Advancements in Ultrasound Simulations Enabled by High-bandwidth GPU Interconnects

Filip Vaverka
Brno University of Technology,
Faculty of Information Technology,
Centre of Excellence IT4Innovations
Brno, Czech Republic
ivaverka@fit.vutbr.cz

Bradley E. Treeby
University College London, Medical
Physics and Biomedical Engineering,
Biomedical Ultrasound Group
London., United Kingdom.
b.treeby@ucl.ac.uk

Jiri Jaros
Brno University of Technology,
Faculty of Information Technology,
Centre of Excellence IT4Innovations
Brno, Czech Republic
jarosjir@fit.vutbr.cz

KEYWORDS

Ultrasound simulations, Local Fourier basis decomposition, k-Wave toolbox, Multi-GPU systems, CUDA, NVlink, MPI.

ACM Reference Format:
Filip Vaverka, Bradley E. Treeby, and Jiri Jaros. 2019. Advancements in Ultrasound Simulations Enabled by High-bandwidth GPU Interconnects. In *Proceedings of Supercomputing Conference (SC'19).* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Realistic ultrasound simulations are becoming integral part of many novel medical procedures ranging from preoperative ultrasound and photoacoustic screening [6, 8] to non-invasive treatment planning such as brain stimulation [4] or ultrasound ablation [1, 10]. The common denominator of all these applications is the need for cheap, fast and relatively large-scale ultrasound simulations with sufficient accuracy. A typical medical application requires a full-wave simulation which taking into account frequency-dependent absorption and non-linearity. Appication of the k-space pseudo-spectral approach (which is regarded to be one of best methods) leads to a system of partial differential equations solved over domains with more than 1024^3 grid points for tens of thousands of time steps. This task would traditionally be solved on a CPU-based cluster since these codes are typically memory and communication heavy. However, at SC 2017 we presented a GPU accelerated simulation code running on a cluster of 512 single-GPU nodes of the Piz Daint supercomputer. Last year at SC 2018, we investigated how our code can benefit from the use of CUDA peer-to-peer (P2P) communications [5] in multi-GPU compute nodes based on PCI-Express interconnect. P2P communications were found to provide significant speedup, which is hindered primarily by GPUs being split between multiple CPU sockets.

In this paper, we investigate the benefits of a high-bandwidth low-latency NVlink interconnect in an Nvidia DGX-2 super-dense multi-GPU server in comparison to a more traditional PCI-E 3.0 based multi-GPU server. Nvidia DGX-2 is a dual socket server based on Intel Xeon Platinum 8168 processors with 2×768 GB of main memory, and 16 Nvidia Tesla V100 Volta GPUs, each of which with 5120 CUDA cores and 32 GB of HBM memory. Our second system is a dual socket PNY server equipped with Intel E5-2620v4 CPUs, 2×256 GB or main memory, and 8 Nvidia Tesla P40 Pascal GPUs, each of which with 3840 CUDA cores and 24 GB of GDDR5X memory. The important distinction between these two systems is the inter-GPU interconnection. DGX-2 uses an NVlink 2.0 which offers 300 GB/s bi-directional bandwidth between two GPUs and 2.4 TB/s bisection bandwidth (all-to-all). The PCI-E 3.0 based system has GPUs grouped into pairs connected through PCI-E hubs, each of which connected to the root hub in one of the CPUs. The sockets are connected together via an Intel QPI providing up to 64 GB/s of bandwidth while each link in the PCI-Express structure has 16 PCI-E 3.0 lanes with up to 32 GB/s.

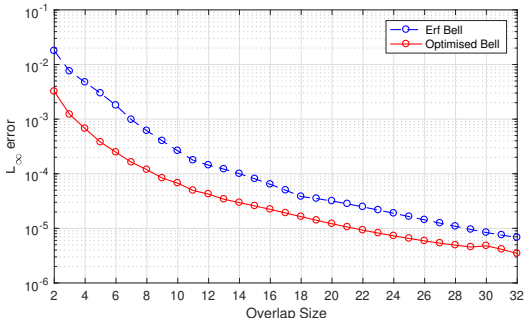

Figure 1: Numerical error introduced by a single interface in the local Fourier basis approach.

2 METHOD AND RESULTS

The key component for an efficient multi-GPU acceleration of the pseudo-spectral simulation codes is the minimization of the communication between GPUs. In the case of the k-Wave acoustic toolbox, this can be achieved by a restriction of the global Fourier basis resulting in a domain decomposition [2] where the 3D simulation domain is partitioned into blocks, each of which assigned to a single GPU [3]. Each partition is expanded to overlap with its neighbors by a specified amount of grid points, and these overlaps are then periodically exchanged during each time step of the simulation. The domain is treated as periodic, thus the neighbor communications form a 1-3 dimensional torus (depending on the rank of the decomposition). The size of the overlaps is a primary determining factor to both the accuracy and the performance of the simulation. For the aforementioned medical applications, a numerical error on the order of 10^{-3} is usually acceptable (see Fig. 1).

The computation of each subdomain is assigned to a single GPU and executed as a mix of CUDA FFT library [9] and custom CUDA kernel calls. The CPU is dedicated to I/O, management, GPU control

SC'19, 2019, Denver, Colorado, USA
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

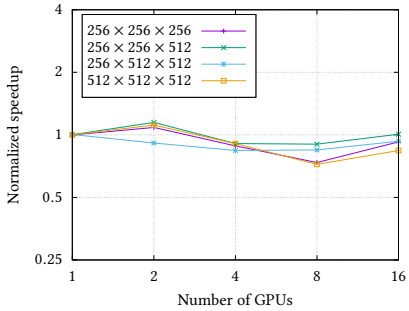

Figure 2: The speedup of the local gradient operator over the global on a DGX-2. The speedup is normalized to a single GPU, and overlaps of 16 points are used.

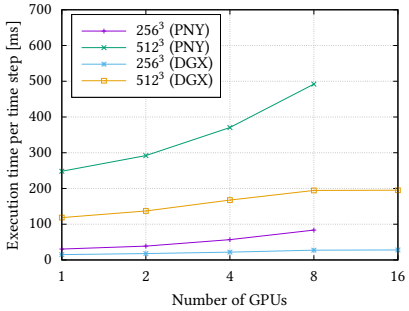

Figure 3: Weak scaling of the local approach using 16 point overlaps.

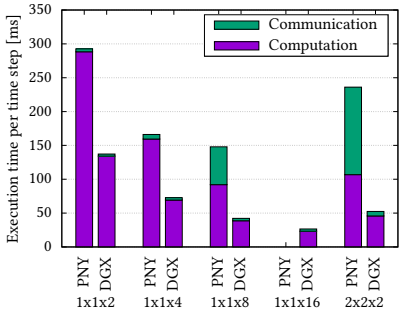

Figure 4: The impact of the communication through the QPI in 512×512×1024 grid point simulation with 16 point overlaps in various decompositions on both machines.

and communication tasks. The data transfers are realized by means of CUDA Peer-to-Peer (P2P) routines where possible, with a fallback to CUDA-Aware MPI [7]. The synchronization is always handled on the CPU side by MPI. The particular overlaps are packaged by CUDA kernels and transferred either by the GPU itself (P2P) or the CPU (MPI) to the target. This approach was successful on clusters with a single GPU per node such as Piz Daint[1] and achieved almost linear scaling.

The same approach yields very good results when used on a PCI-E based multi-GPU server. However, it is not clear, whether this approach is still optimal when a high-bandwidth interconnect is available since the computation and memory overhead may become larger than the communication penalty.

The only part of the simulation which involves communication between GPUs, is the pseudo-spectral gradient operator. Its behavior can easily be approximated by a pair of 3D discrete Fourier Transforms (forward and back). The rest of the simulation is essentially local without the need for inter-GPU communication. In Fig. 2, such an estimation is used to show that a DGX-2 manages to keep the performance of both approaches within a 30% margin. The estimated speedup of the local to the global approach has to be normalized so that both approaches are equal in a single GPU scenario. The normalization coefficient is computed for each domain size independently. Such a convoluted approach is necessary due to the issues with multi-GPU implementation of Real-to-Complex 3D DFT in the CUDA FFT library (Complex-to-Complex variant is used instead). This shows that while the global approach can be used in the future to avoid accuracy issues, it is still beneficial to continue with the local approach.

A single V100 is about 2.1× faster than a P40 in our workload (mostly due to an increase in the memory bandwidth). However, as a system, DGX-2 server achieves 2-4× speedup over the PNY server while using the same number of GPUs. The additional speedup comes from the NVlink 2.0 interconnect implemented in DGX-2, which is almost 10× faster than PCI-E 3.0 ×16 used in the PNY server. Figure 3 shows weak scaling on both machines. Although the scaling is rather poor on both machines due to increasing rank of the decomposition, it can be seen that this behavior is more severe on the PNY server and that the scaling is very good beyond 8 GPUs (where full 3D decomposition is reached). Finally, Fig. 4 illustrates the importance of the decomposition and the necessity to avoid unnecessary communication through QPI, while the NVlink 2.0 interconnect in DGX-2 is immune to these issues.

[1]Up to 512 Nvidia P100 GPUs, CSCS, CH

3 CONCLUSIONS

The main contribution of this paper is the experimental evaluation of the advantages offered by high-bandwidth interconnects such as NVlink in the area of pseudo-spectral simulations. It has been shown that approaches such as global Fourier transforms may become feasible, and can displace local Fourier basis decomposition at a multi-GPU node level. This can enable further scaling (when limited by accuracy) and alleviate some of the memory overhead associated with the local domain overlaps.

The results obtained for various domain decompositions and subdomain mappings to GPUs also show that the direct communication between GPUs in PCI-E based machines may severely suffer if the GPUs are distributed over multiple CPU sockets. The tree structure of the PCI-E may also become somewhat an issue as the number of GPUs per node grows.

Finally, when using only 8 Tesla V100 GPUs, our simulation code achieved 3× speedup over 8 Tesla P40 GPUs in the PCI-E based server, showing the importance of fast communication between GPUs. The speedup is then doubled as expected when all 16 GPUs in DGX-2 are used, and a sufficiently large simulation is considered.

4 ACKNOWLEDGMENT

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.

REFERENCES

[1] Osama Al-Bataineh, Jürgen Jenne, and Peter Huber. 2012. Clinical and future applications of high intensity focused ultrasound in cancer. *Cancer Treatment Reviews* 38, 5 (2012), 346–353. https://doi.org/10.1016/j.ctrv.2011.08.004

[2] M. Israeli, L. Vozovoi, and A. Averbuch. 1993. Spectral multidomain technique with Local Fourier Basis. *Journal of Scientific Computing* 8, 2 (jun 1993), 135–149. https://doi.org/10.1007/BF01060869

[3] Jiri Jaros, Filip Vaverka, and Bradley E Treeby. 2016. Spectral Domain Decomposition Using Local Fourier Basis: Application to Ultrasound Simulation on a Cluster of GPUs. *Supercomputing Frontiers and Innovations* 3, 3 (nov 2016), 39–54. https://doi.org/10.14529/jsfi160305

[4] James L. B. Robertson, Ben T. Cox, J. Jaros, and Bradley E. Treeby. 2017. Accurate simulation of transcranial ultrasound propagation for ultrasonic neuromodulation and stimulation. *The Journal of the Acoustical Society of America* 141, 3 (2017), 1726–1738. https://doi.org/10.1121/1.4976339 arXiv:https://doi.org/10.1121/1.4976339

[5] Cook Shane. 2013. *CUDA Programming*. Elsevier. https://doi.org/10.1016/C2011-0-00029-7 arXiv:arXiv:1011.1669v3

[6] Bradley E. Treeby, Jiri Jaros, and Ben T. Cox. 2016. Advanced photoacoustic image reconstruction using the k-Wave toolbox. In *SPIE Photons Plus Ultrasound: Imaging and Sensing*, Alexander A. Oraevsky and Lihong V. Wang (Eds.), Vol. 9708. 97082P. https://doi.org/10.1117/12.2209254

[7] Hao Wang, Sreeram Potluri, Devendar Bureddy, Carlos Rosales, and Dhabaleswar K. Panda. 2014. GPU-aware MPI on RDMA-enabled clusters: Design, implementation and evaluation. *IEEE Transactions on Parallel and Distributed Systems* 25, 10 (2014), 2595–2605. https://doi.org/10.1109/TPDS.2013.222

[8] Sheng Wang, Jing Lin, Tianfu Wang, Xiaoyuan Chen, and Peng Huang. 2016. Recent advances in photoacoustic imaging for deep-tissue biomedical applications. *Theranostics* 6, 13 (2016), 2394–2413. https://doi.org/10.7150/thno.16715

[9] Xueqin Zhang, Kai Shen, Chengguang Xu, and Kaifang Wang. 2013. Design and Implementation of Parallel FFT on CUDA. In *2013 IEEE 11th International Conference on Dependable, Autonomic and Secure Computing*. IEEE, 583–589. https://doi.org/10.1109/DASC.2013.130

[10] Yu-Feng Zhou, Ali Syed Arbab, and Ronald Xiaorong Xu. 2011. High intensity focused ultrasound in clinical tumor ablation. *World journal of clinical oncology* 2, 1 (2011), 8–27. https://doi.org/10.5306/wjco.v2.i1.8