From Pascal to Blackwell: Benchmarking Multi-GPU k-Wave with a Global FFT Solver

Oliver Kunik, Jiri Jaros
Faculty of Information Technology, Brno University of Technology,
Email: {ikunik, jarosjir}@fit.vut.cz

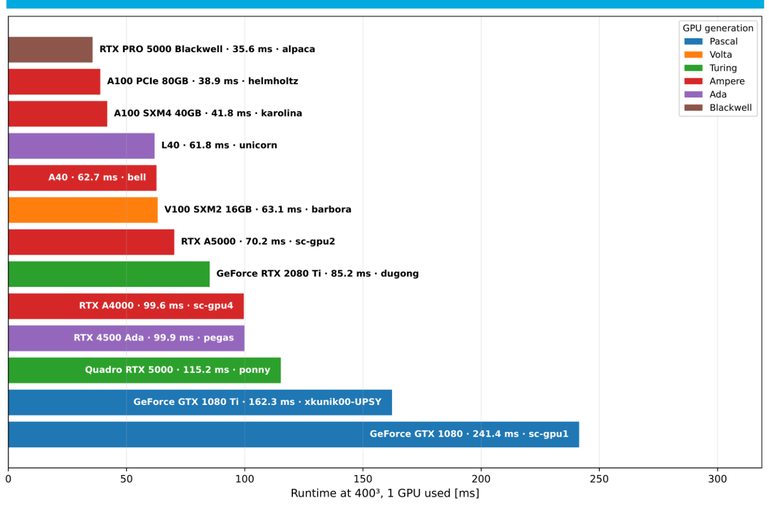
BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Introduction

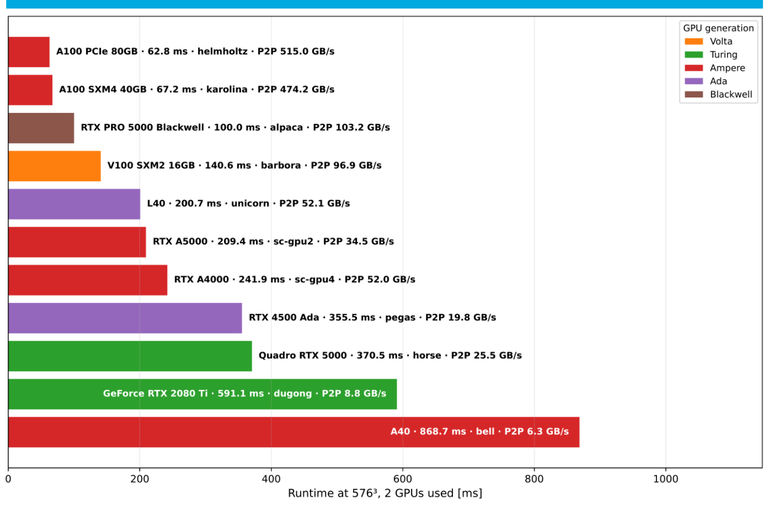
Large three-dimensional k-Wave simulations often exceed the memory and runtime limits of a single GPU, especially in demanding ultrasound applications such as transcranial simulations, where large domains and short turnaround times are required.
To overcome this limit, we developed a **multi-GPU k-Wave Global FFT solver** based on the original single-GPU implementation. The solver preserves the full-domain spectral formulation, but distributes the computation across multiple GPUs, introducing repeated distributed FFTs and global data redistributions. This poster benchmarks the solver on NVIDIA systems ranging from older PCIe workstations to NVLink/NVSwitch servers and PCIe 5.0 Blackwell hardware. Runtime is reported as the average time per simulation step over 400 time steps. FFT-friendly cubic domains are used because similarly sized FFT-unfriendly domains can be up to three times slower.

Server Name	GPU Name	Generation	# GPU	Memory per GPU	Interconnect	PCIe gen	CUDA Architecture	CUDA Driver	Power	P2P Speed
helmholtz	NVIDIA A100 80GB PCIe	Ampere	2	80 GB	NVLink	4	8.0	12.8	300 W	515.0 GB/s
karolina	NVIDIA A100-SXM4-40GB	Ampere	8	40 GB	NVSwitch	4	8.0	13.0	400 W	474.2 GB/s
alpaca	NVIDIA RTX PRO 5000 Blackwell	Blackwell	4	48 GB	PCIe	5	12.0	13.0	300 W	103.2 GB/s
barbora	Tesla V100-SXM2-16GB	Volta	4	16 GB	NVLink	3	7.0	12.2	300 W	96.9 GB/s
unicorn	NVIDIA L40	Ada	7	46 GB	PCIe	4	8.9	13.0	300 W	52.1 GB/s
sc-gpu4	NVIDIA RTX A4000	Ampere	8	16 GB	PCIe	4	8,6	13.0	140 W	52.0 GB/s
sc-gpu2	NVIDIA RTX A5000	Ampere	4	24 GB	PCIe	4	8,6	13.0	230 W	34.5 GB/s
horse	Quadro RTX 5000	Turing	3	16 GB	PCIe	3	7.5	13.0	230 W	25.5 GB/s
ponny	Quadro RTX 5000	Turing	2	16 GB	PCIe	3	7.5	13.0	230 W	19.9 GB/s
pegas	NVIDIA RTX 4500 Ada Generation	Ada	8	24 GB	PCIe	3	8.9	13.0	210 W	19.8 GB/s
dugong	NVIDIA GeForce RTX 2080 Ti	Turing	2	11 GB	PCIe	3	7.5	13.0	250 W	8.8 GB/s
bell	NVIDIA A40	Ampere	2	48 GB	PCIe	3	8.0	12.8	300 W	6.3 GB/s
sc-gpu1	NVIDIA GeForce GTX 1080	Pascal	4	8 GB	PCIe	3	6,1	13.0	180 W	6.3 GB/s
deer	NVIDIA GeForce RTX 2080 Ti	Turing	2	11 GB	PCIe	3	7.5	13.0	300 W	5.9 GB/s
upsy	NVIDIA GeForce GTX 1080 Ti	Pascal	1	11 GB	PCIe	3	6.1	12.6	250 W	

Single-GPU Performance

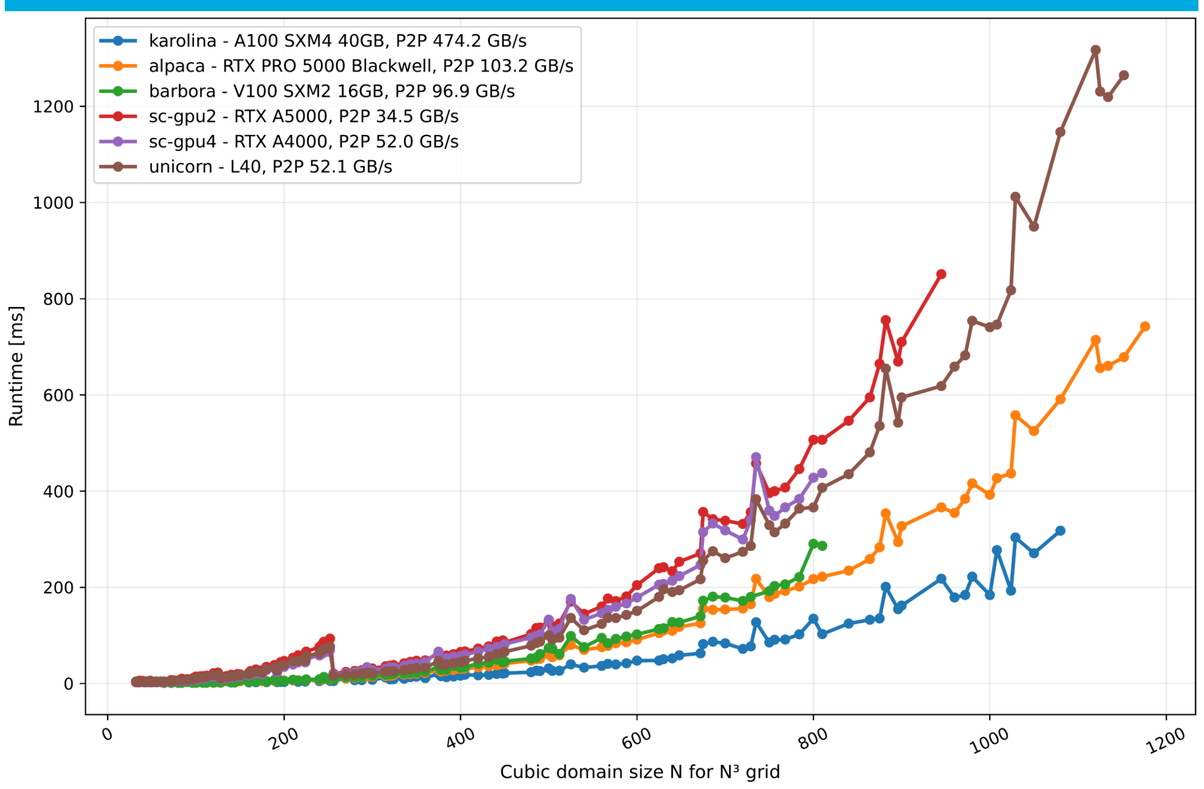


Two-GPU Performance

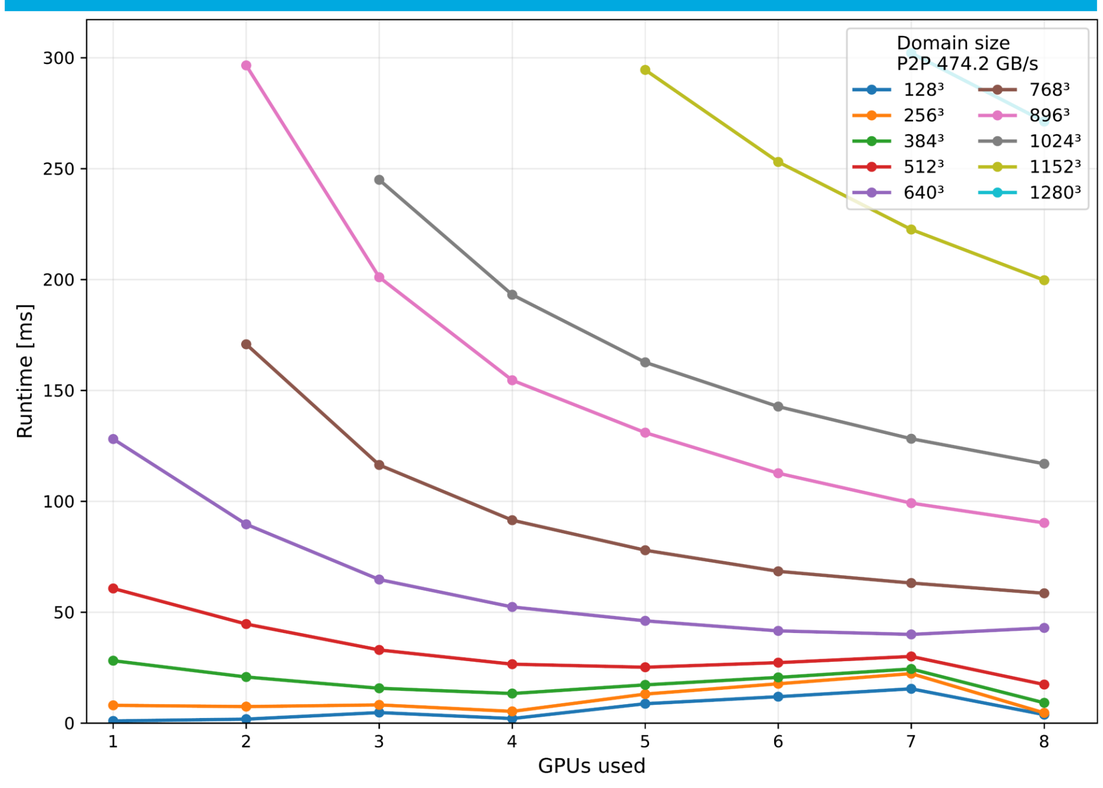


Four-GPU Performance

Four-GPU Runtime across Domain Sizes



Karolina – Strong Scaling



Conclusion

The multi-GPU k-Wave Global FFT solver runs across a wide range of NVIDIA platforms, but performance depends strongly on the hardware balance of each system. Measured P2P bandwidth explains much of the observed multi-GPU behavior, although it is not a complete predictor of runtime; GPU generation, memory bandwidth, memory capacity, topology, and FFT efficiency also contribute.
Multi-GPU execution is most beneficial for sufficiently large domains, where distributed FFT and redistribution costs are amortized by more computation. For smaller domains, communication, synchronization, and launch overheads can reduce the benefit of additional GPUs.
Among the tested systems, **Karolina provides the best full-node performance**, mainly due to A100 SXM GPUs and NVSwitch. **Alpaca is the strongest PCIe-only platform** and can compete with older, more expensive high-end systems when P2P communication is fast and uniform.

k-Wave
A MATLAB toolbox for the time-domain simulation of acoustic wave fields

European Innovation Council

citrus

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