Enhancing Earthquake Simulation Performance and Efficiency Through GPU-Aware Memory Management and Compression Using MVAPICH

Shuzheng Zhang[1], Shijie Wang[1], Akash Palla[1], Yifeng Cui[2]
[1]University of California, San Diego, [2]San Diego Supercomputer Center

Summary

Accurate simulation of earthquake scenarios is essential for advancing seismic hazard analysis and risk mitigation strategies. At the San Diego Supercomputer Center (SDSC), our research focuses on optimizing the performance and reliability of large-scale earthquake simulations using the AWP-ODC software. By implementing GPU-aware MPI calls, we enable direct data processing within GPU memory, eliminating the need for explicit data transfers between CPU and GPU. This GPU-aware MPI achieves nearly ideal parallel efficiency at full scale across both Nvidia and AMD GPUs, leveraging the MVAPICH-Plus 4.0 support on Frontier at Oak Ridge National Laboratory (ORNL) and Vista at the Texas Advanced Computing Center (TACC).

We utilized the MVAPICH-Plus 4.0 compiler to enable on-the-fly ZFP compression, which significantly enhanced inter-node communication efficiency - a critical improvement given the communication bottleneck inherent in large-scale simulations. Our GPU-aware AWP-ODC versions include linear forward, topography and nonlinear Iwan-type solvers with discontinuous mesh support.

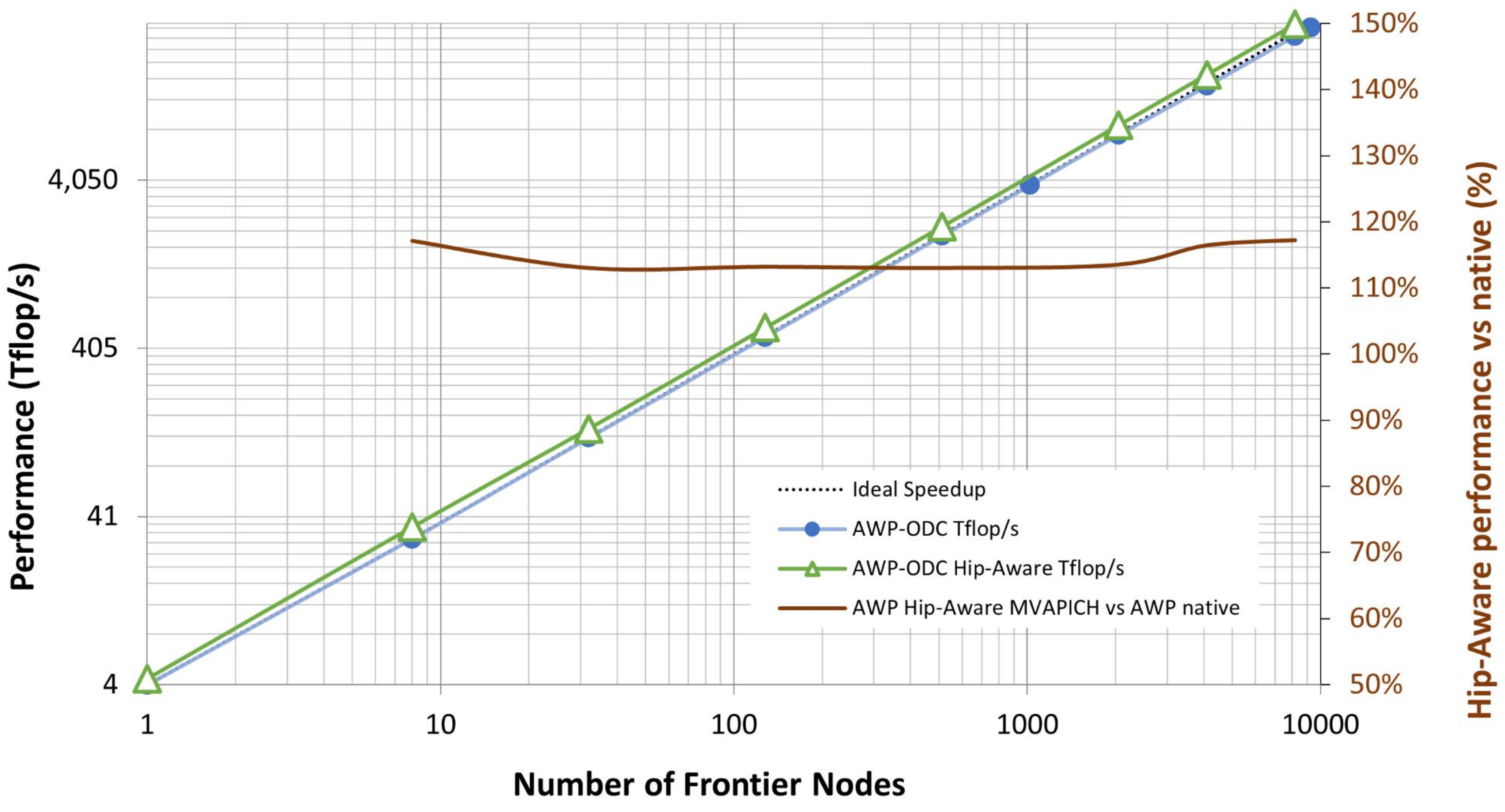
On the Frontier system with MVAPICH 4.0, Hip-aware MPI calls on AMD's MI250X GPUs deliver nearly ideal weak-scaling speedup up to 8,192 nodes for both linear and topography versions. On TACC's Vista system, CUDA-aware MPI calls on GH200 GPU substantially outperform their non GPU-aware counterparts across all three solver versions.

This poster will present a detailed evaluation of GPU-aware AWP-ODC using MVAPICH, including the impact of ZFP message compression compared to the native versions. Our results highlight the importance of MPI library MVAPICH support for accelerating and scaling earthquake simulations.

Accelerating AWP-ODC with GPU-Aware on OLCF Frontier MI250X

Frontier[7] at ORNL is the current No. 2 system in the June 2025 TOP500 list. This system is based on the AMD Instinct GPUs and EPYC CPUs, and it is the first US system with a peak performance exceeding one ExaFlop/s[5]. By porting AWP-ODC linear code to HIP, we saw performance uplifts with MVAPICH-Plus 4.0 using HIP-Aware GDR on the MI250X GPU nodes. We observed consistent speedup when compared to the native code in weak scaling tests, especially pronounced in larger and full-machine scales.

AWP-ODC performance with MVAPICH on OLCF Frontier



Frontier	AWP Native (MVAPICH)			AWP hip-aware (MVAPICH)				Frontier
nodes	AWP Tflop/s	Parallel efficiency	AWP speedup	frontier Tflop/s	Parallel efficiency	AWP speedup	enhanced % vs native	GCDs
1	4.05			4.39				8
8	29.88	100.00%	8	34.99	100.0%	8	117.1%	64
16								128
32	119.07	99.62%	32	134.58	96.2%	31	113.0%	256
64								512
128	473.58	99.06%	127	536.16	95.8%	123	113.2%	1024
256								2048
512	1,890.00	98.83%	506	2,136.19	95.4%	488	113.0%	4096
1024	3,782.54	98.90%	1013					8192
2048	7,536.58	98.53%	2018	8,556.30	95.5%	1956	113.5%	16384
4096	14,690.31	96.02%	3933	17,103.55	95.5%	3910	116.4%	32768
8192	29,168.14	95.33%	7809	34,185.33	95.4%	7816	117.2%	65536

AWP-ODC weak scaling on OLCF Frontier, with 95% parallel efficiency on full machine scale.

MVAPICH2-GDR enhancement improves time-to-solution performance by 17.2% on 8,192 nodes or 65,536 MI250X GCDs.

Weak-scaling test benchmark results for AWP-ODC linear code performance on ORNL Frontier using MVAPICH-Plus 4.0 at a scale of 600x560x200 per node of 8 MI250X each. A GDR with ZFP version was also tested at lower node counts, but no improvement was found over GDR alone, as the faster NIC on Frontier provides up to 800 Gbps of bandwidth[3], rendering the transfer of uncompressed data to be faster than the compression and decompression speed.

Acknowledgements

Shijie Wang and Shuzheng Zhang acknowledges the support provided by the MUG'25 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 #162830509. 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 thank Dr. Sameer Shende in providing technical guidance in using TAU profiling to identify bottlenecks of AWP-ODC. The authors acknowledge OLCF for INCITE allocation; TACC for Vista allocation and CSA allocation.

References

[1] 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.

[2] Cui, Y. , Extreme-scale Earthquake simulation with MVAPICH, MUG'23, Columbus, Aug 21-23, 2023.

[3] OLCF Frontier User Guide: https://www.olcf.ornl.gov/frontier/

[4] TACC Vista User Guide: https://docs.tacc.utexas.edu/hpc/vista/

[5] Q. Zhou, C. Chu, N. S. Kumar, P. Kousha, S. M. Ghazimirsaeed, H. Subramoni and D. K. Panda. Designing High-Performance MPI Libraries with On-the-fly Compression for Modern GPU Clusters*, 2021.

[6] Dhabaleswar K. (DK) Panda (OSU), Co-PIs: Sameer Shende (UO), Ahmad Abdelfattah (UTK), Yifeng Cui (SDSC). CSSI Frameworks: Performance Engineering Scientific Applications with MVAPICH and TAU using Emerging Communication Primitives, 2025.

Accelerating Iwan AWP-ODC with ZFP Compression on TACC Vista GH200

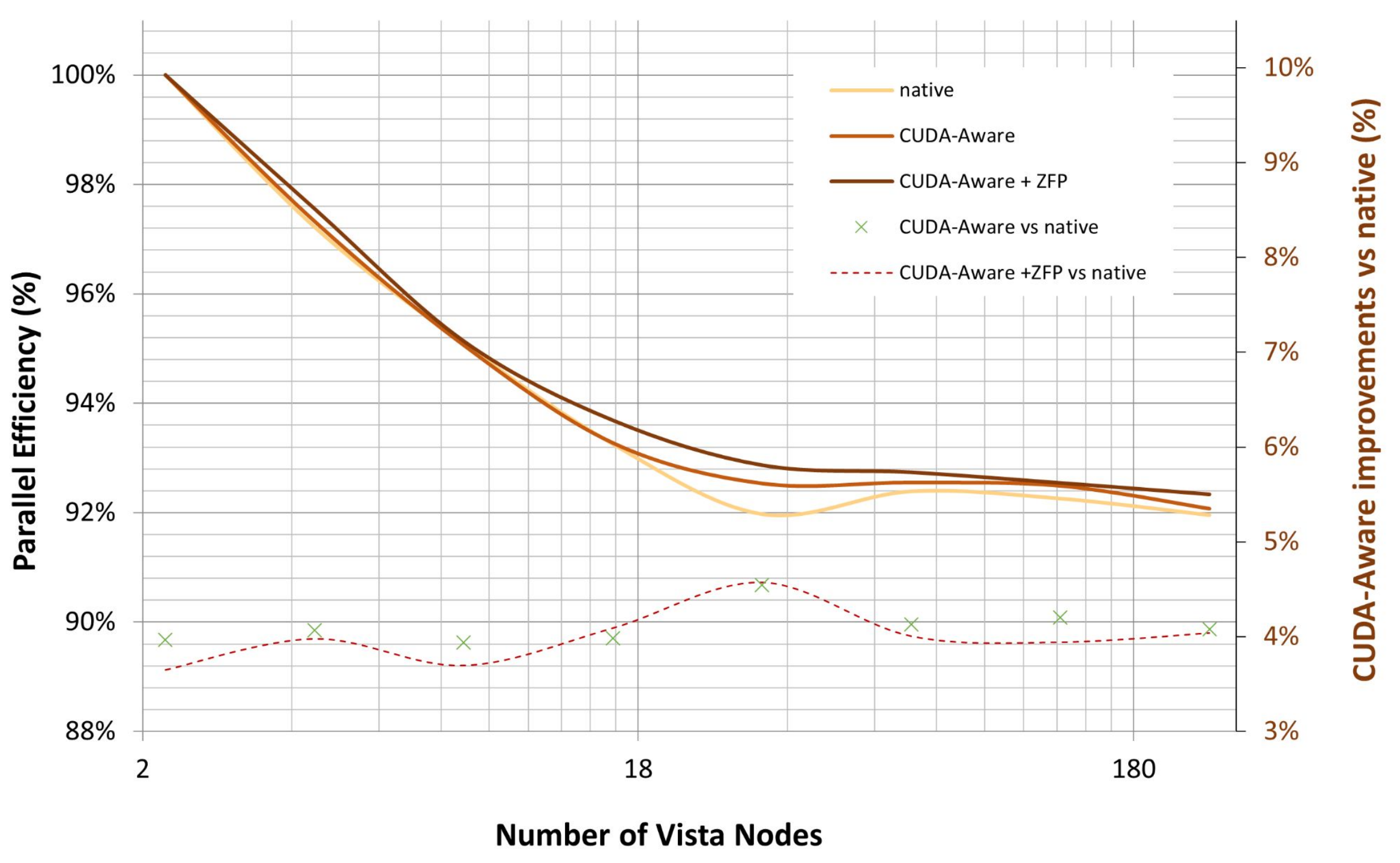
We have implemented the CUDA version of AWP-ODC with Iwan landscape modeling that can run on TACC Vista with GH200 GPUs. CUDA-aware feature has been implemented to the Iwan version of AWP-ODC, leveraging GPU-Direct RDMA (GDR) for enhanced data transfer efficiency between GPUs. We tested AWP-ODC-Iwan with on-the-fly ZFP compression feature of Mvapich-Plus 4.0 and verified its correctness. With the use of ZFP compression and CUDA-Aware MPI calls, AWP-ODC is able to achieve a maximum of 4.1% enhanced performance as shown in the table below.

Vista at TACC is a next-generation, AI-focused supercomputer that bridges TACC's CPU-based Frontera system and the upcoming Horizon exascale system. It is built around ARM-based NVIDIA Grace Hopper (GH200) and dual Grace Superchip nodes, and the nodes are interconnected through Infiniband network[4].

Iwan AWP-ODC benchmarks on TACC Vista with 45x45x80,128,192 per GPU

Vista	AWP Native (MVAPICH)		AWP GDR (MVAPICH)			AWP GDR (MVAPICH ZFP)		
nodes	vista time	Parallel efficiency	vista time	Parallel efficiency	enhanced % vs native	vista time	parallel efficiency	enhanced % vs native
1	0.0328		0.0316			0.0317		
2	0.0338	100.00%	0.0326	100.00%	3.46%	0.0327	100.00%	3.15%
4	0.0347	97.22%	0.0335	97.32%	3.57%	0.0335	97.55%	3.48%
8	0.0355	95.09%	0.0343	95.06%	3.44%	0.0344	95.13%	3.19%
16	0.0362	93.26%	0.0349	93.27%	3.48%	0.0349	93.68%	3.59%
32	0.0367	91.98%	0.0352	92.53%	4.04%	0.0352	92.86%	4.07%
64	0.0365	92.39%	0.0352	92.55%	3.63%	0.0353	92.73%	3.51%
128	0.0366	92.26%	0.0352	92.49%	3.70%	0.0353	92.54%	3.44%
256	0.0367	91.96%	0.0354	92.07%	3.58%	0.0354	92.33%	3.54%

AWP-ODC-Iwan Parallel Efficiency with MVAPICH on TACC Vista



In contrast to the performance of AWP-ODC-Iwan on OLCF Frontier with MI250X GPUs, AWP-ODC-Iwan gained less performance on TACC Vista using GPU-Aware MPI calls. This is primarily due to faster data transfer between CPU and GPU connected via 900 GB/s NVLink[4] in Vista's GH200 node, compared to the 36 GB/s Infinity Fabric[3] in Frontier's MI250X node. Faster data transfer lowers the effect for eliminating explicit data transfer between CPU and GPU.

On-the-fly ZFP Message Compression in Mvapich-Plus 4.0 MPI Library[5]

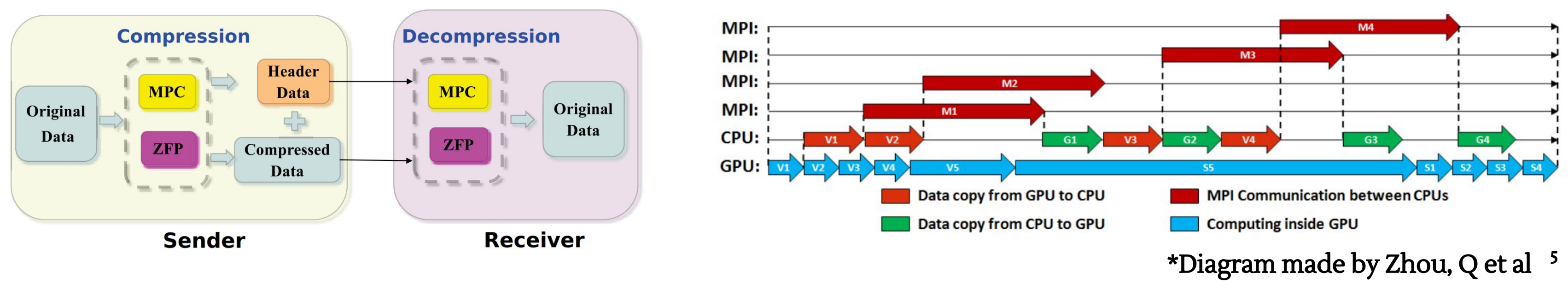
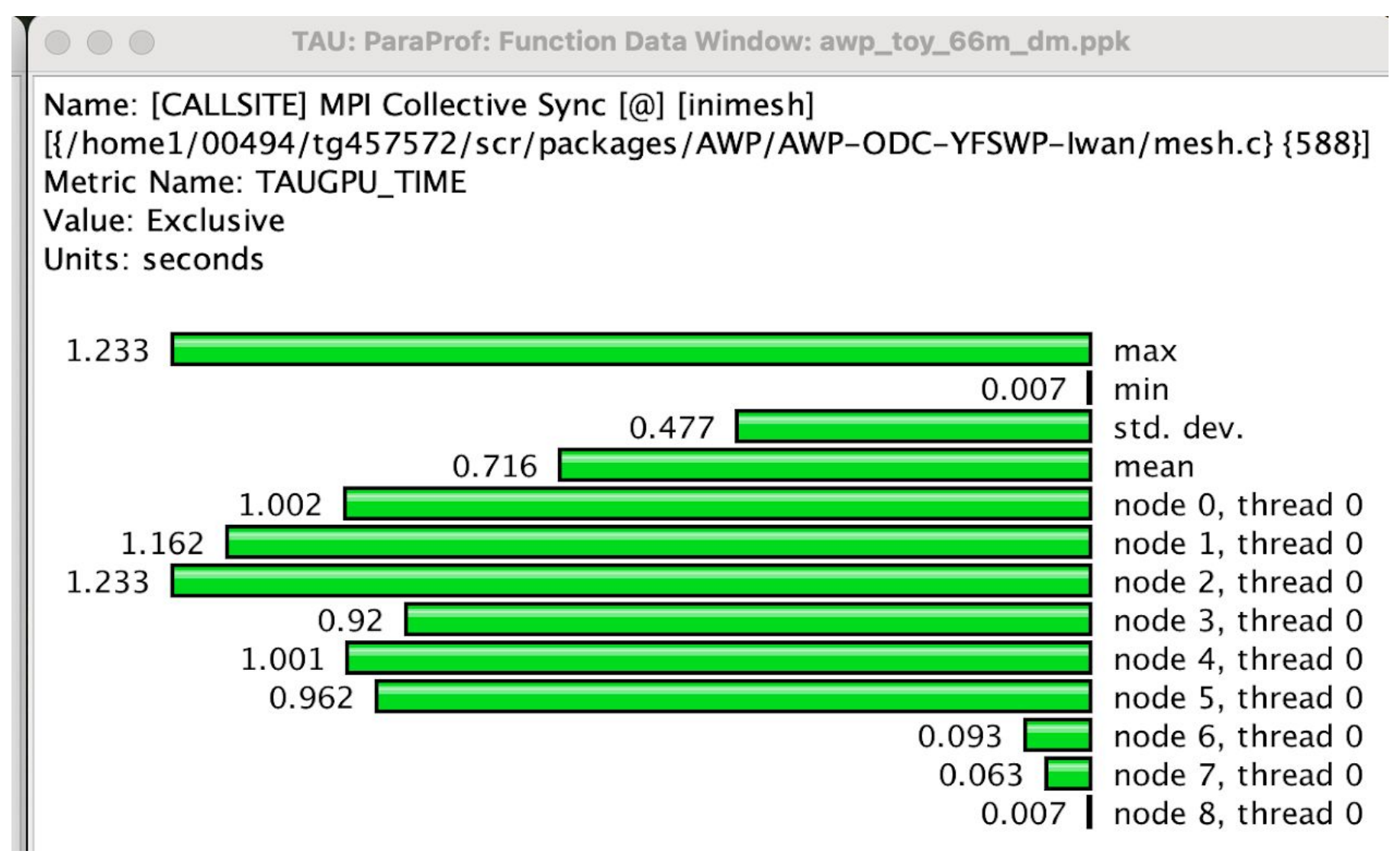


*Diagram made by Zhou, Q et al [5]

TAU Profiling result of MPI Calls in Iwan AWP-ODC[6]



```
Name: [CALLSITE] MPI Collective Sync [@] [inimesh]
[{/home1/00494/tg457572/scr/packages/AWP/AWP-ODC-YFSWP-Iwan/mesh.c} {588}]
Metric Name: TAUGPU_TIME
Value: Exclusive
Units: seconds
```

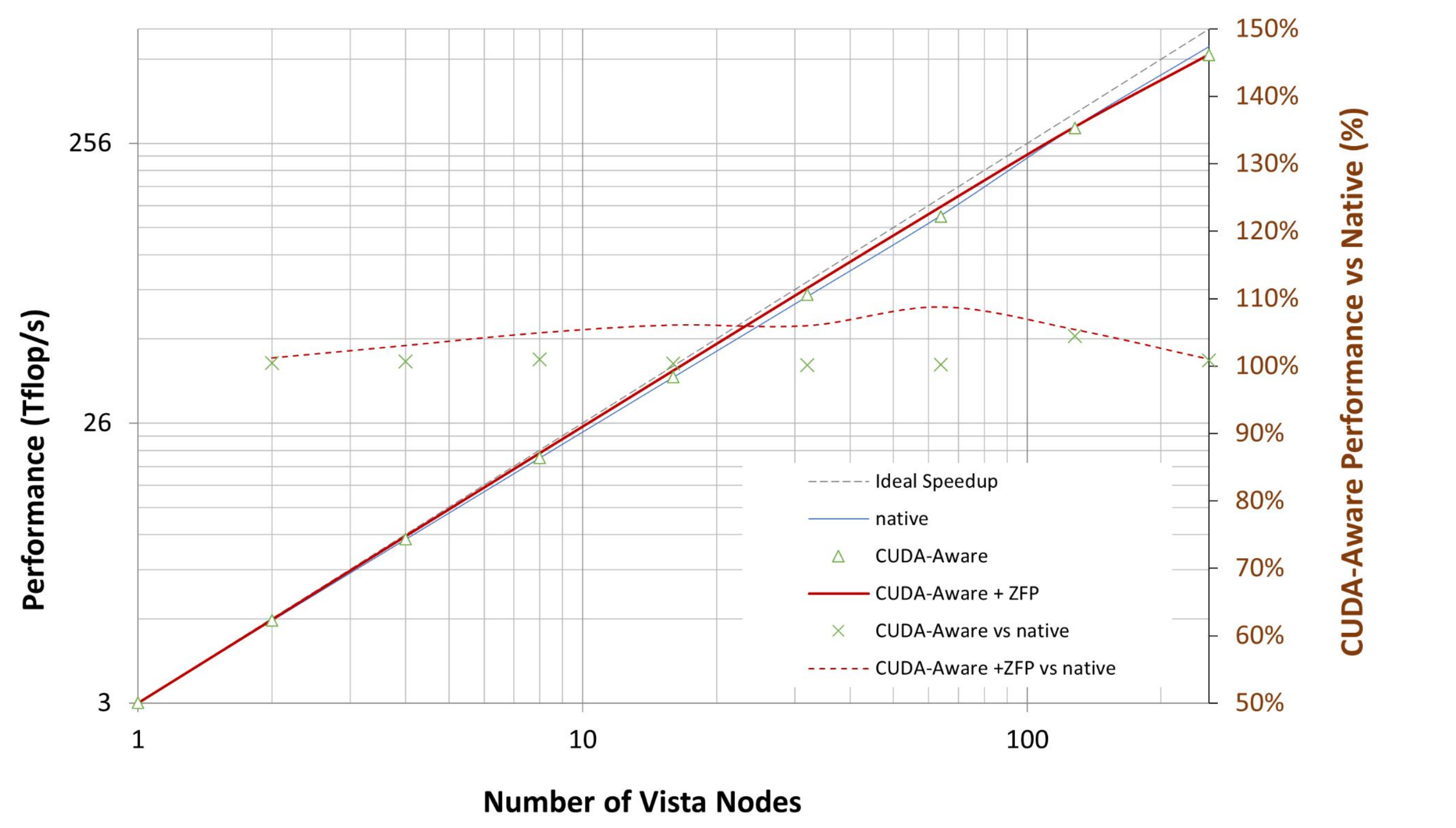
1.233	max
0.007	min
0.477	std. dev.
0.716	mean
1.002	node 0, thread 0
1.162	node 1, thread 0
1.233	node 2, thread 0
0.92	node 3, thread 0
1.001	node 4, thread 0
0.962	node 5, thread 0
0.093	node 6, thread 0
0.063	node 7, thread 0
0.007	node 8, thread 0

Profiling AWP-ODC Iwan testcase on Vista at TACC with TAU showing the time spent in a collective operation stalled at an MPI_Barrier call across multiple MPI ranks[6].

Accelerating Linear AWP-ODC with ZFP Compression on TACC Vista GH200

We have implemented the CUDA version of linear AWP-ODC that can run on TACC Vista with GH200 GPUs. GPU-aware feature has been implemented to the linear version of AWP-ODC, leveraging GPU-Direct RDMA (GDR) for enhanced data transfer efficiency between GPUs. We tested AWP-ODC-Iwan with on-the-fly ZFP compression feature of Mvapich-Plus 4.0 and verified its correctness.

AWP-ODC Tflop/s and CUDA-Aware Performance with MVAPICH on TACC Vista



Linear AWP-ODC Benchmarks on TACC Vista with 640x640x2048 per GPU

Vista	AWP Native (MVAPICH)		AWP GDR (MVAPICH)			AWP GDR (MVAPICH ZFP)		
nodes	vista Tflop/s	parallel efficiency	vista Tflop/s	Parallel efficiency	enhanced % vs native	vista Tflop/s	parallel efficiency	enhanced % vs native
1	2.57		2.57			2.56		
2	5.03	100.00%	5.05	100.00%	100.46%	5.09	100.00%	101.20%
4	9.82	97.68%	9.89	97.86%	100.65%	10.12	99.46%	103.05%
8	19.09	94.91%	19.28	95.42%	101.00%	20.03	98.41%	104.93%
16	37.27	92.63%	37.40	92.55%	100.37%	39.54	97.12%	106.10%
32	73.52	91.38%	73.60	91.05%	100.10%	77.92	95.70%	105.99%
64	140.09	87.06%	140.34	86.81%	100.18%	152.38	93.57%	108.78%
128	278.60	86.57%	290.99	90.00%	104.45%	293.71	90.18%	105.42%
256	526.32	81.77%	530.58	82.05%	100.81%	531.69	81.62%	101.02%

Linear AWP-ODC on TACC Vista, with 82% parallel efficiency on 256-nodes scale.

Little improvement was found over GDR GPU-Aware alone, due to the high data transfer rate between CPU and GPU connected via NVLink in GH200. GDR CUDA-Aware combined with the on-the-fly ZFP compression improves time-to-solution performance by 8.8% on 64 GH200 nodes.