Accelerating AWP-ODC for Large-scale Earthquake Simulations Using MVAPICH2

Arnav Talreja (atalreja@ucsd.edu)[1], Akash Palla [1], Daniel Roten [2], Qinghua Zhou[3], Yifeng Cui [2]
[1]University of California, San Diego, [2]San Diego Supercomputer Center, [3]Ohio State University

Summary

AWP-ODC is a 4th-order finite difference code used for linear wave propagation[1,3,8], Iwan-type nonlinear dynamic rupture and wave propagation[10,11], and Strain Green Tensor simulation[2]. We have ported and verified the linear and topography version of AWP-ODC, with discontinuous mesh as well as topography, to HIP so that it can also run on AMD GPUs [5,6]. The topography code achieved a 99.6% parallel efficiency on 4,096 nodes on Frontier, a Leadership Computing Facility at Oak Ridge National Laboratory. We have also implemented CUDA-aware features and on-the-fly GDR compression in the linear version of the ported HIP code. These enhancements significantly improve data transfer efficiency between GPUs, reducing communication overhead and boosting overall performance. We have also extended CUDA-aware features to the topography version and are actively working on incorporating GDR compression into this version as well. We see 154% benefits over IMPI in MVAPICH2-GDR with CUDA-aware support and on-the-fly compression for linear AWP-ODC on 16 Lonestar6 A100 nodes. Furthermore, we have successfully integrated a checkpointing feature into the nonlinear IWAN version of AWP-ODC, prepared for future extreme-scale simulation during Texascale Days of Frontera at TACC.

AWP-ODC Software Progress on Heterogeneous Machines

Newly added feature in AWP-ODC[4] includes CUDA-Aware, which supports passing GPU buffers directly to MPI calls, with 15% performance gain observed compared to the original configuration setup. This feature has been implemented on both HIP and CUDA AWP-ODC versions. OSU NOWLAB's on-the-fly message compression is enabled in AWP-ODC through the enhancement of lossless and lossy compression algorithms, MPC and ZFP, respectively[14]. The redesign in MVAPICH2 MPI library results in 19% and 37% improvement in the GPU computing flops in AWP on V100s, with enhanced MPC-OPT and ZFP-OPT schemes respectively[14]. This is the first work that leverages the GPU-based compression techniques to significantly improve the GPU communication performance in a real application[14]. On TACC Lonestar-6 A100s, we observed 48%-64% benefits using on-the-fly MPC compression using MPC over GDR[6]. Combined MVAPICH2-GDR enhancement over IMPI, including both CUDA-aware support and on-the-fly compression, improves application performance by 154% on 16 nodes[6]. We have also ported AWP-ODC to NVIDIA's latest GH200 on Vista at TACC, the initial benchmarking results are presented in the MLUPS Table below.

AWP-ODC benchmarks on TACC Lonestar-6 A100 nodes with 160x160x2048 per GPU configuration*

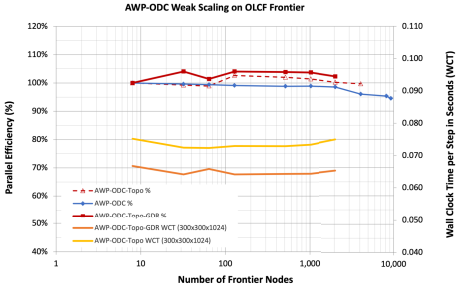
Lonestar6	mvapich2-2.3.7			mvapich2-2.3.7-gdr			mvapich2-2.3.7-gdr-compresson		
a100	gcc11.2.0			gcc11.2.0			gcc11.2.0		
nodes	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.
2	2.0250	0.0488	100.0%	2.2960	0.0399	100.0%	3.7710	0.0261	100.0%
4	4.0270	0.0494	99.4%	4.5260	0.0436	98.6%	6.8510	0.0288	90.8%
8	7.8250	0.0510	96.6%	9.3250	0.0425	101.5%	13.7560	0.0288	91.2%
16	14.4130	0.1543	89.0%	17.1360	0.0460	93.3%	27.5580	0.0288	91.3%
	impi19.0.9			mvapich2-plus-3.0a2			mvapich2-plus-latest		
	gcc11.2.0			gcc11.2.0			gcc11.2.0		
	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.
2	1.6800	0.0585	100.0%	2.391	0.0411	100.0%	3.151	0.0311	100.0%
4	3.4800	0.0572	103.6%	4.579	0.0431	95.8%	5.399	0.0366	85.7%
8	5.8170	0.0686	86.6%	7.796	0.0509	81.5%	10.136	0.0391	80.4%
16	10.8380	0.0737	80.6%	15.214	0.0523	79.5%	20.097	0.0395	79.7%

*mvapich2-2.3.7-gdr: runs with CUDA-aware; mvapich2-2.3.7-gdr-compresson: runs with CUDA-aware+on-the-fly compression (MPC)

AWP-ODC	K20X	KNL7250	V100 (NVLink)	V100 (PCIe)	V100 (PCIe+Opt)	A100 (NVLink)	H100 (PCIe)	H100 (PCIe+Opt)	MI250X (Slingshot)	GH200 (PCIe+Opt)
MLUPS**	552	1092	1598	1074	2009	1937	3713	5145	1711	8548
Speedup	1x	1.98x	2.89x	1.95x	3.64x	3.51x	6.72x	9.32x	3.10x	15.36x

**Millions of lattice point update completed per second

CUDA-aware feature has been implemented to the topography version of AWP-ODC, leveraging GPU-Direct RDMA (GDR) for enhanced data transfer efficiency between GPUs. Tests show up to a 13% improvement in performance on ORNL's Frontier system, equipped with AMD MI250X GPUs. The results of the benchmarking are presented in the table beside. We are currently working on implementing and verifying on-the-fly GDR compression using MVAPICH 2.3.7-gdr_compression to this version of AWP-ODC.



Benchmarks of different versions of AWP-ODC on OLCF's Frontier with 300x300x1024 weak scaling

Nodes	Without GDR		With GDR		% gain
	Sec./step	Efficiency	Sec./step	Efficiency	
2	0.0720	100.0%	0.0679	100.0%	6.0%
8	0.0752	95.7%	0.0668	101.6%	12.6%
32	0.0725	99.3%	0.0642	105.7%	12.9%
128	0.0730	98.6%	0.0642	105.7%	13.7%
512	0.0729	98.7%	0.0643	105.5%	13.3%
2048	0.0750	96.0%	0.0653	103.9%	14.8%

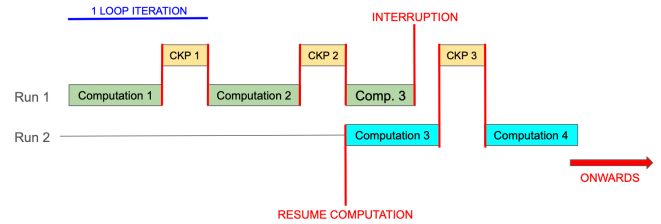
Benchmarks of topography version of AWP-ODC on OLCF's Frontier with 300x300x1024 weak scaling with and without GPU-Aware

Implementation and Verification of Checkpointing feature in AWP-ODC IWAN

We have successfully implemented and verified (up to 4 nodes) checkpointing feature for the non-linear IWAN CPU version of AWP-ODC, for both single-node and multi-node simulations, to improve the resilience and reliability of long-running simulations.

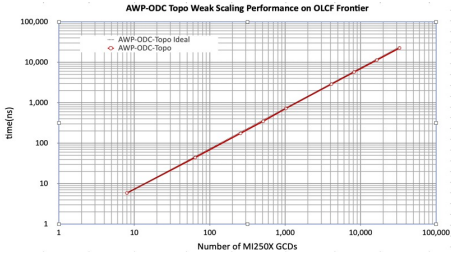
Checkpointing is a crucial functionality that allows the simulation to periodically save its state, enabling recovery from potential failures without the need to restart from the beginning. This is particularly important for large-scale simulations that can run for extended periods, where even minor disruptions could otherwise lead to substantial losses in computational effort and time.

Our current efforts are focused on optimizing the checkpointing process to reduce the I/O throughput during large-scale runs, which can become a significant bottleneck as the number of nodes increases. By evaluating which variables are essential to save and which can be recomputed with minimal overhead, we are able to significantly reduce the volume of data that needs to be written to disk.



Visualization of the checkpointing feature: During each iteration of the main loop, necessary variables are saved after computation is completed, before moving on to the next loop iteration. In case of an error/interruption, computation can be resumed from the last saved checkpoint.

AWP-ODC Topography Ported to HIP on AMD MI250X and Verified



Benchmarks of topography version of AWP-ODC on OLCF's Frontier with 300x300x1024 weak scaling

Frontier[7] at ORNL is the current No. 1 system in the June 2024 TOP500 list. This HPE Cray EX system based on the AMD Radeon Instinct GPUs and EPYC CPUs is the first US system with a peak performance exceeding one ExaFlop/s[5]. AWP-ODC Topography, featuring discontinuous mesh feature along with non-linear computation capability, has been ported and verified on this AMD MI250X based system. *Frontier* weak scaling efficiency is achieved in 99.6% on 4,096 nodes. This implementation also includes CUDA-aware MPI. We are currently verifying on-the-fly GDR compression using MVAPICH 2.3.7-gdr_compression to this version of AWP-ODC on Lonestar-6 at TACC, and working with NOWLAB team for experimentation with the latest MVAPICH-PLUS.

Acknowledgements

Arnav Talreja acknowledges the support provided by the MUG'24 Conference through the travel grant through funding from the U.S. NSF. We thank for the funding supports provided by the NSF under Award # 2311833, the NSF CSA program under Award #2139536, and SCEC RSA #16230509. The authors thank Dr. Ossian O'Reilly of AMD for providing HIP porting advice as well as providing help in debugging the CUDA-Aware implementation for topography code, and Prof. DK Panda NOWLAB team of OSU, including Lang Xu, along with Dr. Te-Yang Yeh of SDSU, for CUDA-Aware and on-the-fly message compression support. The authors acknowledge OLCF for INCITE allocation; TACC for Frontera LSCP allocation and CSA allocation.

References

[1] Cui, Y., K.B. Olsen, T.H. Jordan, K. Lee, J. Zhou,., P. Small, G. Ely, D. Roten, DK Panda, A. Chourasia, J. Levesque, S.M. Day and P. Maechling, Scalable Earthquake Simulation on Petascale Supercomputers, SC10, New Orleans, Nov, 13-19, 2010
[2] Cui, Y., Poyraz, E., Zhou, J., Callaghan, S., Maechling, P., Jordan, T., Shih, L. and Chen, P., Accelerating CyberShake Calculations on XE6/XK7 Platforms of Blue Waters, Proceeding of 2013 Extreme Scaling Workshop, 8-17, I EEE Xplore Digital Library, ISSN 2381-1986, doi: 10.1109/XSW.2013.6, 20, August 15-16, Boulder, 2013.
[3] Cui, Y., E. Poyraz, K.B. Olsen, J. Zhou, K. Withers, S. Callaghan, J. Larkin, C. Guest, D. Choi, A. Chourasia, Z. Shi, S.M. Day, P.J. Maechling, T.H. Jordan (2013), Physics-based seismic hazard analysis on petascale heterogeneous supercomputers, SC13, Denver, CO, November 18-21, 2013.
[4] Cui, Y., Zhou, J., Poyraz, E., Choi, D. J. (2016). AWP-ODC-OS (v1.0), Open source releases under BSD-2 clause license, available from https://github.com/HPGeoC/awp-odc-os
[5] Cui, Y., D. Roten, A. Palla, A. Govind, S. Callaghan, M. Norman, L. Koesterke, W. Zhang and P. Maechling. Progress of porting AWP-ODC to next generation HPC architectures and a 4-Hz Iwan-type nonlinear dynamic simulation of the ShakeOut scenario on TACC Frontera, Sept 11-13, Palm Springs, 2023.
[6] Cui, Y. , Extreme-scale Earthquake simulation with MVAPICH, MUG'23, Columbus, Aug 21-23, 2023.
[7] Frontier: https://www.olcf.ornl.gov/frontier/
[8] Olsen, K.B., Simulation of Three-Dimensional Wave Propagation in the Salt Lake Basin, doctoral dissertation, Univ. of Utah, 1994, p. 157.
[9] Palla, A. and Y. Cui. Internet2/Azure Accelerator 2022 Research Project Final Presentation, Dec 7, 2022. https://internet2.edu/internet2-microsoft-azure-accelerator-2022-researchers-in-action/.
[10] Roten, D., Y. Cui, K. Olsen, S. Day, K. Withers, W. Savran, P. Wang and D. Mu, High-frequency nonlinear earthquake simulations on petascale heterogeneous supercomputers, SC'16, 1-10, Nov 13-18, Salt Lake City, 2016.
[11] Roten, D., T. Yeh, K. Olsen, S. Day and Y. Cui. Implementation of Iwan-type nonlinear rheology in a 3D high-order staggered-grid finite-difference method. BSSA, 2023.
[12] Roten, D. and Cui Y., Nonlinear dynamic modeling for a M7.8 earthquake on the southern San Andreas fault, Frontera User Meeting, Austin, Aug 3-4 2023.
[13] TACC CSA Award In the News. Available at: https://www.tacc.utexas.edu/-/21-scientific-codes-selected-for-new-high-performance-software-improvement-program.
[14] Zhou, Q. N. Kumar, P. Kousha, S. Ghazimirsaeed, H. Subramoni and DK Panda, Designing High-Performance MPI Libraries with On-the-fly Compression for Modern GPU Clusters, *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2021, pp. 444-453, doi: 10.1109/IPDPS49936.2021.00053.