# Performance Engineering of Scientific Applications with MVAPICH and TAU using Emerging Communication Primitives

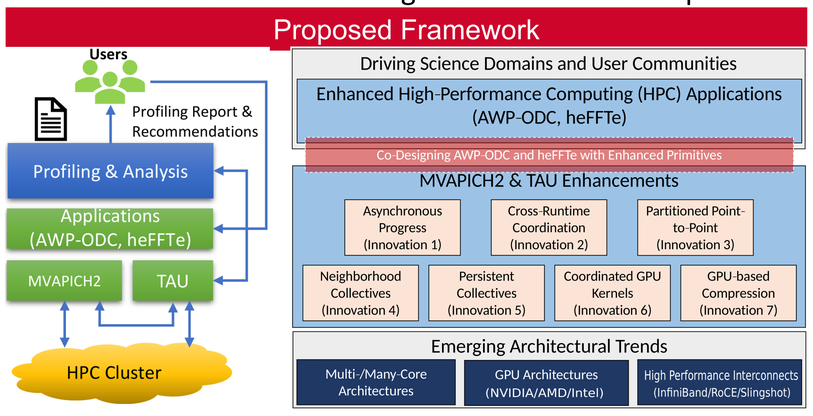
Dhabaleswar K. (DK) Panda[(1)], Sameer Shende[(2)], Ahmad Abdelfattah[(3)], Yifeng Cui[(4)]

[(1)]The Ohio State University (OSU), [(2)]University of Oregon (UO), [(3)]University of Tennessee (UTK), [(4)]San Diego Supercomputer Center (SDSC)

## Abstract

We propose a co-design approach that integrates two powerful tools – MVAPICH and TAU – to demonstrate the new possibilities for performance-guided control and optimization for two large scale applications — AWP-ODC and heFFTe. AWP-ODC is a highly scalable parallel finite-difference application with point-to-point operations that enables 3D earthquake calculations, while heFFTe is a massively parallel application that provides scalable and efficient implementations of the widely used Fast Fourier Transform using several MPI primitives. Through a deep integration between MVAPICH and TAU, the two applications can identify their performance bottlenecks on various supercomputers with different architectures. AWP-ODC and heFFTe can also act as representative real-world benchmarks to MVAPICH and TAU. We show how the co-design approach enables AWP-ODC and heFFTe to deliver better performance on cutting-edge HPC architectures. This is achieved using (1) more optimized and fine-tuned collective operations, and (2) reduced network traffic through real-time data compression.

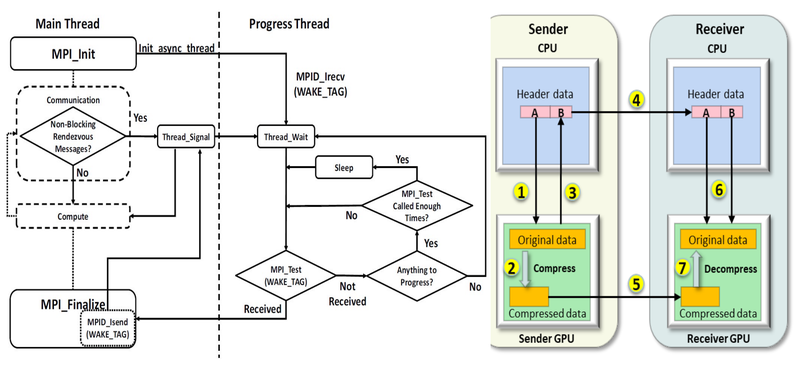
## Proposed Framework




Proposed integrated MPI profiling and performance engineering framework for MVAPICH and TAU, aiding HPC applications (AWP-ODC and heFFTe)

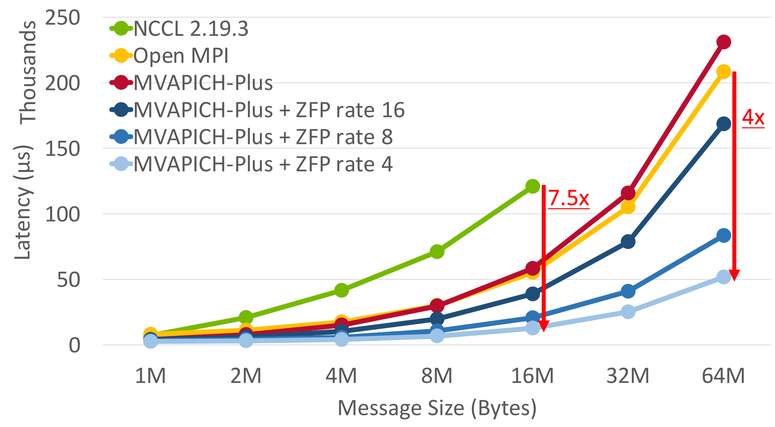
## Proposed Innovations

1. Load-aware designs for MPI asynchronous communication
2. Cross runtime coordination for MPI+X applications
3. Partitioned point-to-point primitives for efficient communication
4. Coordinated communication kernels on GPUs
5. On-the-fly compression for accelerating scientific applications



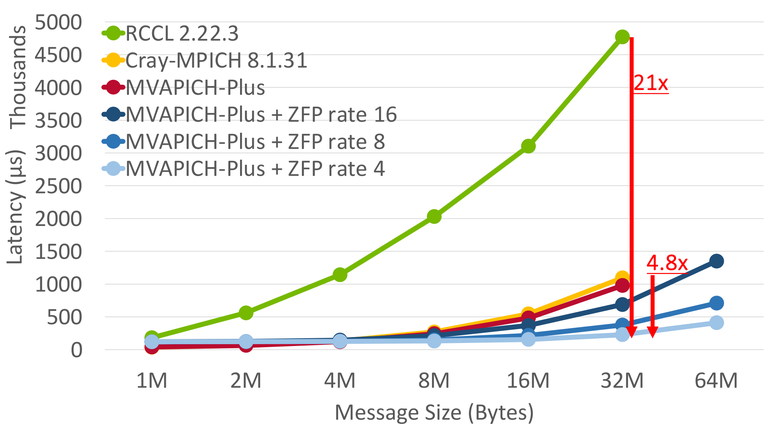

**Asynchronous Progress** **On-the-fly Compression**

## MVAPICH-Plus Benchmark-level Results




**OMB Alltoall on Vista - 128 Nodes (128 NVIDIA GH200s)**

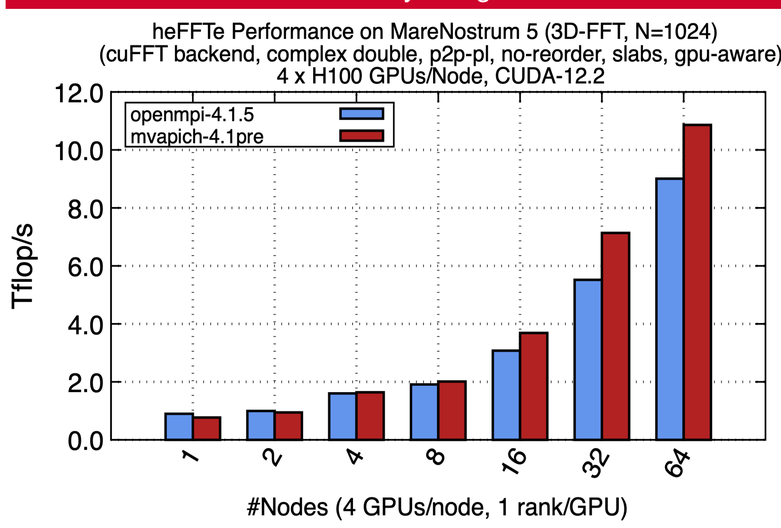
The proposed compression runtime is 7.5x faster than NCCL at 16MB and 4x faster than Open MPI at 64MB.




**OMB Alltoall on Frontier - 32 Nodes (256 AMD MI250Xs)**

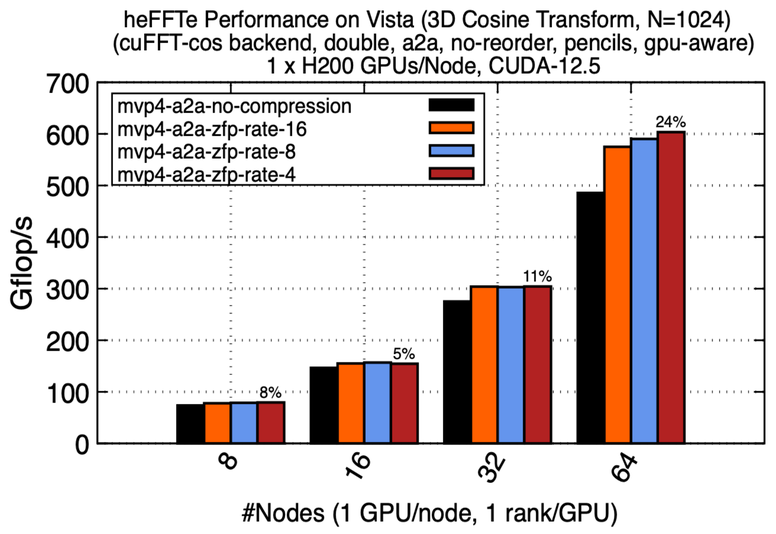
The proposed compression runtime delivers a 21x speedup over RCCL at 32MB and a 4.8x improvement over Cray MPICH at 32MB.

## heFFTe Scalability using MVAPICH



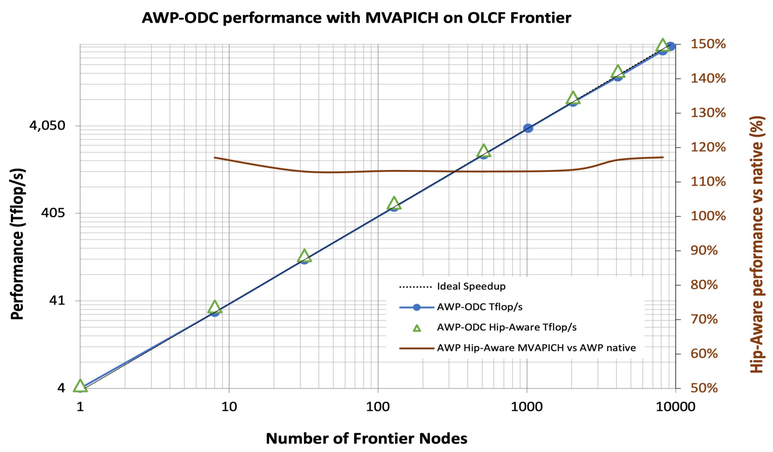

heFFTe strong scaling results on MareNostrum 5 show that using MVAPICH-Plus as a backend achieves up to 29% gain against OpenMPI on up to 64 nodes (256 GPUs)

## heFFTe Preliminary Results Using MVAPICH Data Compression



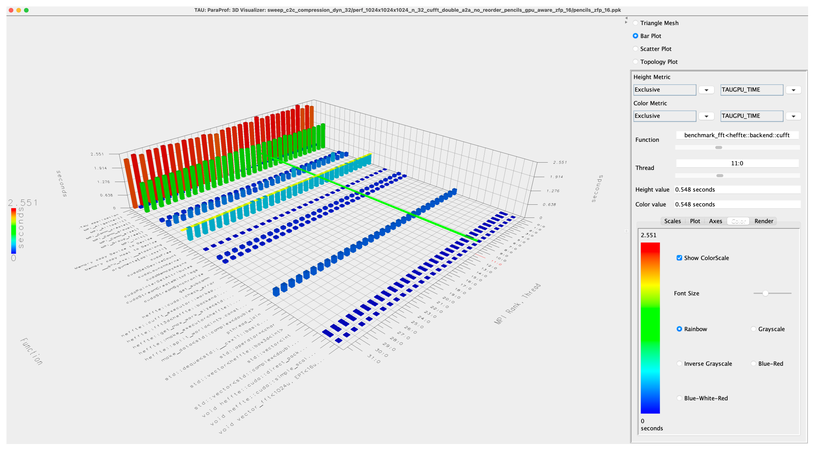

heFFTe strong scaling results on Vista show that MVAPICH-Plus with ZFP achieves throughput improvements of 8%, 5%, 11%, and up to 24% across 8 to 64 nodes.

## AWP-ODC Weak Scaling Results Using MVAPICH



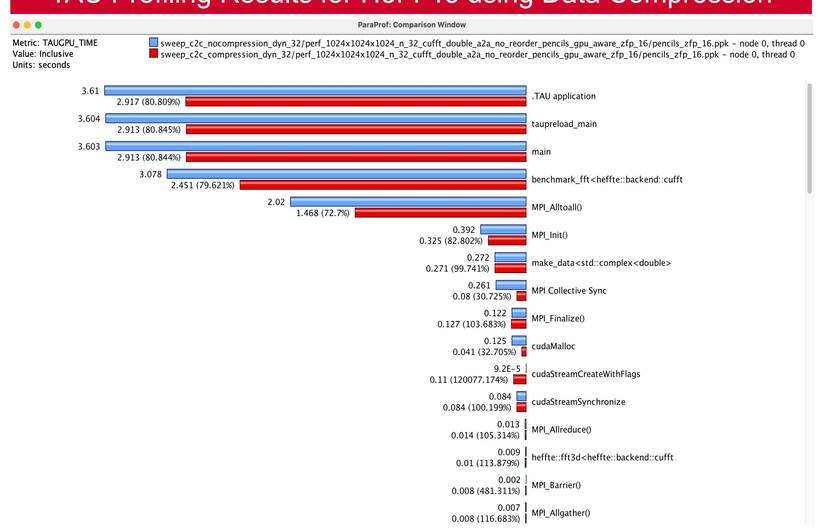

AWP-ODC weak scaling on OLCF Frontier, with 95% parallel efficiency on full machine scale. Its MVAPICH2-GDR enhancement improves time-to-solution performance by 17.2% on 8,192 nodes or 65,536 MI250X GCDs.
TACC Vista (GH200) sees a 3.5% gain for nonlinear solver with on-the-fly compression at 256 nodes.

## TAU's perfomance visualization of HeFFTe using paraprof



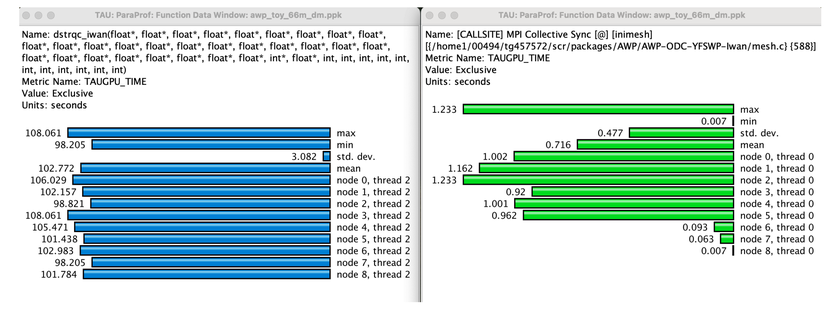

TAU's paraprof 3D visualization shows the shape of the HeFFTe profile run over 32 MPI ranks on Vista where HeFFTe is instrumented using the DyinstAPI library by binary rewriting and MPI level data is extracted from MVAPICH.

## TAU Profiling Results for HeFFTe using Data Compression

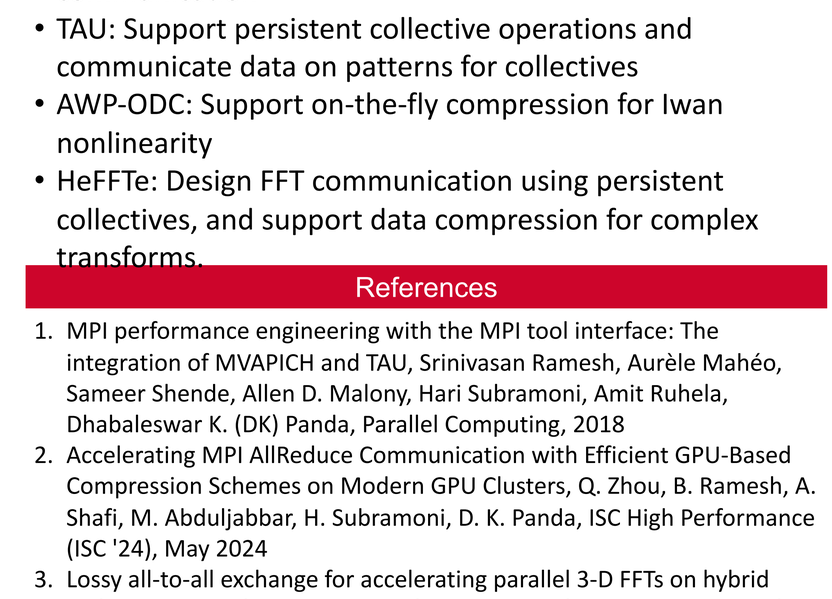



heFFTe performance data viewed in TAU's paraprof for sweep c2c testcase on 32 MPI ranks on Vista at TACC shows performance improvement with compression enabled in MVAPICH.

## TAU Profiling Results for AWP-ODC




Profiling AWP-ODC Iwan testcase on Vista at TACC with TAU showing the distribution of performance of a key GPU kernel across multiple GH200 GPUs (left) and the time spent in a collective operation stalled at an MPI_Barrier call (right) across multiple MPI ranks.




TAU traces of AWP-ODC running on Vista visualized in Perfetto.dev showing timeline views of kernels.

## Future Work

- MVAPICH: Support for adaptive persistent collective communication
- MVAPICH: Application-aware neighborhood collective communication
- TAU: Support persistent collective operations and communicate data on patterns for collectives
- AWP-ODC: Support on-the-fly compression for Iwan nonlinearity
- HeFFTe: Design FFT communication using persistent collectives, and support data compression for complex transforms.

## References

1. MPI performance engineering with the MPI tool interface: The integration of MVAPICH and TAU, Srinivasan Ramesh, Aurèle Mahéo, Sameer Shende, Allen D. Malony, Hari Subramoni, Amit Ruhela, Dhabaleswar K. (DK) Panda, Parallel Computing, 2018
2. Accelerating MPI AllReduce Communication with Efficient GPU-Based Compression Schemes on Modern GPU Clusters, Q. Zhou, B. Ramesh, A. Shafi, M. Abduljabbar, H. Subramoni, D. K. Panda, ISC High Performance (ISC '24), May 2024
3. Lossy all-to-all exchange for accelerating parallel 3-D FFTs on hybrid architectures with GPUs, S. Cayrols, J. Li, G. Bosilca, S. Tomov, A. Ayala and J. Dongarra, 2022 IEEE International Conference on Cluster Computing (CLUSTER), 2022
4. Porting topography version of AWP-ODC to OLCF Frontier, with ROCm-Aware support to improve the performance of collective communication, Cui Y, Palla A, Talreja A, Koesterke L, Zhang W, Yeh T, Maechling P, SCEC'24, Palm Springs, Sept 8-11, 2024.
5. OMB-Compr: An Extension to OSU Micro Benchmarks for Collective Compression Error Measurement, J. Queiser, N. Contini, H. Subramoni, DK Panda, Practice and Experience in Advanced Research Computing (PEARC 2025), 2025

**Supported by: OAC-2311830, OAC-2311831, OAC-2311832, OAC-2311833**

MVAPICH heFFTe AWP-ODC