Porting topography version of AWP-ODC to OLCF Frontier, with ROCm-Aware Support to improve the performance of collective communication

Yifeng Cui (yfcui@sdsc.edu)[1], Daniel Roten [1], Akash Palla[1], Arnav Talreja[1], Lars Koesterke[2], Wenyang Zhang[2], Te-Yang Yeh [3], Philip Maechling[4]

[1]San Diego Supercomputer Center, [2]Texas Advanced Computing Center[2], [3]San Diego State University,[4]University of Southern California

Summary

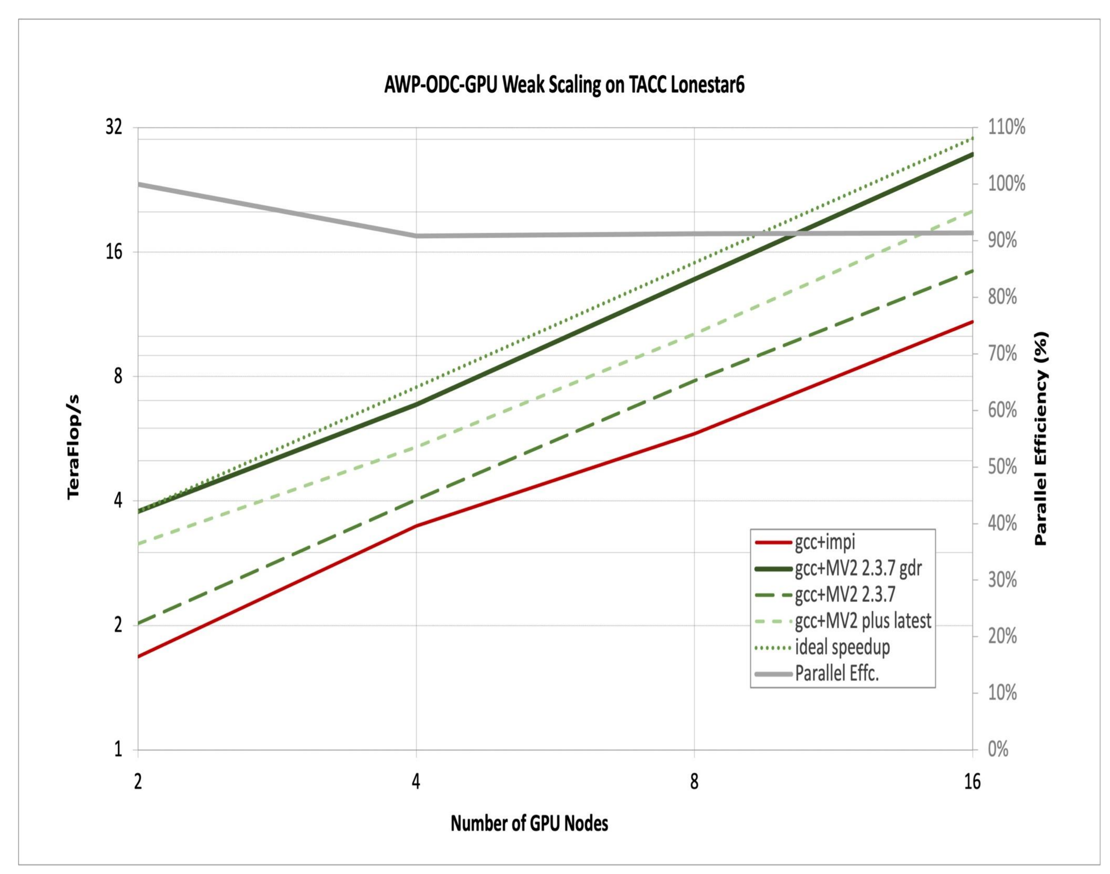
AWP-ODC is a 4th-order finite difference code used by the SCEC community for linear wave propagation, Iwan-type nonlinear dynamic rupture and wave propagation, and Strain Green Tensor simulation. We have ported and verified the topography version of AWP-ODC, with discontinuous mesh feature enabled, to HIP so that it runs on AMD MI250X GPUs. 103.3% parallel efficiency was benchmarked on Frontier between 8 and 4,096 nodes or up to 32,768 GCDs. Frontier is a two exaflop/s computing system based on the AMD Radeon Instinct GPUs and EPYC CPUs, a Leadership Computing Facility at Oak Ridge National Laboratory (ORNL). This HIP topography code has been used in the production runs on Frontier, a primary computing engine currently utilizing the 2024 SCEC INCITE allocation, a 700K node-hours supercomputing time award. Furthermore, we implemented ROCm-Aware GPU direct support in the topo code, and demonstrated 14% additional reduction in time-to-solution up to 4,096 nodes. The AWP-ODC-Topo code is also tuned on TACC Vista, an Arm-based NVIDIA GH200 Grace Hopper Superchip, with excellent performance demonstrated. This poster will demonstrate the studies of weak scaling and the performance characteristics on GPUs. We discuss the efforts of verifying the ROCm-Aware development, and utilizing high-performance MVAPICH libraries with the on-the-fly compression on modern GPU clusters.

NSF CSSI Project: AWP-ODC Optimizations Using MVAPICH

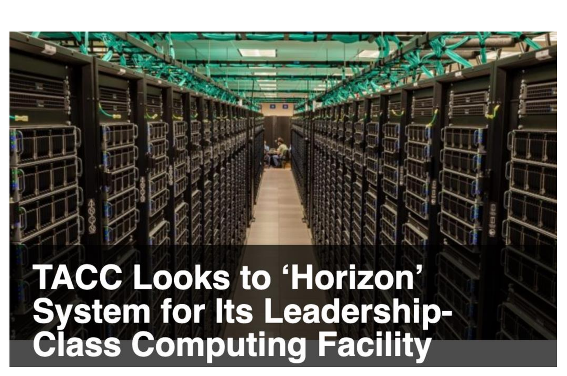
The NSF CSSI project aims to investigate and develop the following innovations by co-designing MVAPICH2 and TAU libraries to scale driving science domains—including AWP-ODC and heFFTe: 1) Load-aware designs for MPI asynchronous communication, 2) Cross runtime coordination for MPI+X applications, 3) Partitioned point-to-point primitives, 4) Application-aware neighborhood collective communication, 5) Support for adaptive persistent collective communication, 6) Coordinating communication kernels on GPUs, and 7) On-the-fly compression.

We profiled AWP-ODC using TAU, identified tensor arrays to leverage asynchronous communication model in the GPU version, enhanced the CPU code with fault tolerance through enabling checkpointing capability. TAU profiling analysis helped to identify the performance bottleneck of HIP topo version of AWP-ODC, resulting in 10x speedup in performance and made possible to run large-scale production simulations with AWP-ODC on Frontier. We tuned the GPU-based AWP-ODC on Nvidia GPU-based systems to run efficiently using MVAPICH2. On TACC Lonestar-6 A100 nodes, we demonstrated 48%-64% benefits using on-the-fly lossless MPC compression over GDR. Combined MVAPICH2-GDR enhancement over IMPI, including both CUDA-aware support and on-the-fly compression, improves the AWP-ODC performance by 154% on 16 Lonestar-6 nodes.

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



NSF *Characteristic Scientific Applications* Project: Porting to GH100



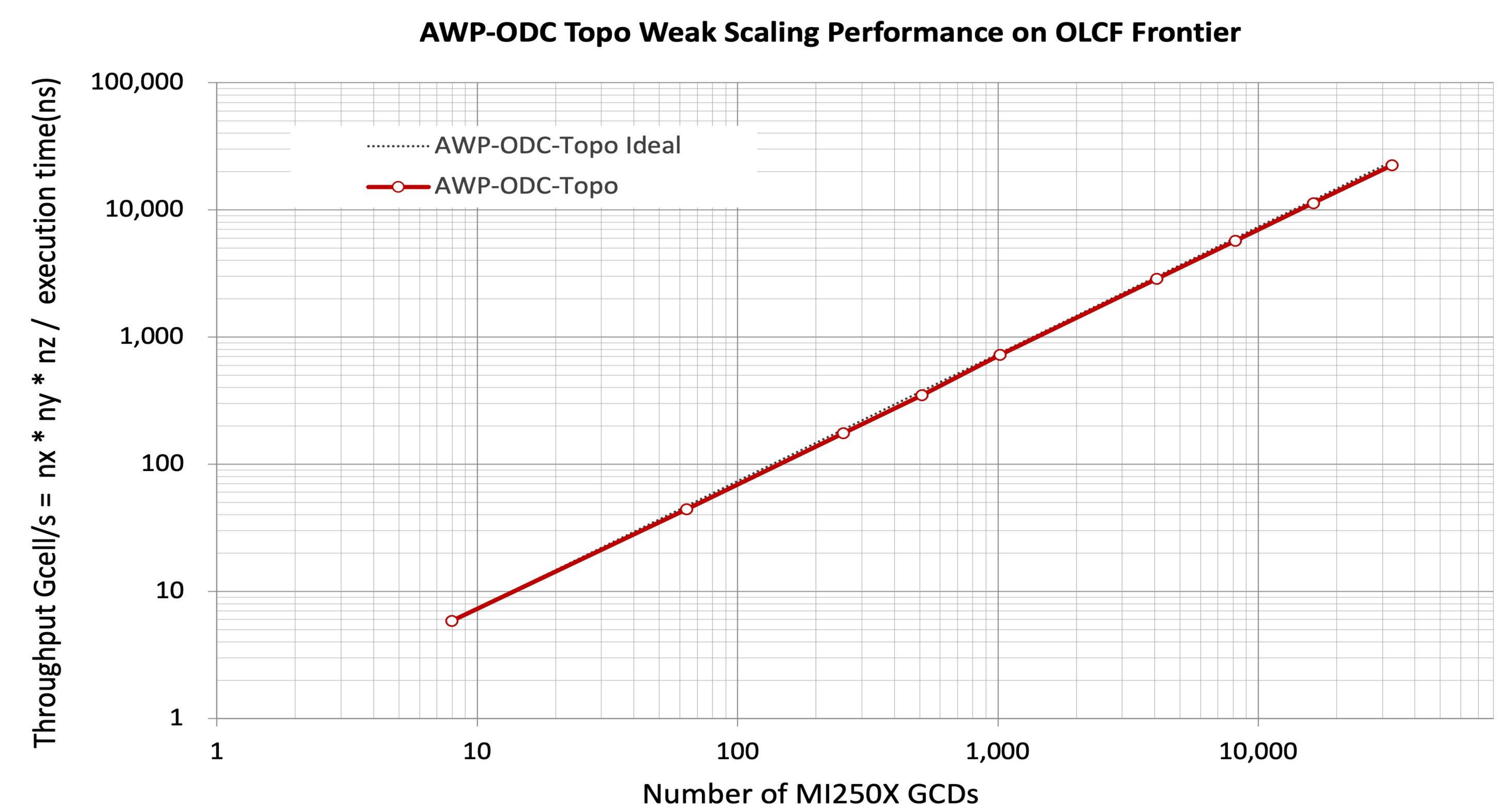
The AWP-ODC-Topo code is tuned on TACC Vista, an Arm-based NVIDIA GH200 Grace Hopper Superchip, with excellent performance demonstrated. This is to prepare for the next generation *Horizon* system to be deployed at TACC.

We added checkpointing capability on the CPU version of AWP-ODC Iwan code, to provide fault tolerance support. This code was used to run a full-machine scale dynamic modeling of San Andreas fault ShakeOut scenario during Texascale Days on Frontera, with rupture dynamics and wave propagation combined in a single step.

Table: AWP-ODC-Topo benchmarks on Vista in weak scaling, using openMPI (execution time seconds per step)

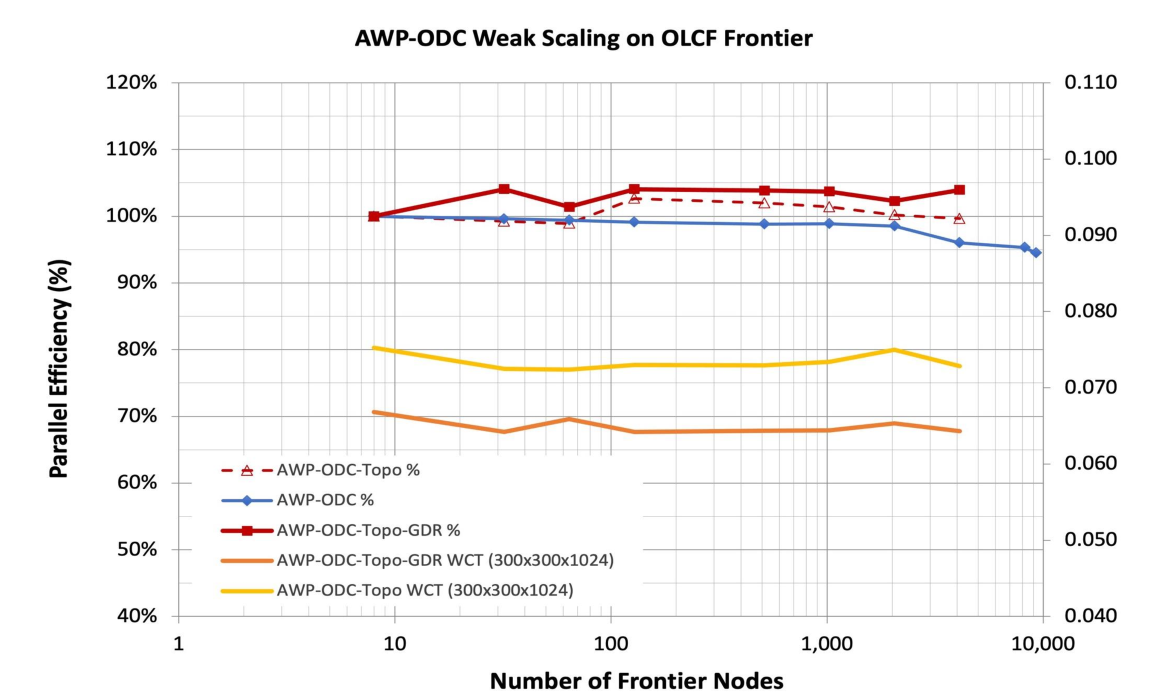
	3-nodes	6-nodes	12-nodes	24-nodes	48-nodes
300x300x128	0.0182	0.0201	0.0207	0.0210	0.0214

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



Frontier at ORNL is the current No. 1 system in the June 2023 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. CUDA based AWP-ODC-Topo has been ported and verified on this AMD MI250X based system. *Frontier* weak scaling efficiency is achieved in 99.6% on 4,096 nodes. The Implementation also includes GPU-aware MPI. The Iwan and discontinuous-mesh based AWP code is being ported to HIP.

Furthermore, we added CUDA-Aware and ROCm-Aware feature to the latest topography version code, which supports for passing GPU buffers directly to MPI calls, and demonstrated 14% performance gain up to 4,096 nodes on Frontier, compared to the original configuration setup.



Acknowledgements

This work is supported by the NSF CSSI Award #2311833 , NSF CSA program under TACC Award #2139536, and SCEC RSA #162830509. SCEC is funded by National Science Foundation (NSF) Cooperative Agreement EAR-1600087 and U.S. Geological Survey (USGS) Cooperative Agreement G17AC00047. The authors thank Dr. Ossian O'Reilly of AMD for providing HIP porting advice, and Prof. DK Panda NOWLAB team of OSU, especially Qinghua Zhou and Lang Xu, for CUDA-Aware and on-the-fly message compression support. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725; and of NSF LSCP and CSA allocations on Frontera, Vista and Lonestar-6 at Texas Advanced Computing Center; and of Director's Discretionary allocation on Expanse at SDSC. This work is also supported by SDSC core program.

References

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

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.

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.

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

Cui, Y. , T.Y. Yeh, Extreme-scale Earthquake simulation with MVAPICH, MUG'24, Columbus, Aug 19-22, 2024.

Frontier: https://www.olcf.ornl.gov/frontier/

Olsen, K.B., Simulation of Three-Dimensional Wave Propagation in the Salt Lake Basin, doctoral dissertation, Univ. of Utah, 1994, p. 157.

O'Reilly, O., T-Y Yeh, K.B. Olsen, Z.F. Zhu, A. Breuer, D. Roten, C. Goulet (2022). A high-order finite-difference method on staggered curvilinear grids for seismic wave propagation applications with topography, BSSA, 112, 3-22, 2022.

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.

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. accepted to BSSA, 2023.

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.

TACC CSA Award In the News. Available at: https://www.tacc.utexas.edu/-/21-scientific-codes-selected-for-new-high-performance-software-improvement-program.

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.