MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC'23)

by

Hari Subramoni

The Ohio State University

E-mail: subramon@cse.ohio-state.edu

https://web.cse.ohio-state.edu/~subramoni.1/

History of MVAPICH

- A long time ago, in a galaxy far, far away…. (actually 22 years ago), there existed…
- MPICH
  - High performance and widely portable implementation of MPI standard
  - From ANL
- MVICH
  - Implementation of MPICH ADI-2 for VIA
  - VIA – Virtual Interface Architecture (precursor to InfiniBand)
  - From LBL
- VAPI
  - Verbs level API
  - Initial InfiniBand API from IB Vendors (older version of OFED/IB verbs)

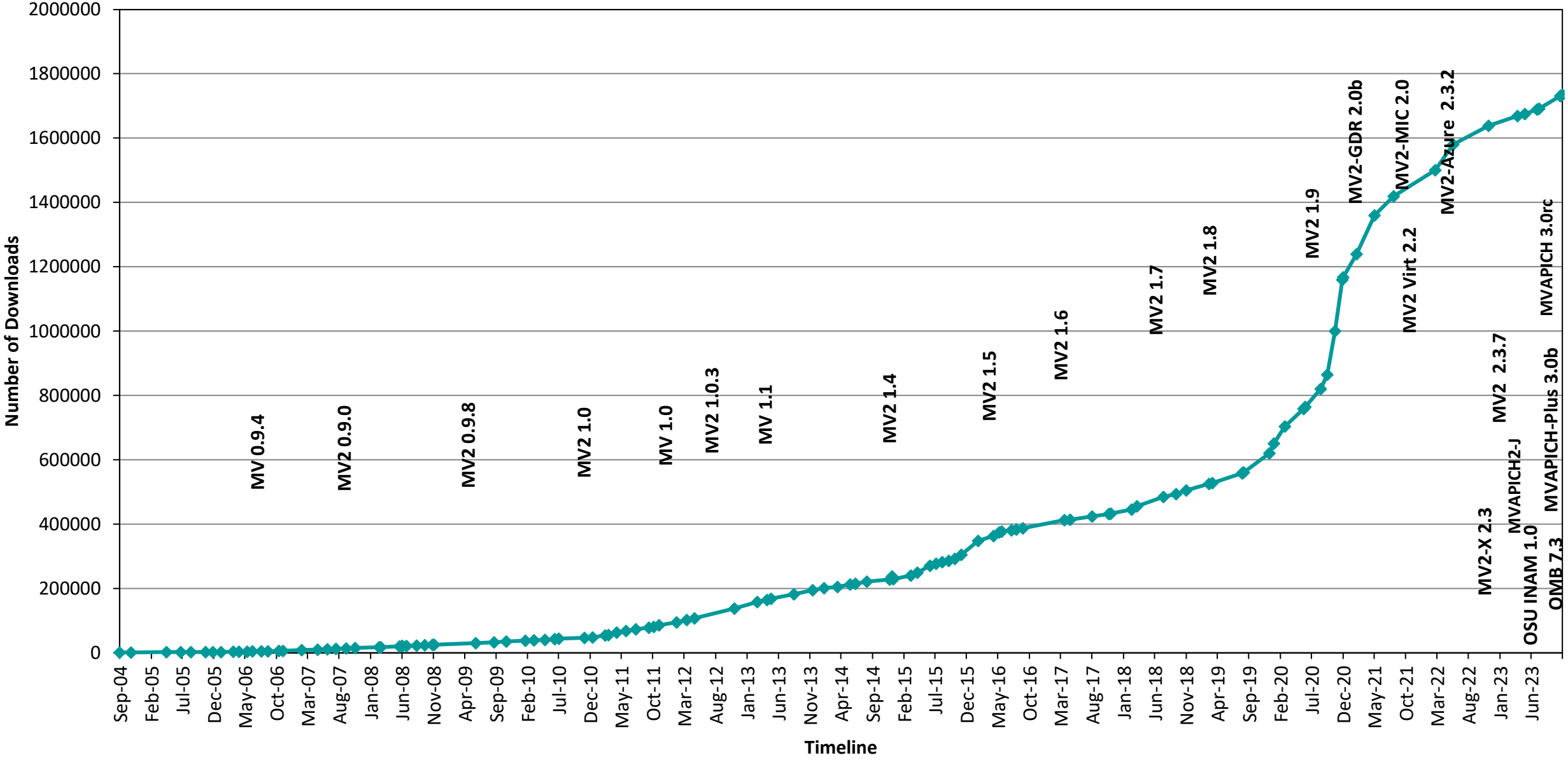
MPICH + MVICH + VAPI = MVAPICH

Overview of the MVAPICH2 Project

- High Performance open-source MPI Library
- Support for multiple interconnects
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, OPX, Broadcom RoCE, Intel Ethernet, Rockport Networks, Slingshot 10/11
- Support for multiple platforms
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and AMD)
- Started in 2001, first open-source version demonstrated at SC '02
- Supports the latest MPI-3.1 standard
- http://mvapich.cse.ohio-state.edu
- Additional optimized versions for different systems/environments:
  - MVAPICH2-X (Advanced MPI + PGAS), since 2011
  - MVAPICH2-GDR with support for NVIDIA (since 2014) and AMD (since 2020) GPUs
  - MVAPICH2-MIC with support for Intel Xeon-Phi, since 2014
  - MVAPICH2-Virt with virtualization support, since 2015
  - MVAPICH2-EA with support for Energy-Awareness, since 2015
  - MVAPICH2-Azure for Azure HPC IB instances, since 2019
  - MVAPICH2-X-AWS for AWS HPC+EFA instances, since 2019
- Tools:
  - OSU MPI Micro-Benchmarks (OMB), since 2003
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015

- Used by more than 3,325 organizations in 90 countries
- More than 1.73 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '23 ranking)
  - 11th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 29th , 448, 448 cores (Frontera) at TACC
  - 46th, 288,288 cores (Lassen) at LLNL
  - 61st, 570,020 cores (Nurion) in South Korea and many others
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)
- Partner in the 21st ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

MVAPICH2 Release Timeline and Downloads

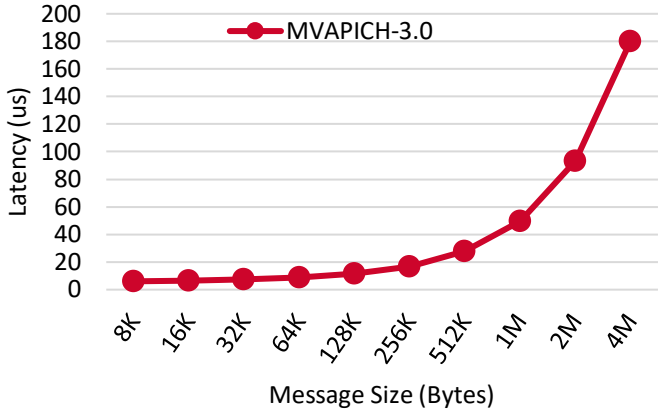


MVAPICH2 Software Family

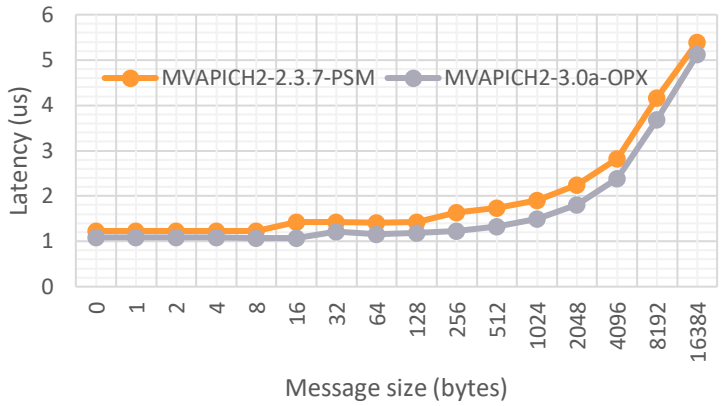
High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters with NVIDIA GPUs
MVAPICH2-Plus	Advanced MPI with unified MVAPICH2-GDR and MVAPICH2-X features for HPC, DL, ML, Big Data and Data Science applications
MVAPICH2-J	Java bindings for MVAPICH2 family of libraries
Microbenchmarks	
OMB	Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs
Tools	
OSU INAM	Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration
OEMT	Utility to measure the energy consumption of MPI applications

MVAPICH-3.0 Point-to-Point

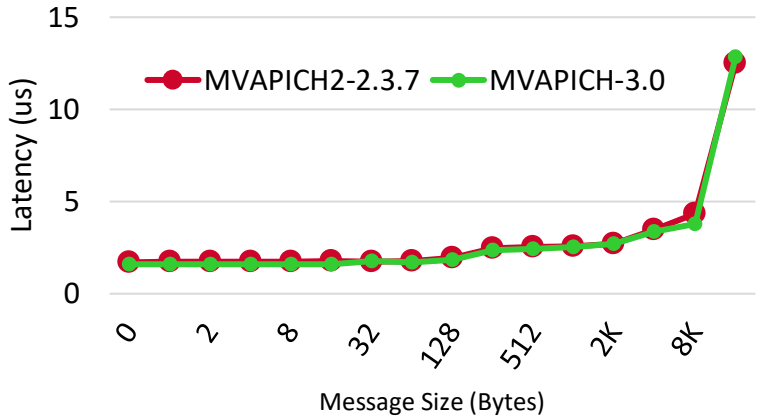
Latency on Slingshot 11



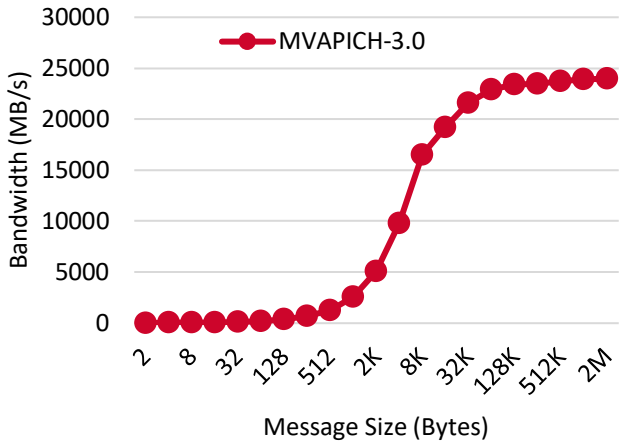
Latency on OPX



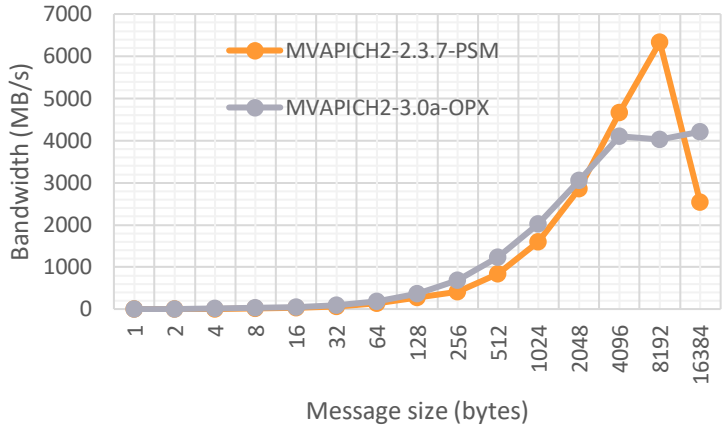
Latency on IB



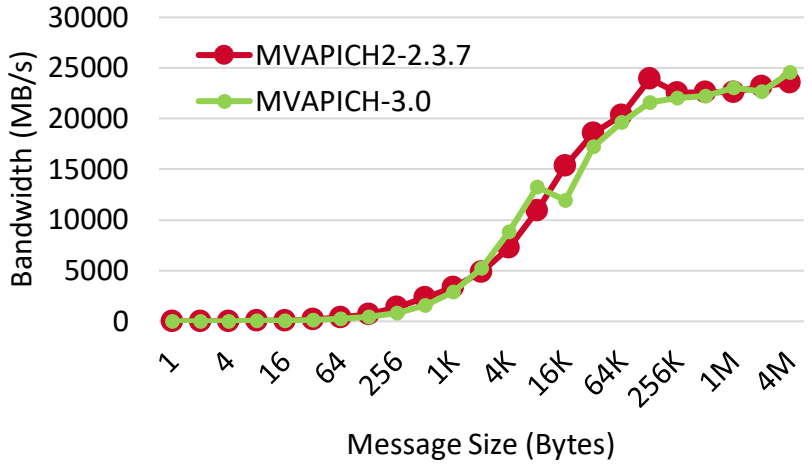
Bandwidth on Slingshot 11



Bandwidth on OPX



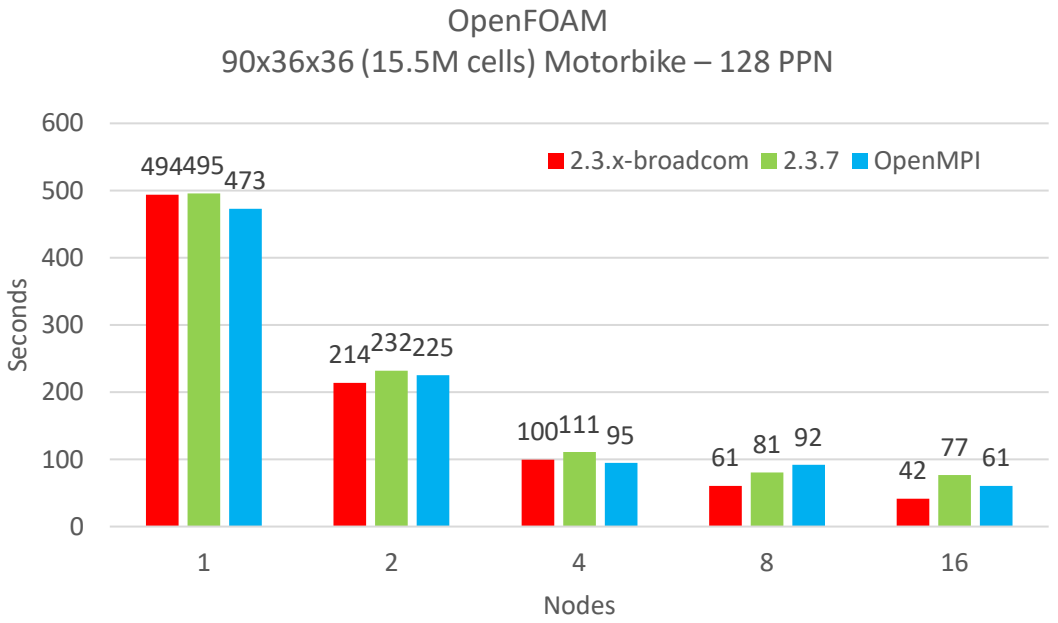
Bandwidth on IB

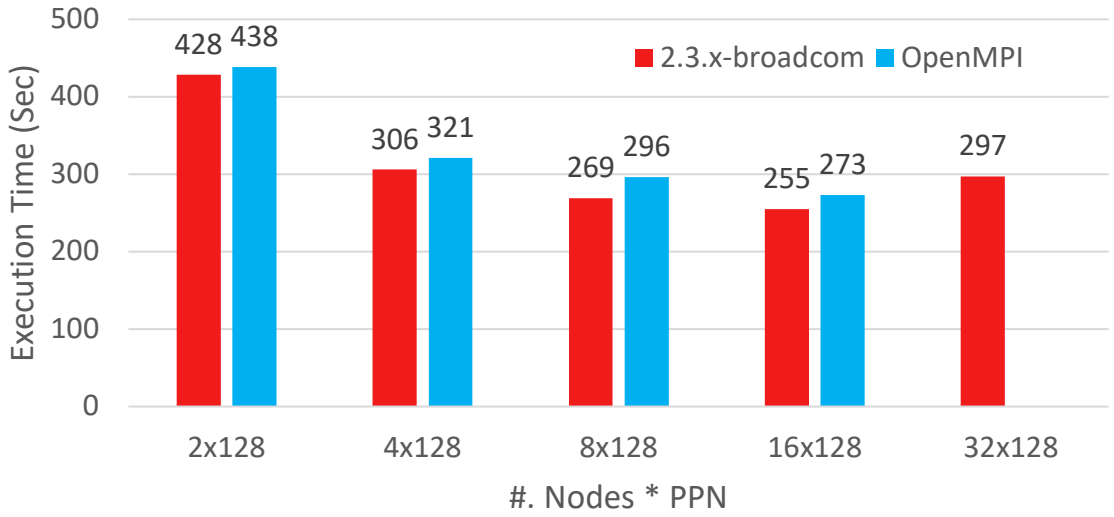


MVAPICH2-Broadcom (Highly optimized for Broadcom's RoCEv2 NIC)

- The MVAPICH2-Broadcom collaboration brings highly scalable enhancements to the Ethernet world through:
  - IB level optimizations targeting message rate and latency
  - Selective Coalescing
  - Collective and point-to-point optimizations
  - Advanced process binding policies and much more
- HPC Applications gains:
  - Reduce up to 45% execution time of OpenFOAM Motorbike on 16 nodes 128 PPN scale
  - Reduce up to 51% execution time of GROMACS benchPEP on 64 nodes 128 PPN scale
  - Reduce up to 9.2% execution time of CP2K H2O-dft-ls (NREP4)

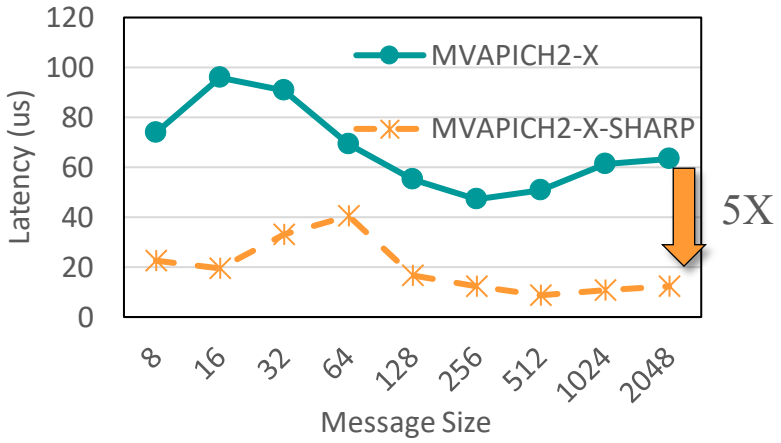
- Regular code-drops to appear in
  - MVAPICH2-2.3.8
  - MVAPICH-3.0



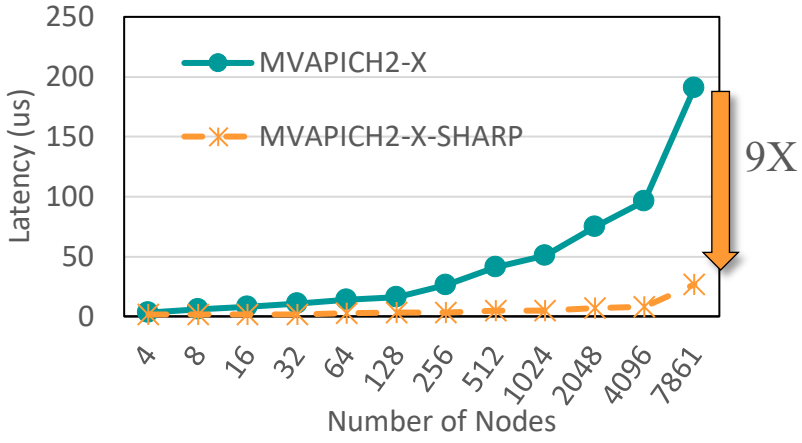


MVAPICH2-X – Advanced MPI + PGAS + Tools

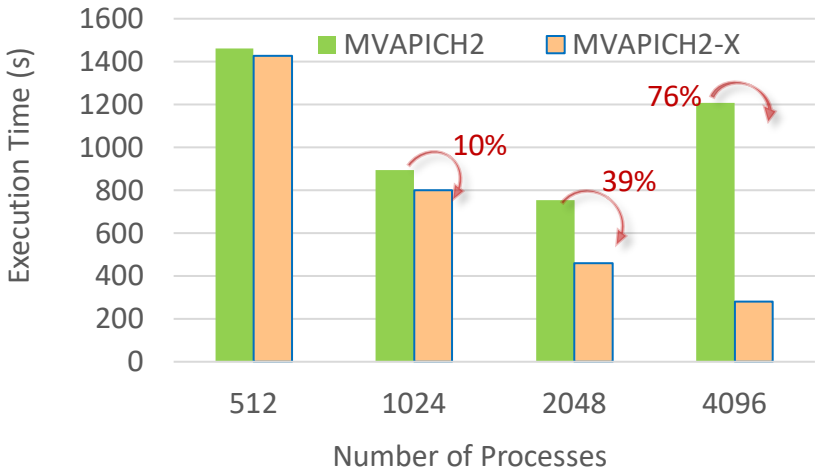
MPI_Allreduce using SHARP on Frontera (1ppn, 7,861 nodes)



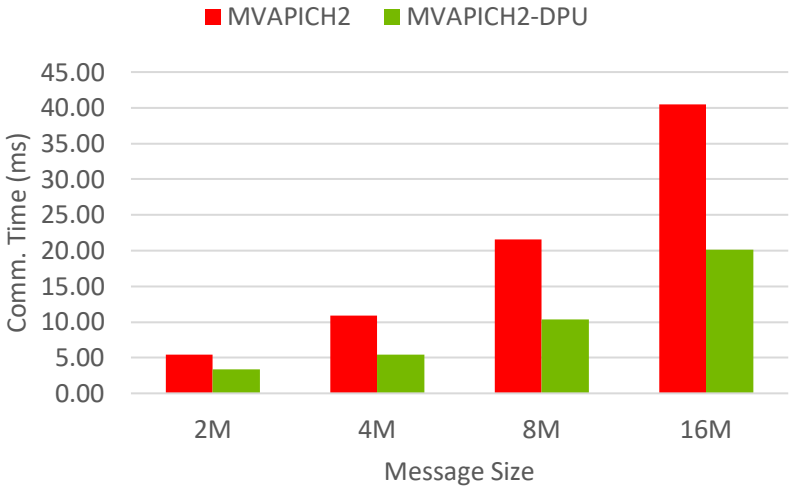
MPI_Barrier using SHARP on Frontera (1ppn, 7,861 nodes)



Impact of Transport Protocol Selection

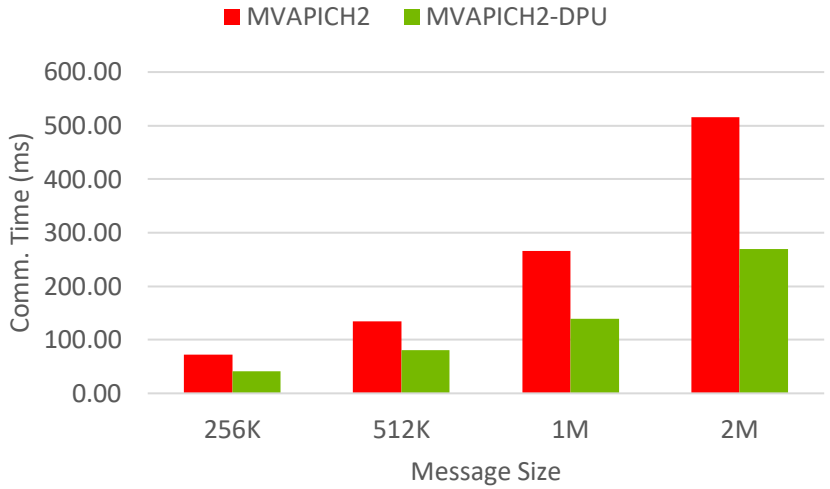


Total Execution Time, BF-2 (osu_ibcast)



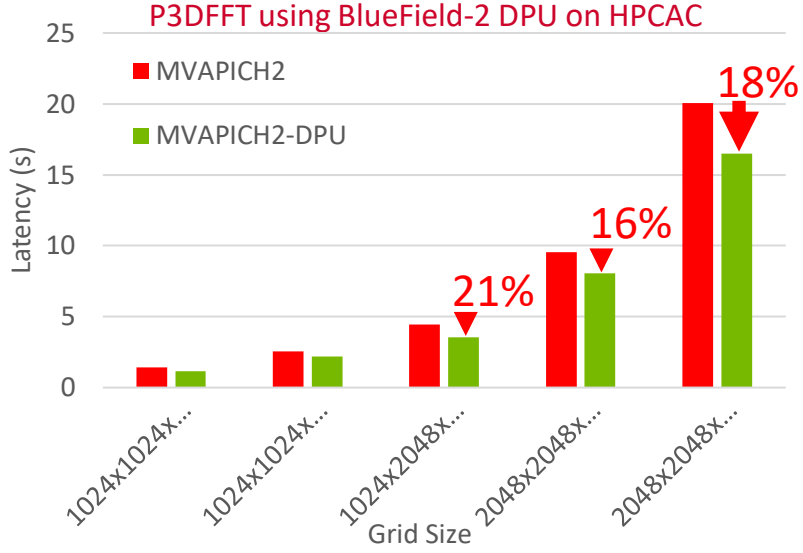
16 Nodes, 32 PPN

Total Execution Time, BF-2 (osu_iallgather)



16 Nodes, 16 PPN

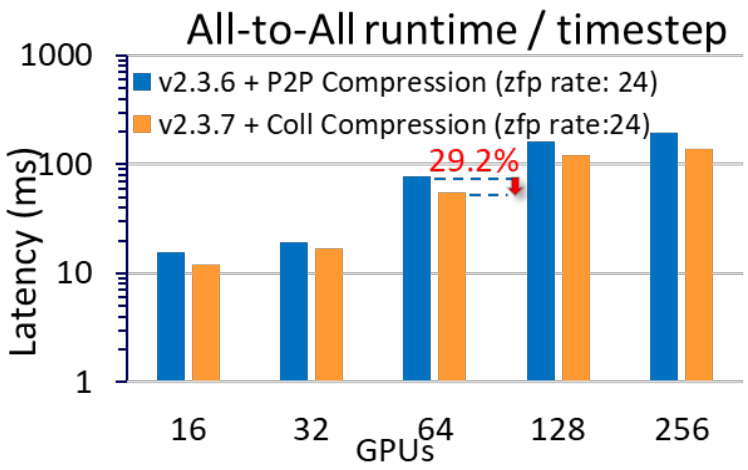
P3DFFT using BlueField-2 DPU on HPCAC



32 Nodes, 32 PPN

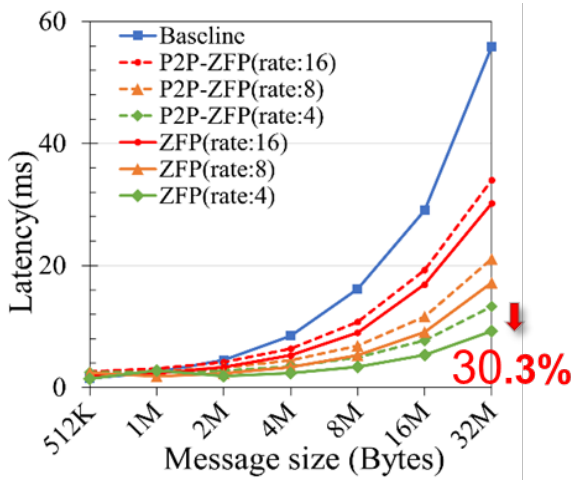
MVAPICH2-GDR – On-the-Fly GPU-based Compression

Performance of AlltoAll with Compression



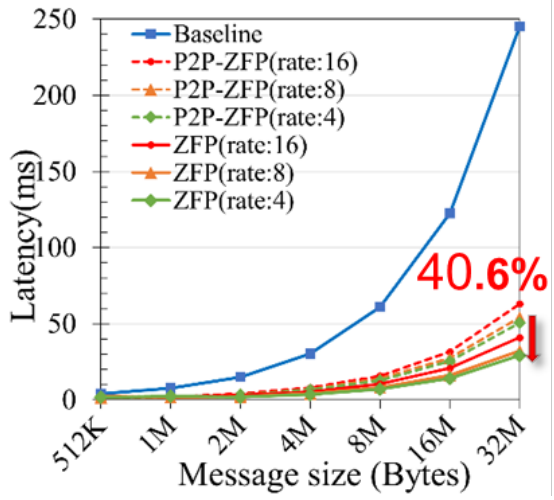
3D-FFT kernel of PSDNS on Lassen (V100)

Performance of Allgather with Compression

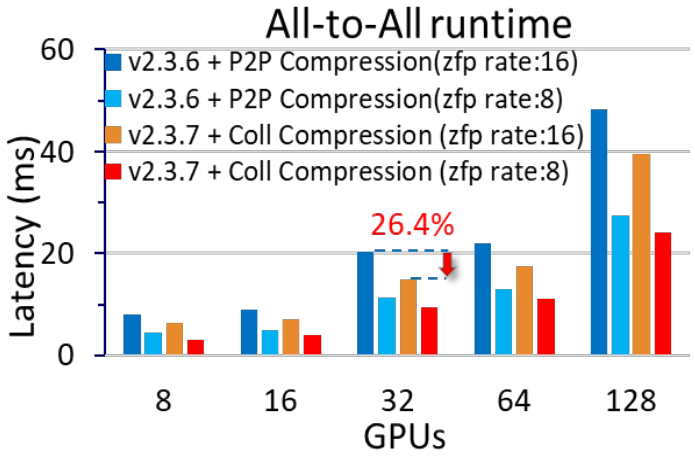


MPI_Allgather on Longhorn (4N4ppn, V100)

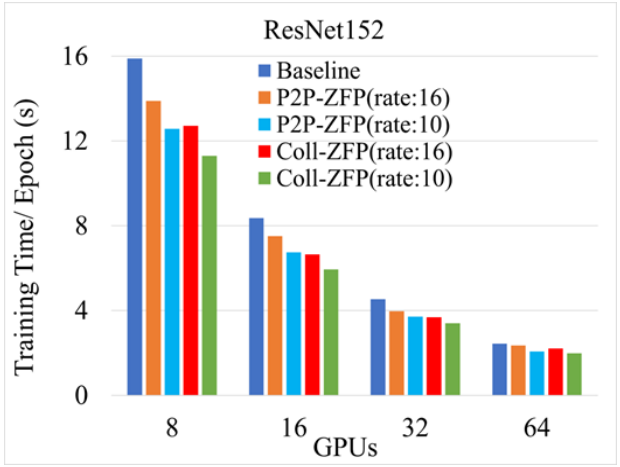
Performance of Reduce-Scatter with Compression

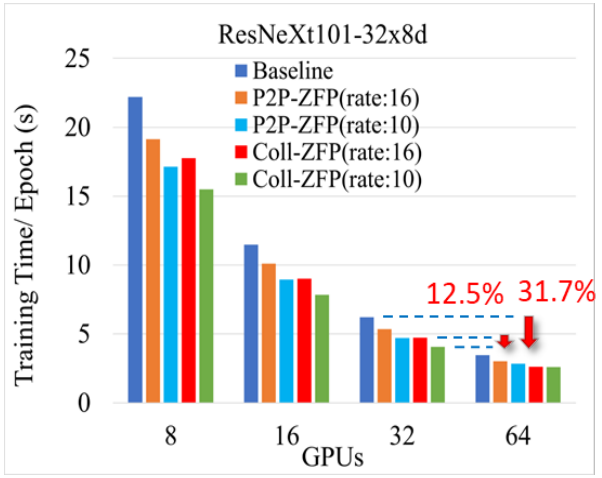


MPI_Reduce_scatter on Longhorn (4N4ppn, V100)



DeepSpeed Benchmark on Lassen (V100)

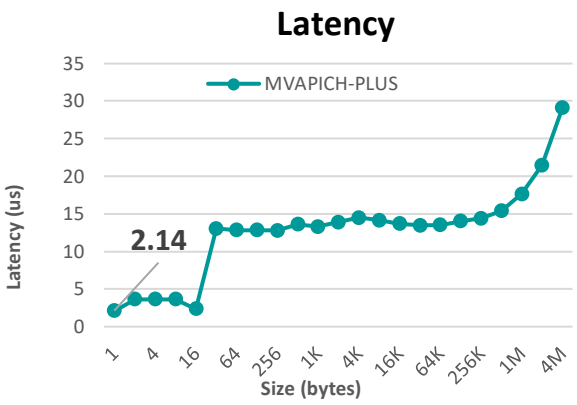


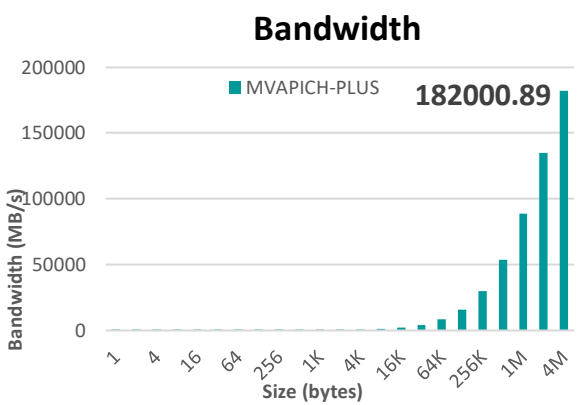


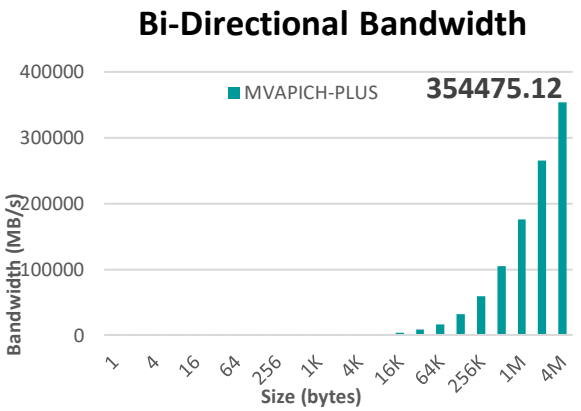
PyTorch FSDP Training with MPI_Allgather and MPI_Reduce_scatter on Longhorn (V100)

MVAPICH-PLUS – NVIDIA GPU Performance + IB

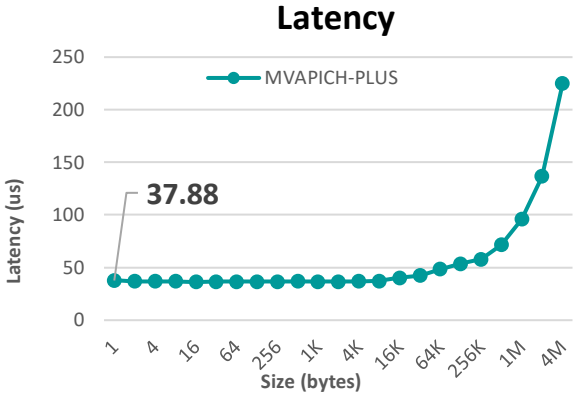
Intra-Node Point-to-Point

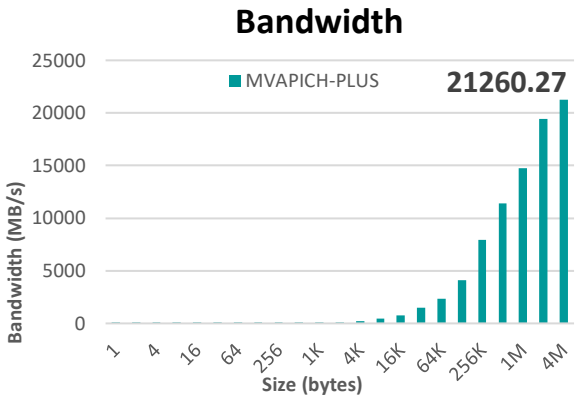


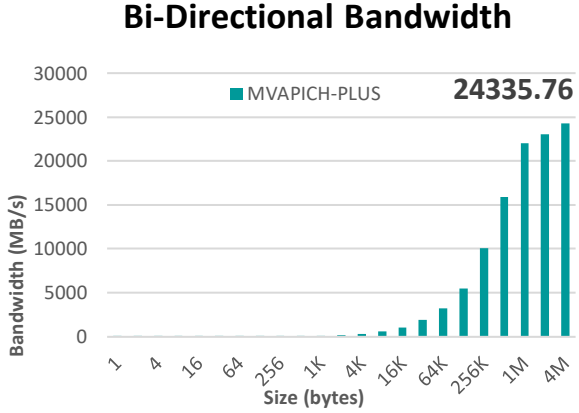




Inter-Node Point-to-Point

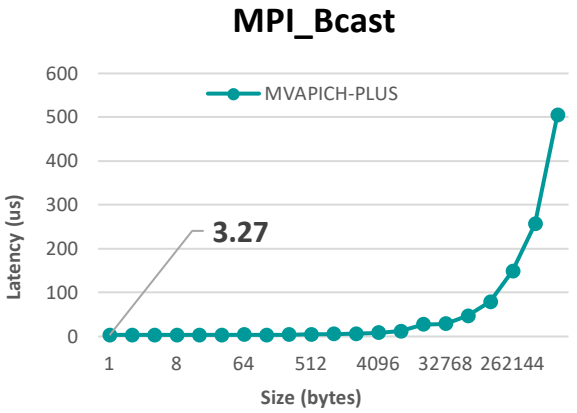


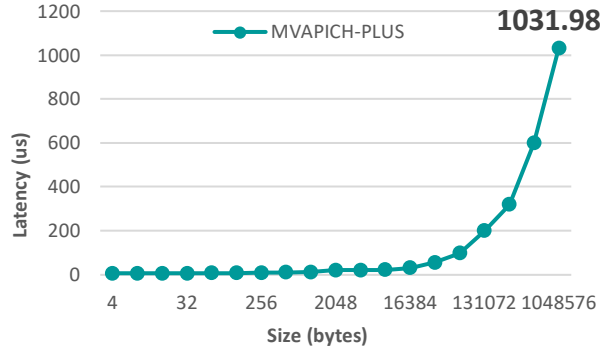




NVIDIA A100 GPUs, InfiniBand Networking, and CUDA 11.5 (ThetaGPU - ALCF)

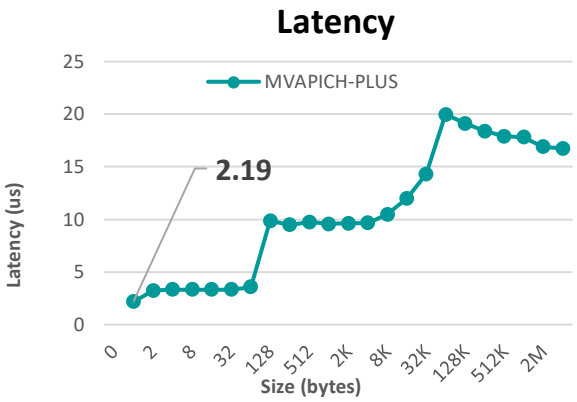
MPI Collectives – 16 GPUs

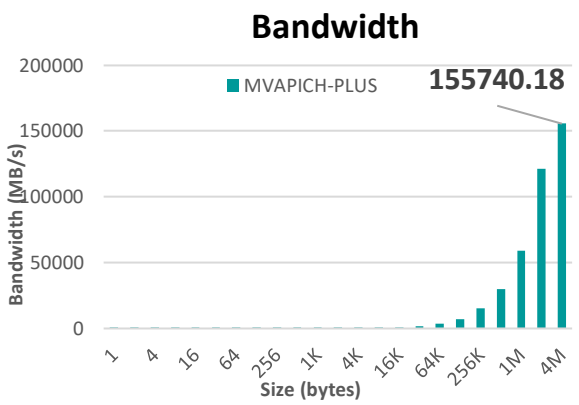


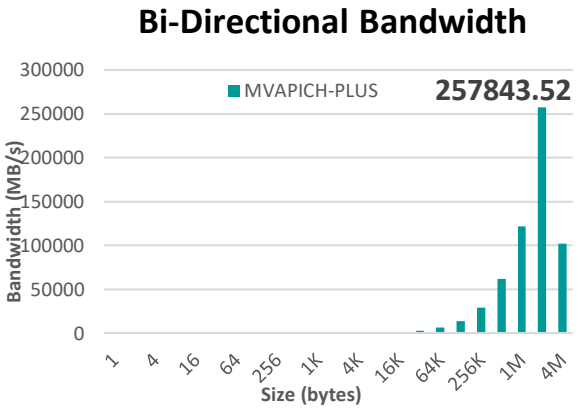


MVAPICH-PLUS – AMD GPU Performance + IB

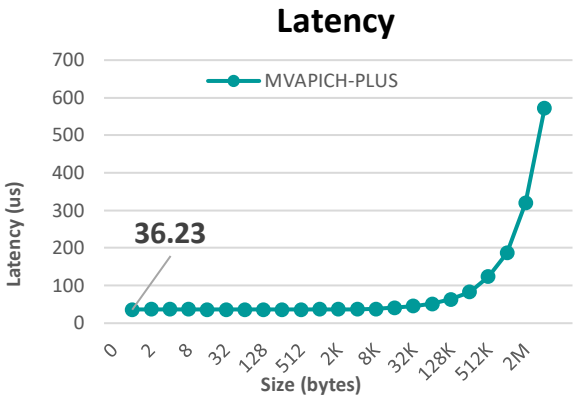
Intra-Node Point-to-Point

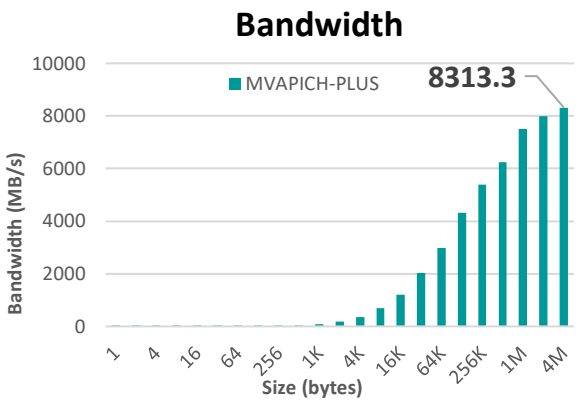


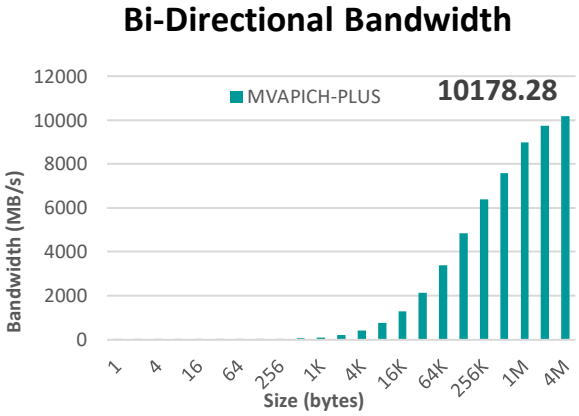




Inter-Node Point-to-Point

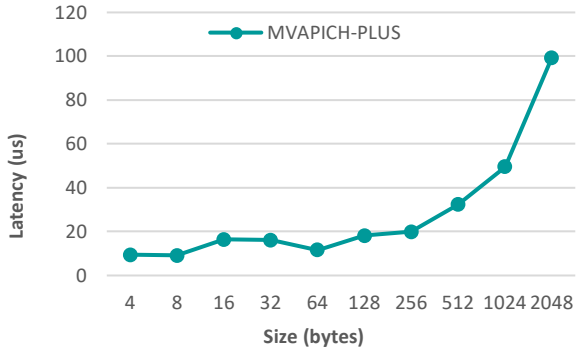




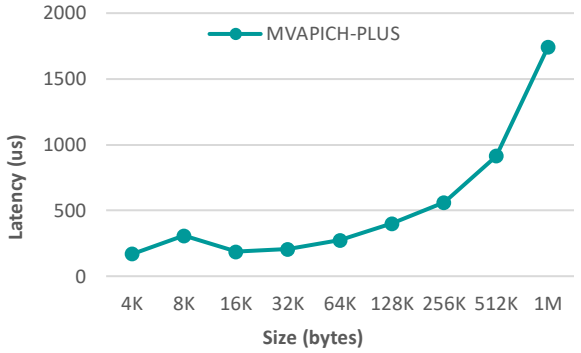


AMD MI-100 GPUs, InfiniBand Networking, and ROCm 5.6.0 (MRI)

MPI_Allreduce – 16 GPUs



Small Message Allreduce

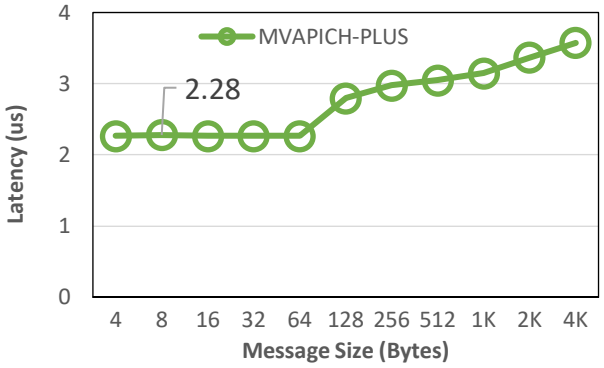


Large Message Allreduce

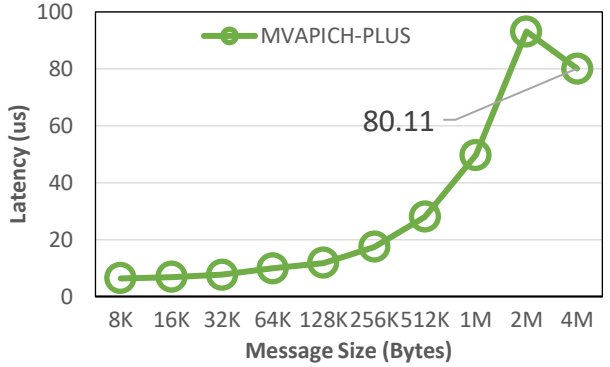
MVAPICH-PLUS – GPU Performance + Slingshot-11

Inter-Node Point-to-Point

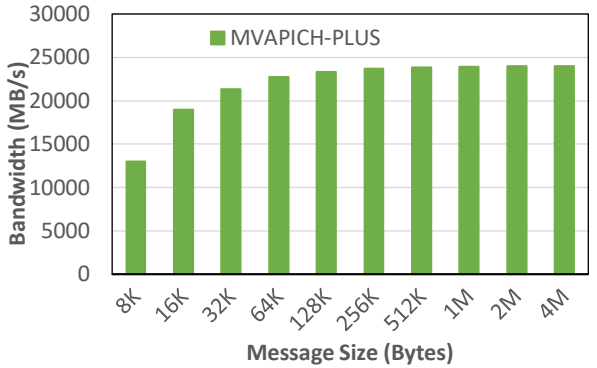
Small Message Latency



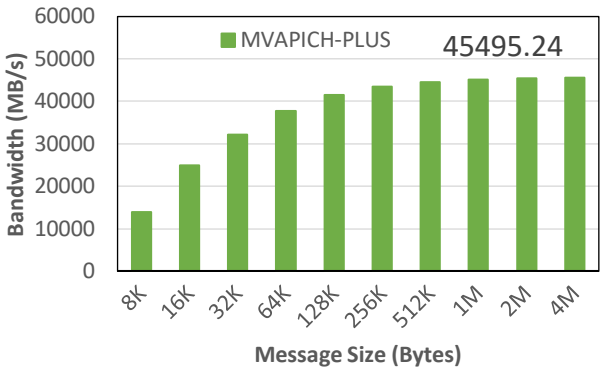
Large Message Latency



Bandwidth

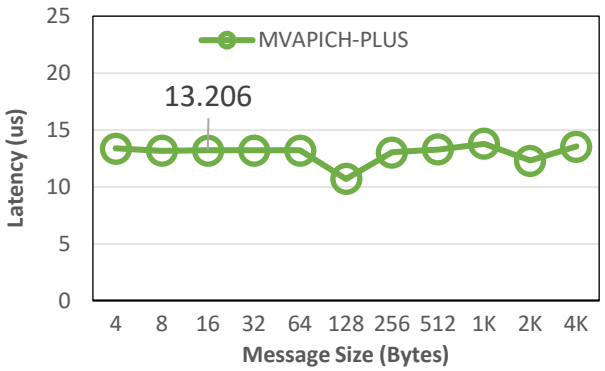


Bi-Directional Bandwidth

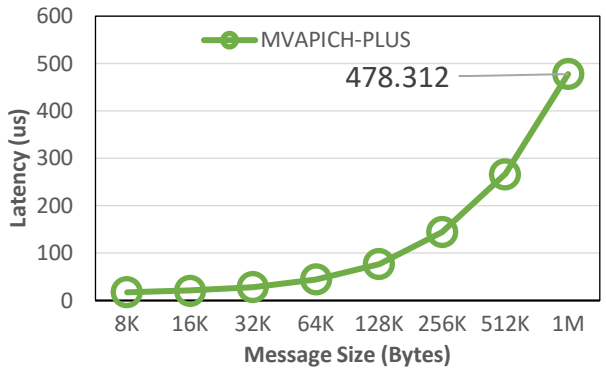


MPI Collectives – 32 GPUs (4 Nodes)

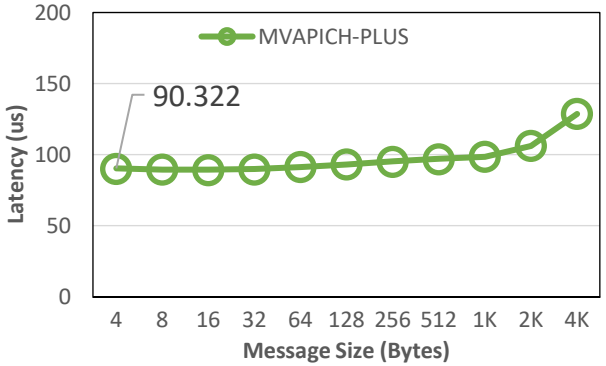
Small Message Broadcast



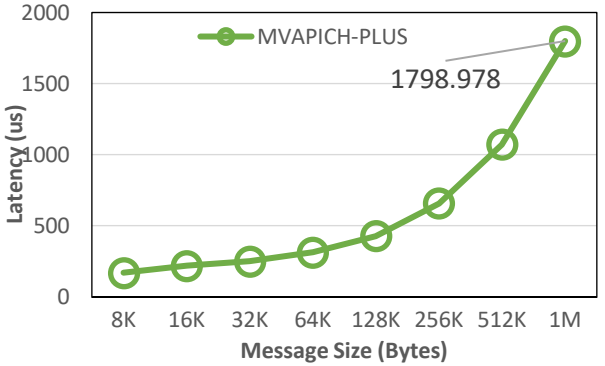
Large Message Broadcast



Small Message Allreduce



Large Message Allreduce



AMD MI250X GPUs, Slingshot-11 Networking, ROCm-5.6.0 (Tioga - LLNL)

MVAPICH – Moving to MPI 4.1

- MPI 4.1 released early this month
  - 4.0 released last year
- MPICH support for MPI 4.1 coming soon
- MVAPICH total re-write for MPICH 4.2+
- Migrating MVAPICH to a "rolling update" model to keep up with the latest features in MPICH
  - Regular syncing with upstream MPICH to keep us current with the community and the standard
  - Regular attendance at MPICH developer meetings
- Integrated optimizations that "play nice" with existing designs
- Enhanced shared memory optimizations, tuning, and new IB netmod leveraging 23 years of experience in IB networking

MVAPICH 4.0 Release Timeline

- Q1 2024 – MVAPICH-Plus 4.0 Initial version
  - MPI 4.1 support + Key MVAPICH designs
  - Shared memory optimizations for pt2pt and collectives
    - CMA/XPMEM aware collectives
  - Enhanced GPU support
    - Intra-node optimizations
    - On the fly compression
  - Enhanced collective tuning
  - New, next-gen job launcher/process manager
- Late Q1 2024 – MVAPICH 4.0 Initial version
  - Subset of MVAPICH-Plus features
- Q2 2024 - MVAPICH-Plus 4.1
  - Improved IB netmod
  - GPU Direct RDMA support
  - Further intra-node enhancements
    - Loopback, DMABuff
  - Adaptive tuning

Acknowledgments to all the Heroes (Past/Current Students and Staffs)

Current Students (Graduate)

- N. Alnaasan (Ph.D.)
- Q. Anthony (Ph.D.)
- C.-C. Chun (Ph.D.)
- N. Contini (Ph.D.)
- T. Chen (Ph.D.)
- J. Hatef (Ph.D.)
- H-R. Huang (Ph.D)
- P. Kousha (Ph.D.)
- S. Lee (Ph.D.)
- B. Michalowicz (Ph.D.)
- A. Potlapally (Ph. D.)
- B. Ramesh (Ph.D.)
- K. K. Suresh (Ph.D.)
- A. H. Tu (Ph.D.)
- S. Xu (Ph.D.)
- Q. Zhou (Ph.D.)
- K. Al Attar (Ph.D.)
- L. Xu (Ph.D.)
- G. Kuncham (Ph.D.)
- R. Vaidya (M.S.)
- J. Yao (P.hD.)
- M. Han (M.S.)

Current Research Scientists

- M. Abduljabbar
- A. Shafi

Past Students (Undergrads)

- J. Sulewski
- T. Chen

Current Faculty

- H. Subramoni

Current Software Engineers

- N. Pavuk
- N. Shineman
- M. Lieber
- A. Gupta

Current Research Specialist

- R. Motlagh

Past Students

- A. Awan (Ph.D.)
- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- M. Bayatpour (Ph.D.)
- R. Biswas (M.S.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- S. Chakraborthy (Ph.D.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- C.-H. Chu (Ph.D.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- A. Guptha (M.S.)
- J. Hashmi (Ph.D.)
- W. Huang (Ph.D.)
- A. Jain (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- K. S. Khorassani (Ph.D.)
- M. Kedia (M.S.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- M. Li (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- K. Raj (M.S.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- N. Sarkauskas (B.S. and M.S)
- N. Senthil Kumar (M.S.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Srivastava (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)
- J. Zhang (Ph.D.)

Past Research Scientists

- K. Hamidouche
- S. Sur
- X. Lu

Past Senior Research Associate

- J. Hashmi

Past Programmers

- A. Reifsteck
- D. Bureddy
- J. Perkins
- B. Seeds

Past Post-Docs

- D. Banerjee
- X. Besseron
- M. S. Ghazimirsaeed
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- K. Manian
- S. Marcarelli
- A. Ruhela
- J. Vienne
- H. Wang

Past Research Specialist

- M. Arnold
- J. Smith

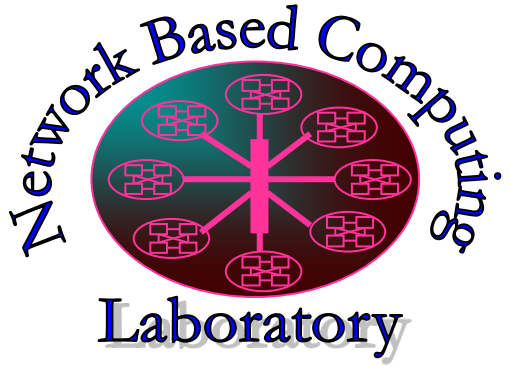
Join us for Multiple Events at SC '23

- Presentations at OSU and X-Scale Booth (#1581)
  - Members of the MVAPICH, HiBD and HiDL members
  - External speakers
- Presentations at SC main program (Tutorials, Workshops, BoFs, Posters, and Doctoral Showcase)
- Presentation at many other booths and satellite events
- Complete details available at

  http://mvapich.cse.ohio-state.edu/conference/964/talks/

Thank You!

subramon@cse.ohio-state.edu



Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

Follow us on

https://twitter.com/mvapich

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

HiBD
High-Performance
Big Data

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

HiDL
High-Performance
Deep Learning

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/