all communication is reduced to nearest neighbors communication.

Piz Daint: GPU Accelerated Cluster (MPI + CUDA)

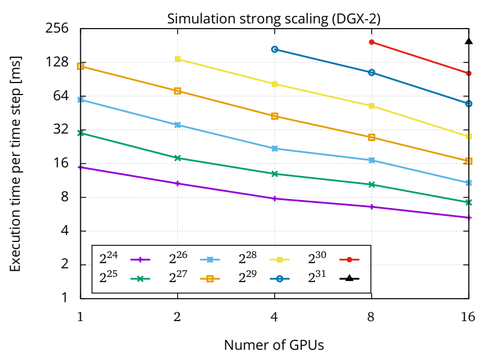
Each GPU accelerated node of Piz Daint hosts single Nvidia P100 GPU with 16 GB of HBM 2 memory. Nodes are interconnected by a DragonFly network running on the Cray Aries technology.

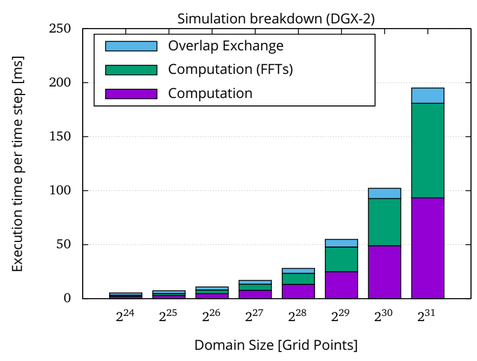




Nvidia DGX-2: Multi-GPU Dense Node (MPI + CUDA + NVlink)

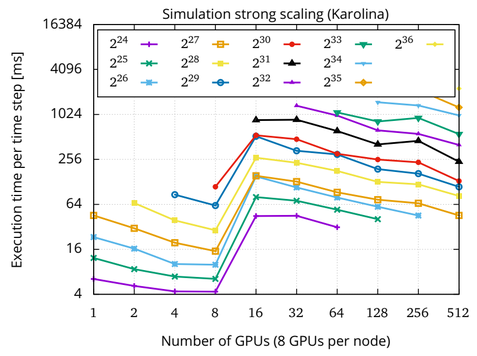
Nvidia DGX-2 compute node contains 16 Nvidia V100 GPUs with 32 GB of HBM 2 memory at 900 GB/s each. All 16 GPUs are connected together via NVlink 2.0, which provides 300 GB/s of connectivity to each GPU and bisection bandwidth of 2.4 TB/s.

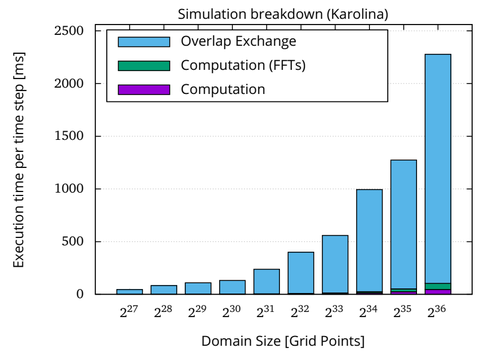




Karolina: Multi-GPU Accelerated Cluster (MPI + CUDA + NVlink)

Each GPU accelerated compute node of Karolina cluster consists of 8 Nvidia A100 GPUs with 40 GB of HBM 2 memory. The strong scaling plot (left) shows strong performance of each node (2 to 4× faster than 8 GPUs of DGX-2). However the inter-node communication bears surprisingly high overhead (considering 4× 200 Gbit IB ports per node), which is confirmed by simulation time breakdown (right).





This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140). This work was supported by Brno University of Technology under project number FIT-S-20-6309.