MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC'22)

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)

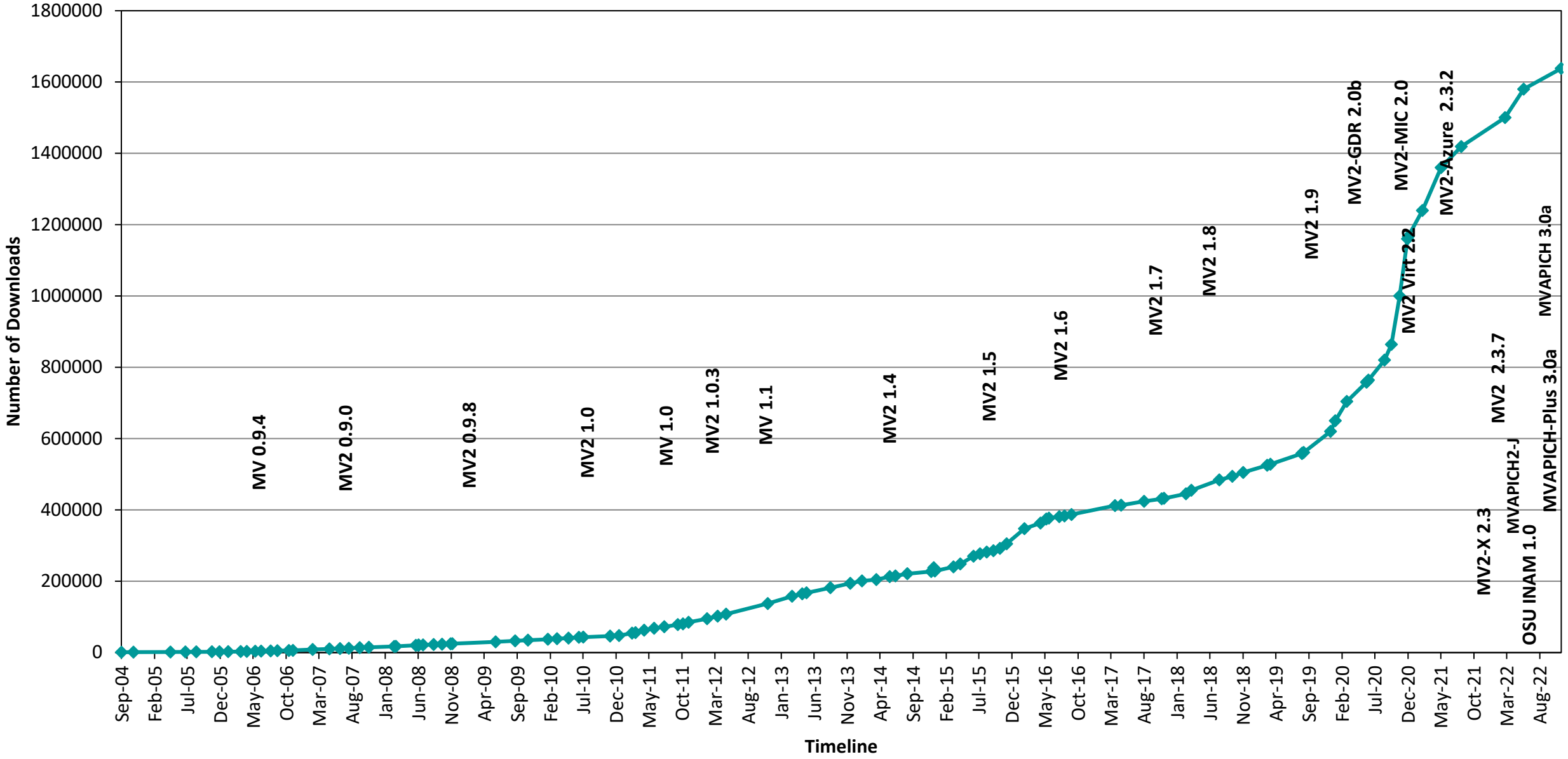
MPICH + MVICH + VAPI = MVAPICH

Overview of the MVAPICH2 Project

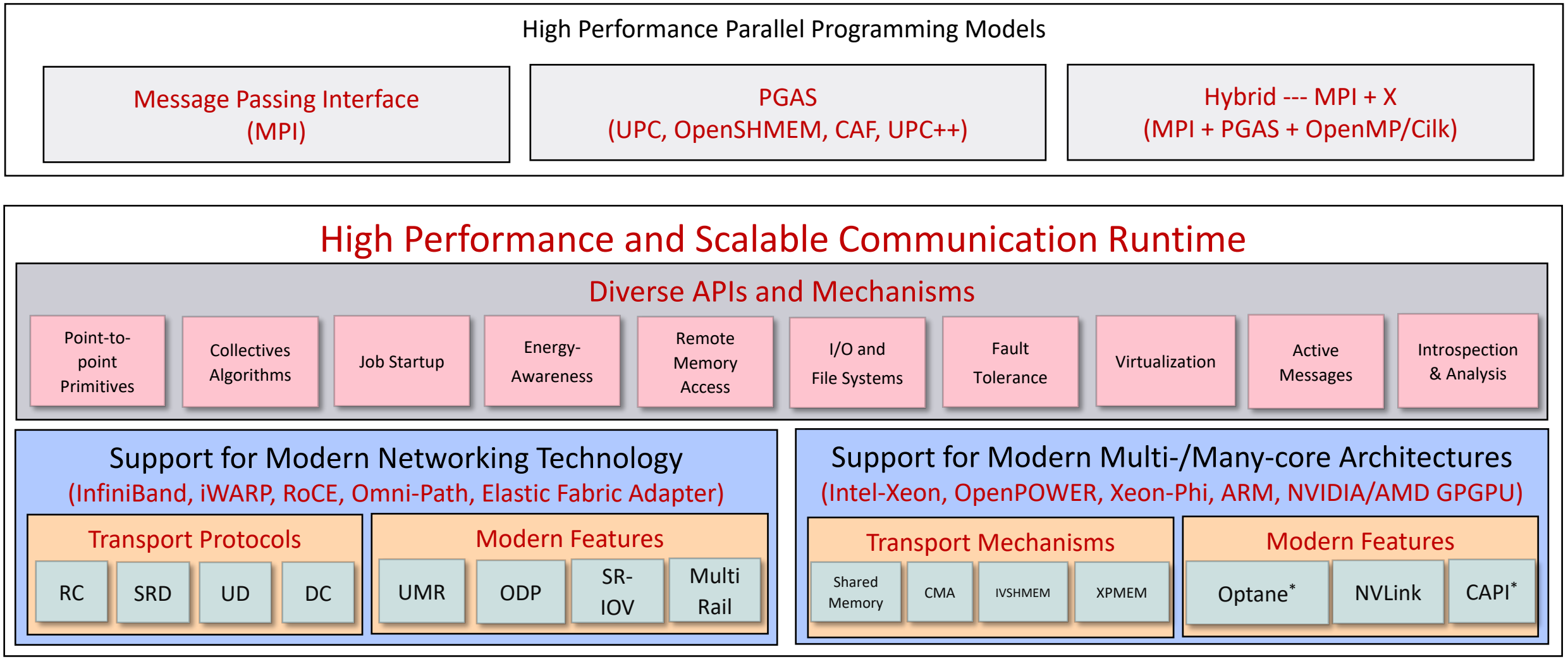
- High Performance open-source MPI Library
- Support for multiple interconnects
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, Rockport Networks, and Slingshot10/11, Broadcom, Cornelis Networks OPX
- Support for multiple platforms
  - x86, OpenPOWER, ARM, Xeon-Phi, 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 3,275 organizations in 90 countries
- More than 1.64 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '22 ranking)
  - 7th, 10,649,600-core (Sunway TaihuLight) at NSC in Wuxi, China
  - 19th, 448, 448 cores (Frontera) at TACC
  - 34th, 288,288 cores (Lassen) at LLNL
  - 46th, 570,020 cores (Nurion) in South Korea
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)
- Partner in the 16th ranked TACC Frontera system
- Empowering Top500 systems for more than 20 years

MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family for HPC and DL/ML



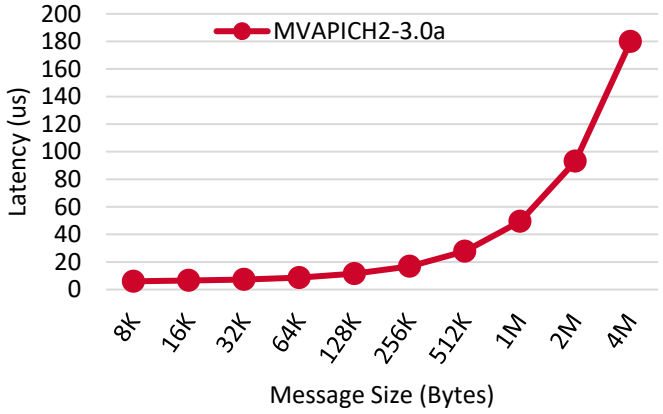
* Upcoming

MVAPICH2 Software Family

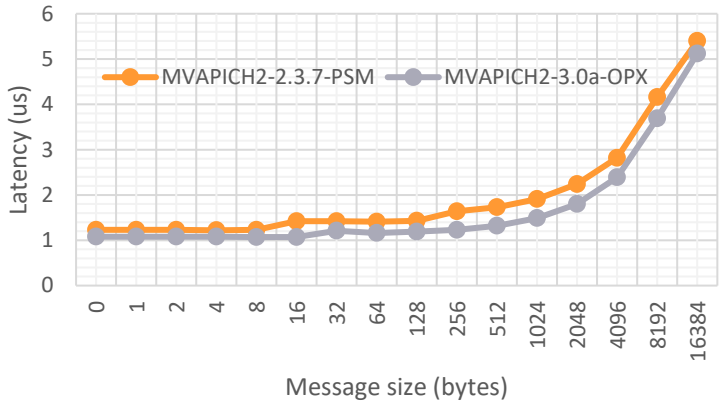
High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters 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

MVAPICH2-3.0a Point-to-Point

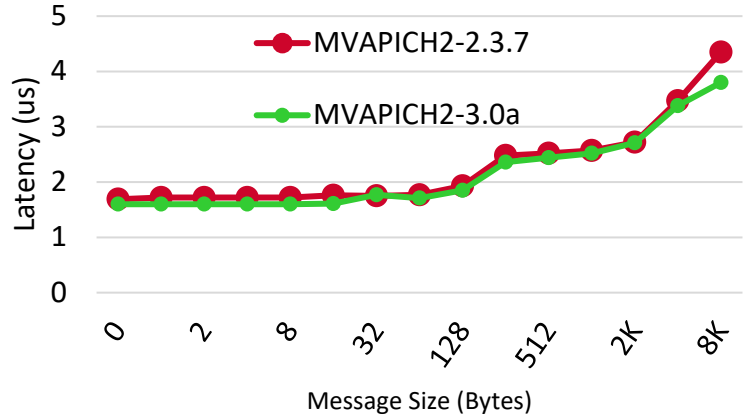
Latency on Slingshot 11



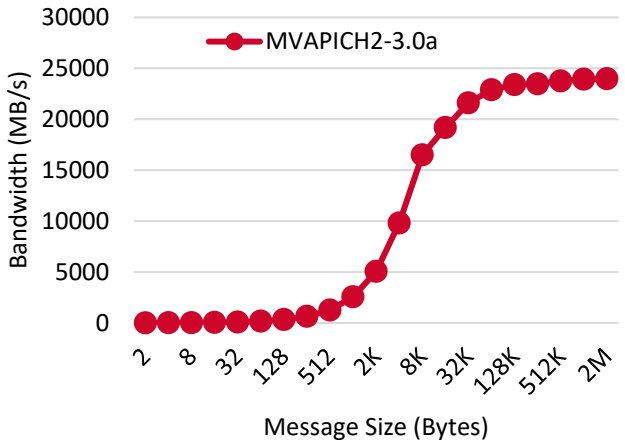
Latency on OPX



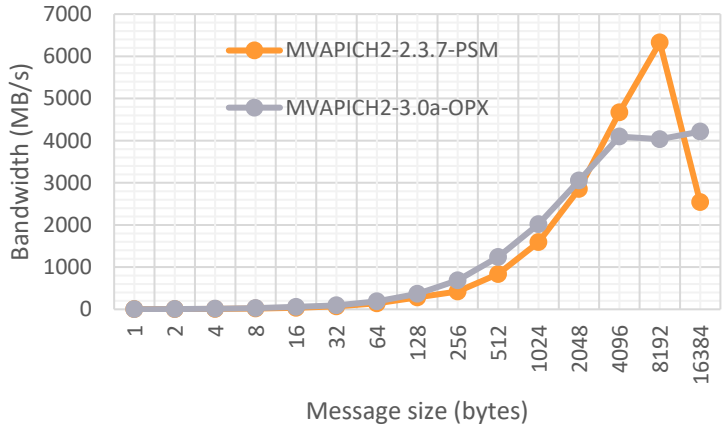
Latency on IB



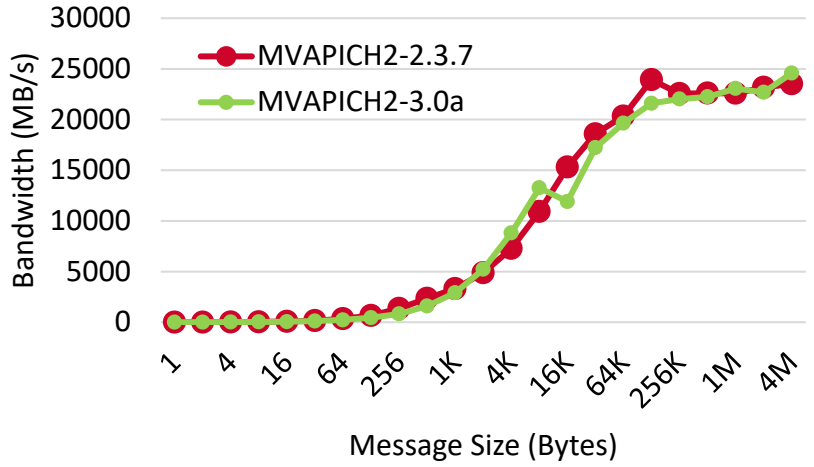
Bandwidth on Slingshot 11



Bandwidth on OPX

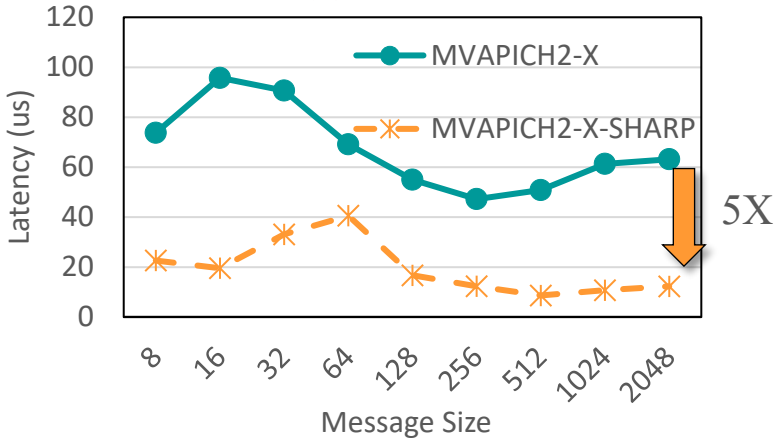


Bandwidth on IB

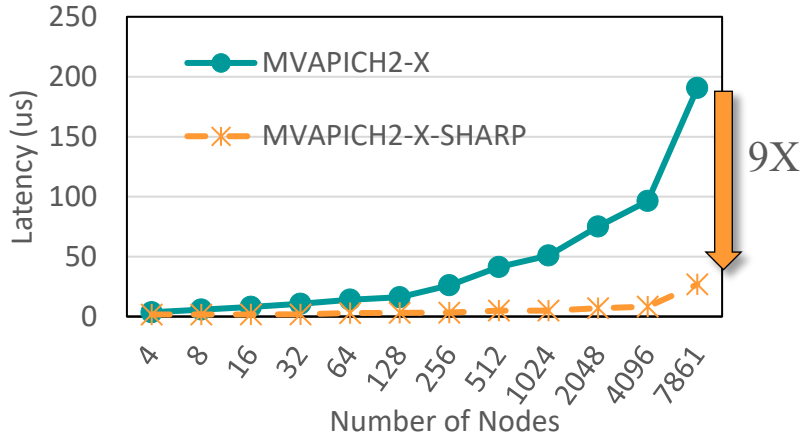


MVAPICH2-X – Advanced MPI + PGAS + Tools

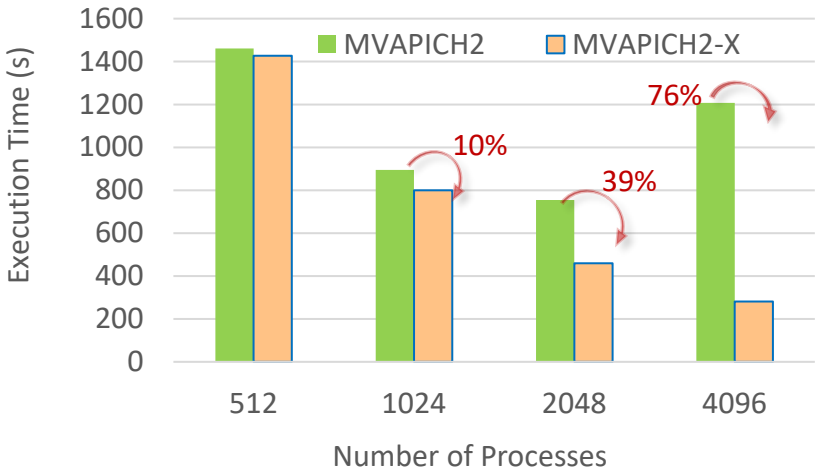
MPI_Allreduce using SHARP on Frontera (1ppn, 7,861 nodes)



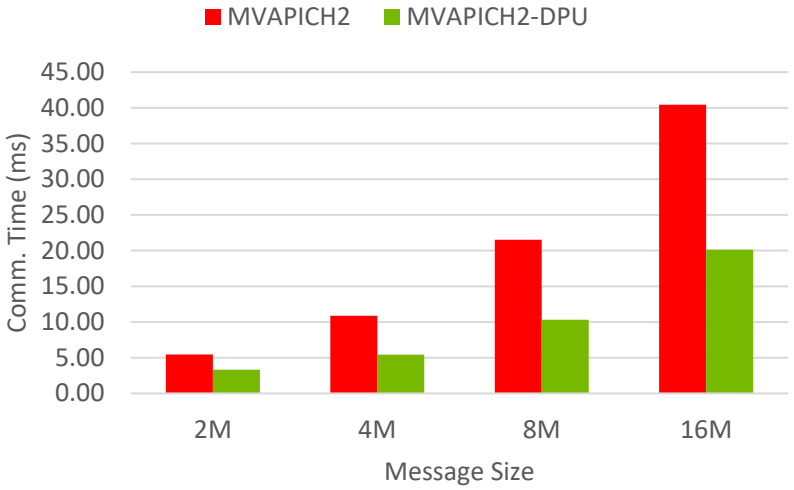
MPI_Barrier using SHARP on Frontera (1ppn, 7,861 nodes)



Impact of Transport Protocol Selection

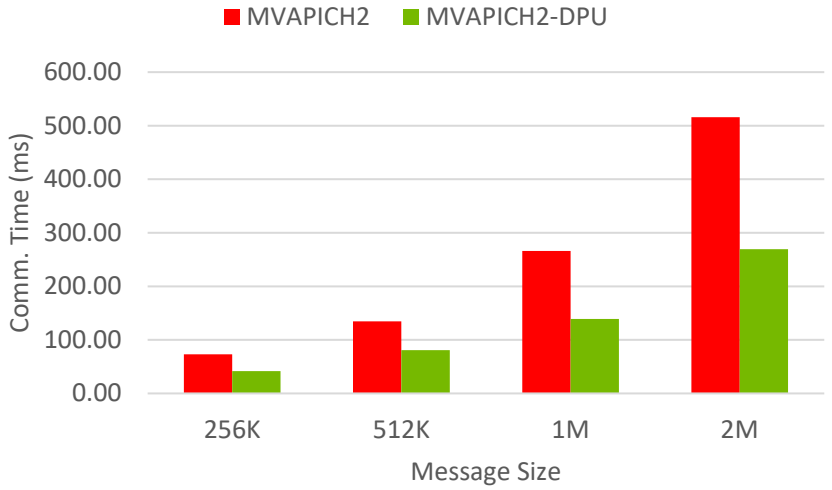


Total Execution Time, BF-2 (osu_ibcast)



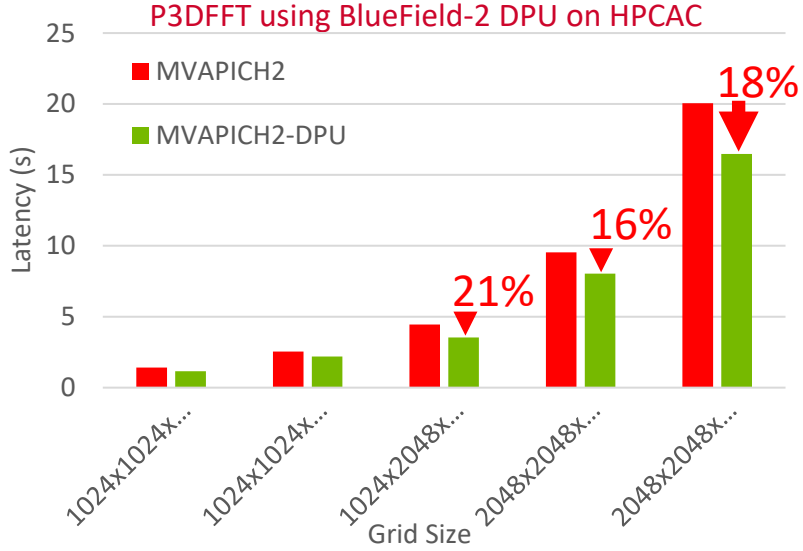
16 Nodes, 32 PPN

Total Execution Time, BF-2 (osu_iallgather)



16 Nodes, 16 PPN

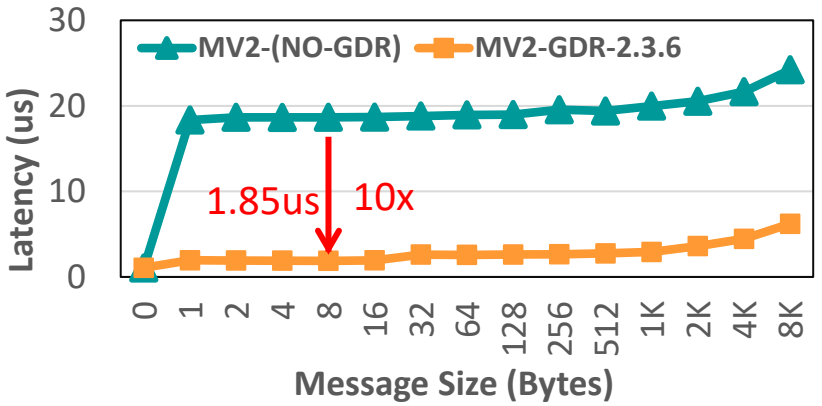
P3DFFT using BlueField-2 DPU on HPCAC



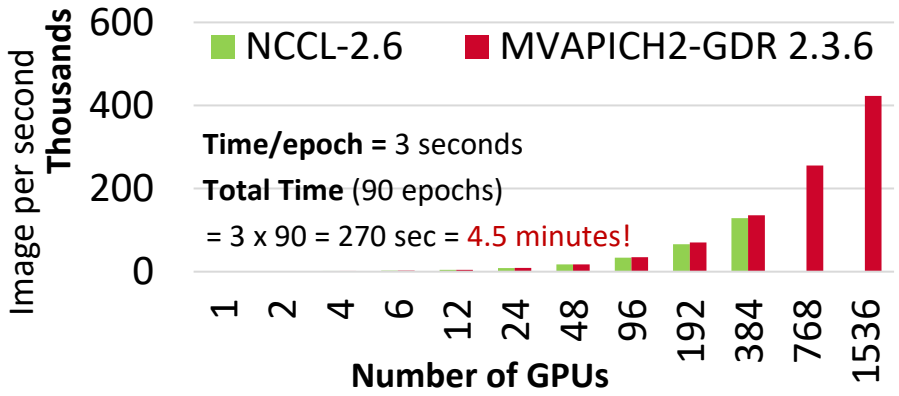
32 Nodes, 32 PPN

MVAPICH2-GDR – Optimized MPI for clusters with NVIDIA and AMD GPUs

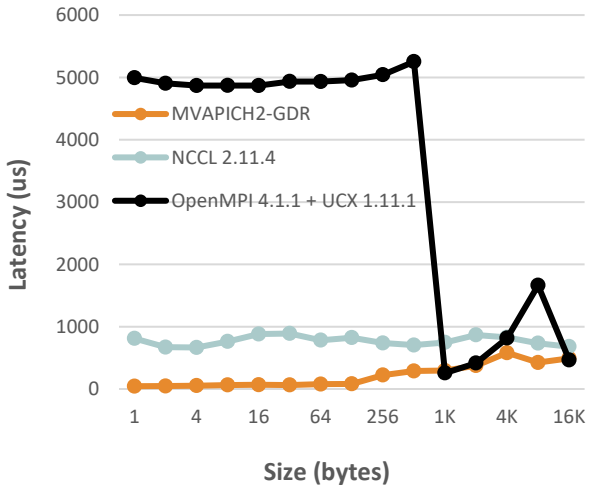
Best Performance for GPU-based Transfers



TensorFlow Training with MVAPICH2-GDR on Summit



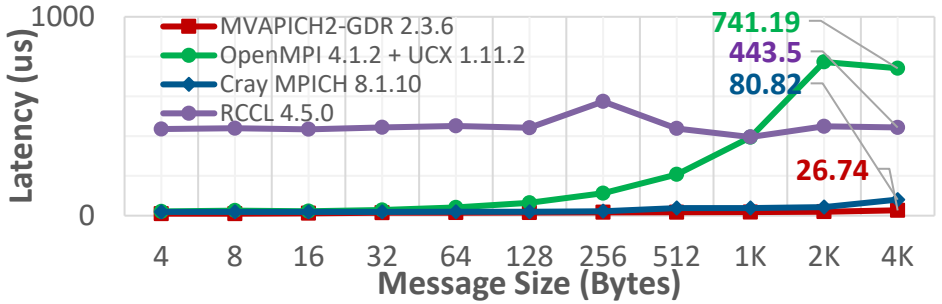
Enhanced Alltoall on DGX2-A100



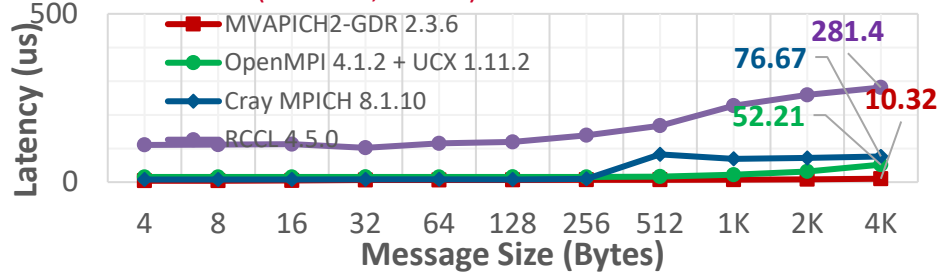
MVAPICH2-GDR on Slingshot-10 and AMD GPUs

OLCF Spock Cluster – ROCm 5.0 + MI100 GPUS (4 Nodes 4 PPN – 16 GPUs)

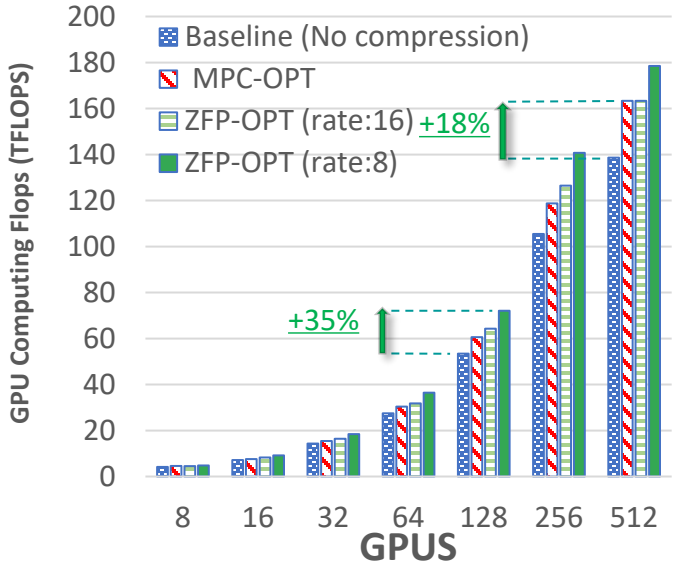
Allreduce 16 GPUs (4 Nodes, 4 GPN)



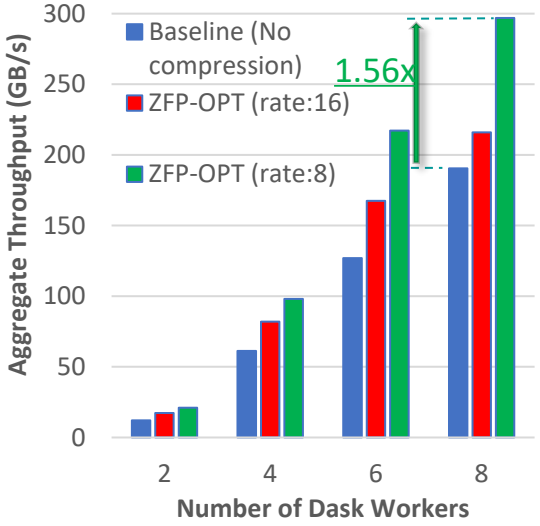
Broadcast 16 GPUs (4 Nodes, 4 GPN)



"On-the-fly" Compression Support (AWP-ODC Earthquake Sim App)



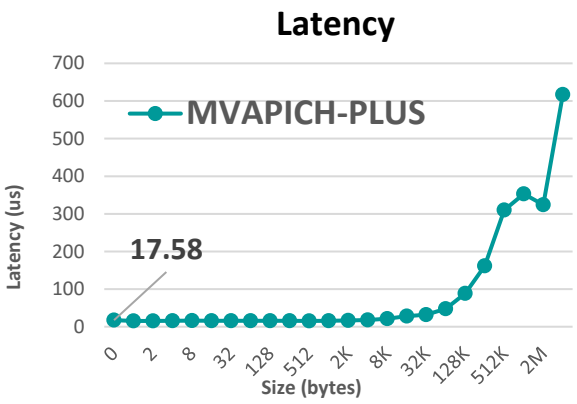
"On-the-fly" Compression Support (DASK for Data Science)

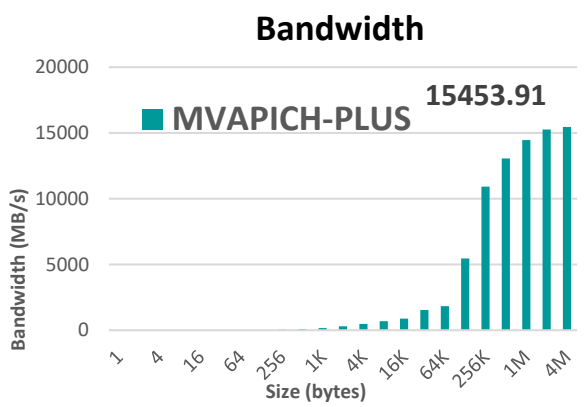


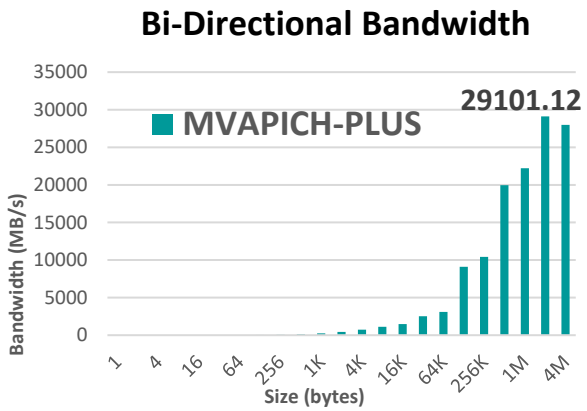
(cuPy Dims: 10Kx10K, Chunk size: 1K)

MVAPICH-PLUS – NVIDIA GPU Performance + IB

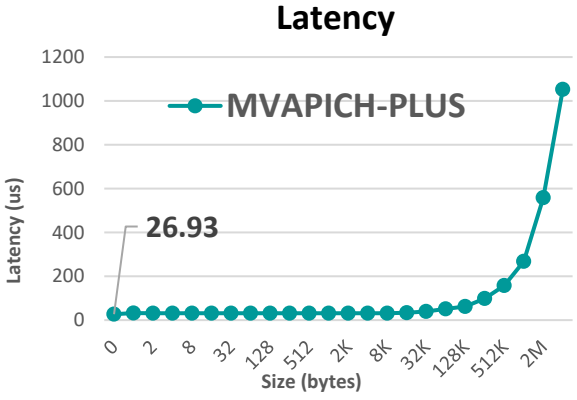
Intra-Node Point-to-Point

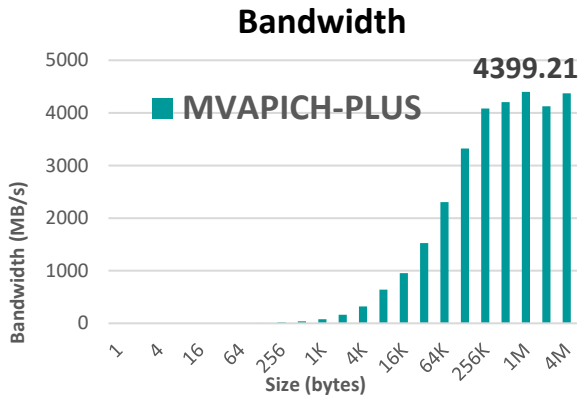


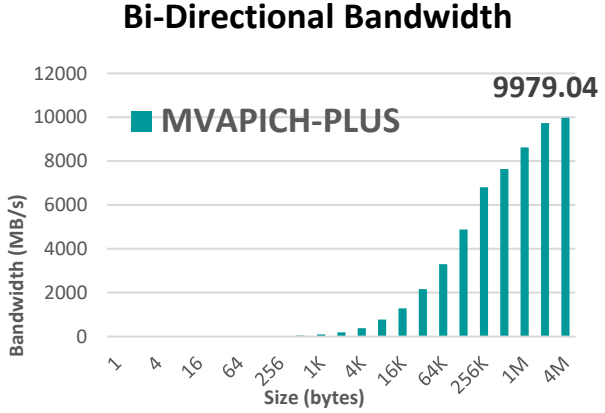




Inter-Node Point-to-Point

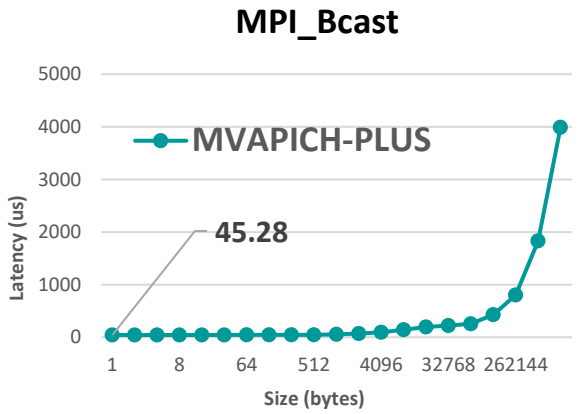


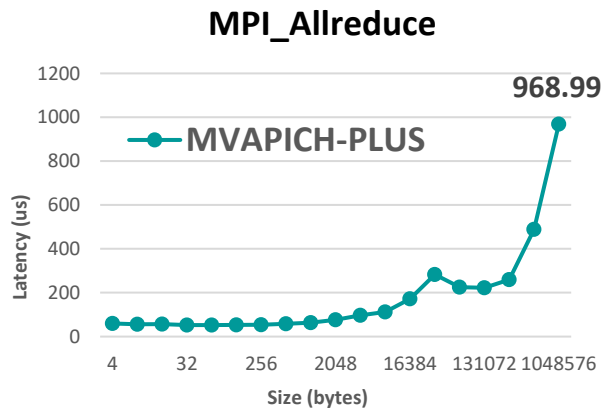




NVIDIA A100 GPUs, InfiniBand Networking, and CUDA 11.5

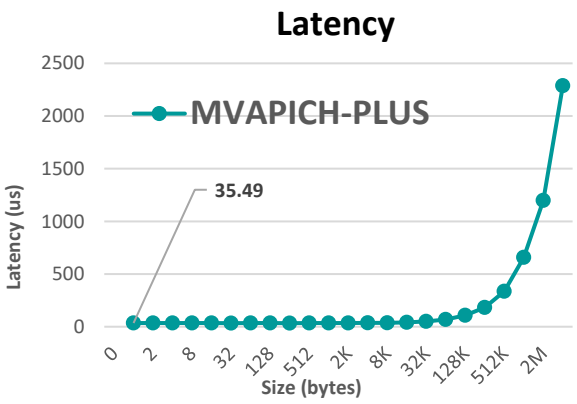
MPI Collectives – 16 GPUs

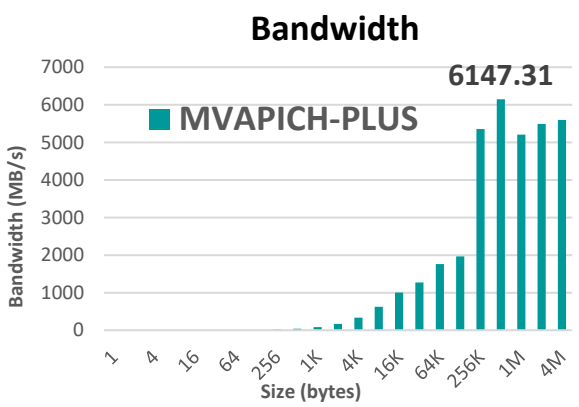


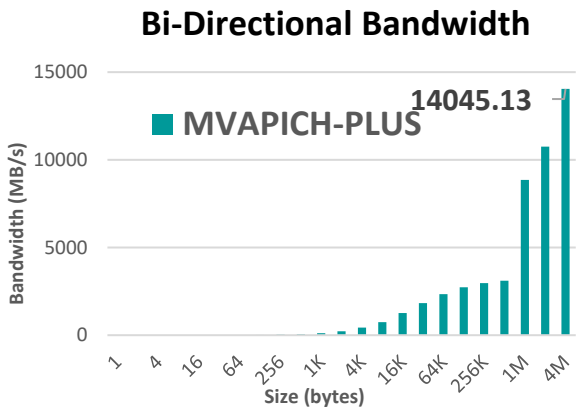


MVAPICH-PLUS – AMD GPU Performance + IB

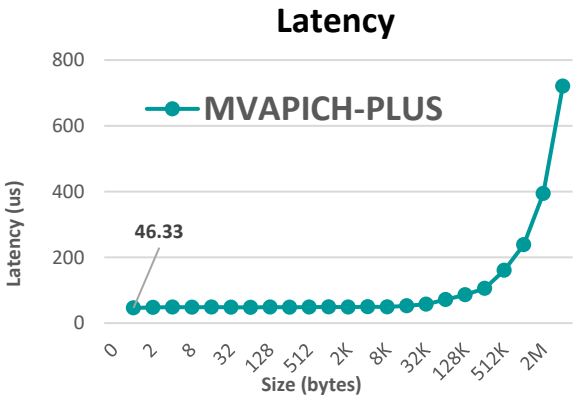
Intra-Node Point-to-Point

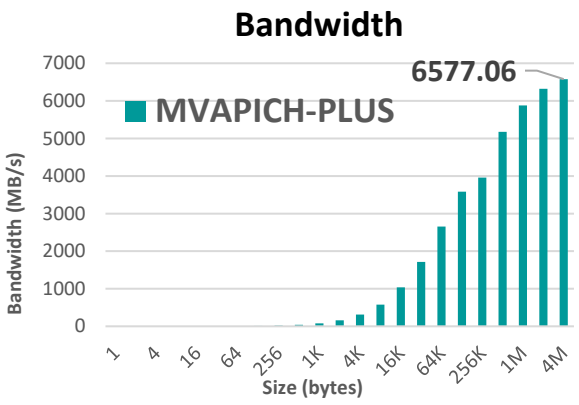


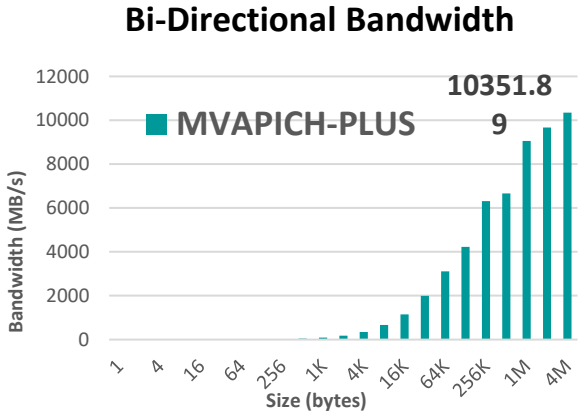




Inter-Node Point-to-Point

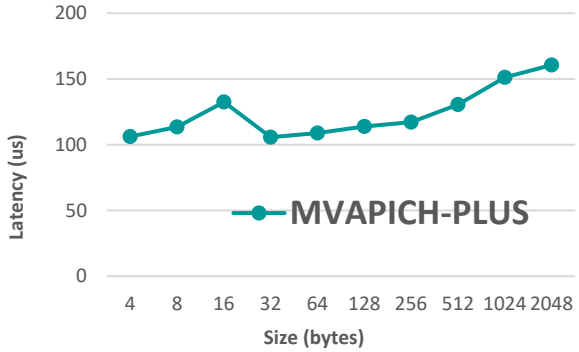




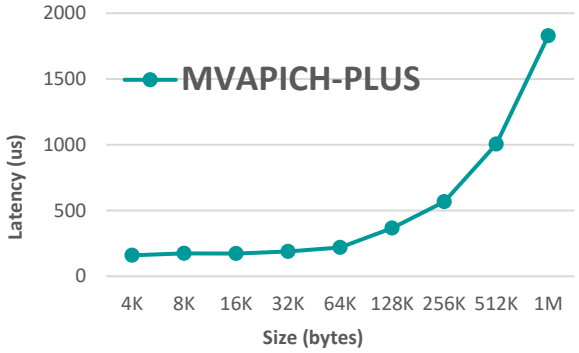


AMD MI-100 GPUs, InfiniBand Networking, and ROCm 5.1.3

MPI_Allreduce – 16 GPUs



Small Message Allreduce

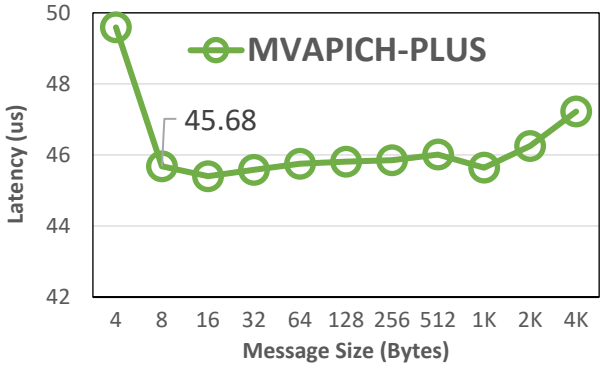


Large Message Allreduce

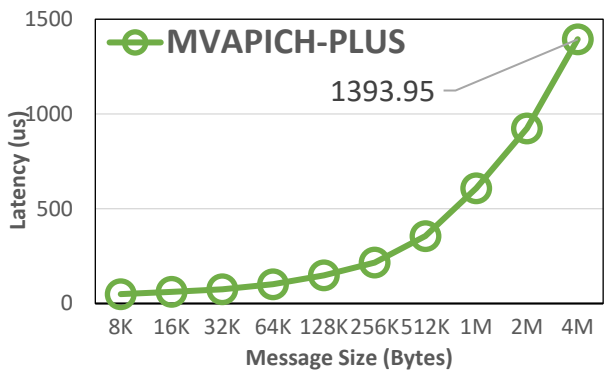
MVAPICH-PLUS – GPU Performance + Slingshot-11

Inter-Node Point-to-Point

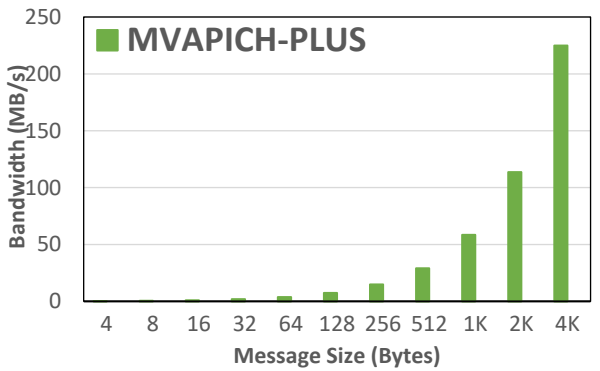
Small Message Latency



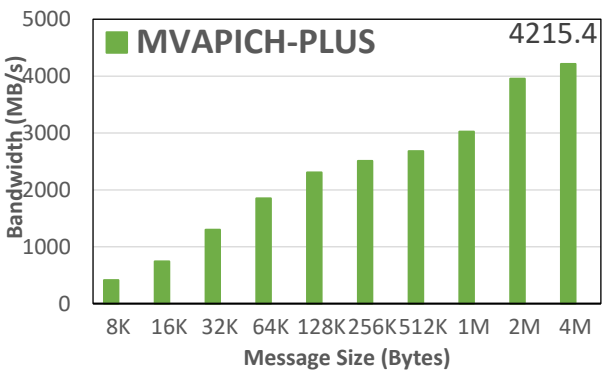
Large Message Latency



Bandwidth

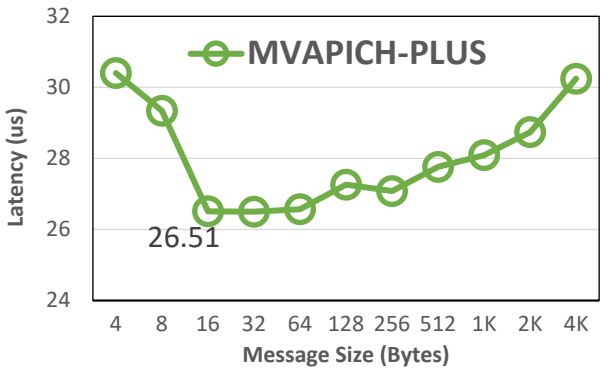


Bi-Directional Bandwidth

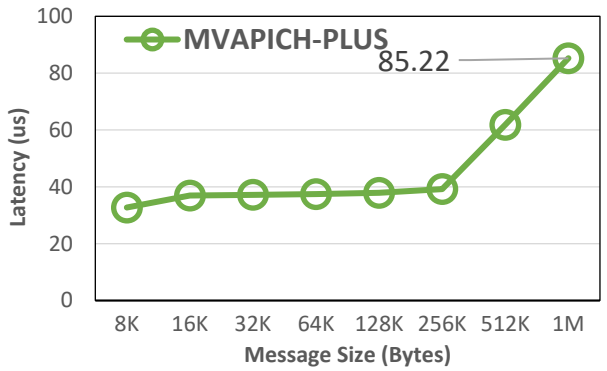


MPI Collectives – 8 GPUs

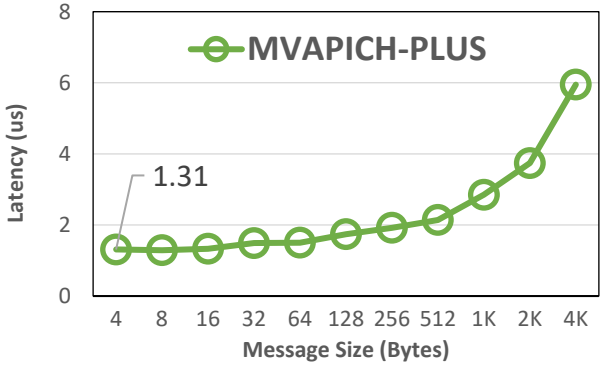
Small Message Broadcast



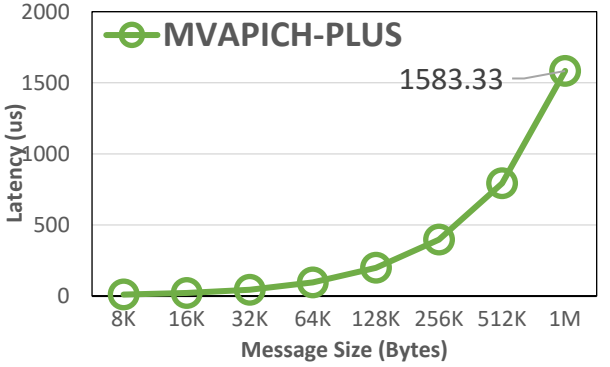
Large Message Broadcast



Small Message Allreduce



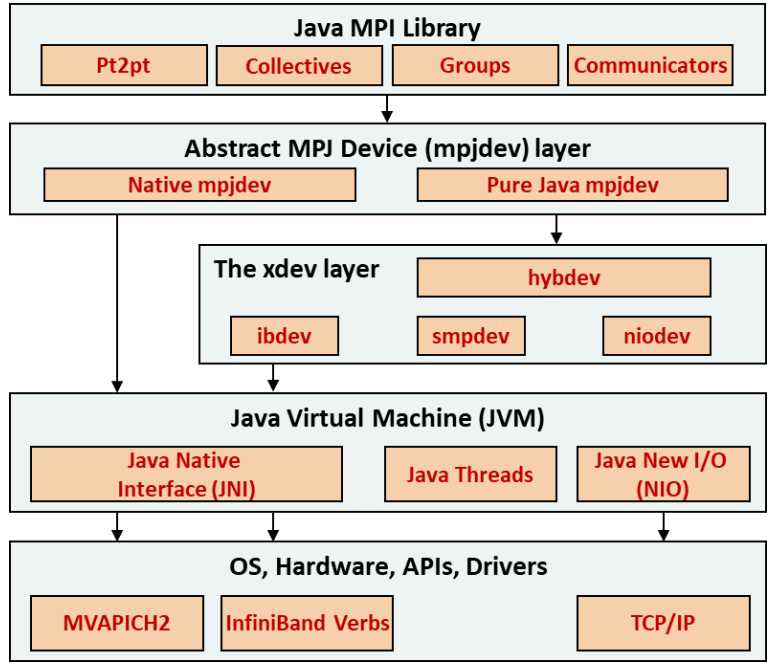
Large Message Allreduce



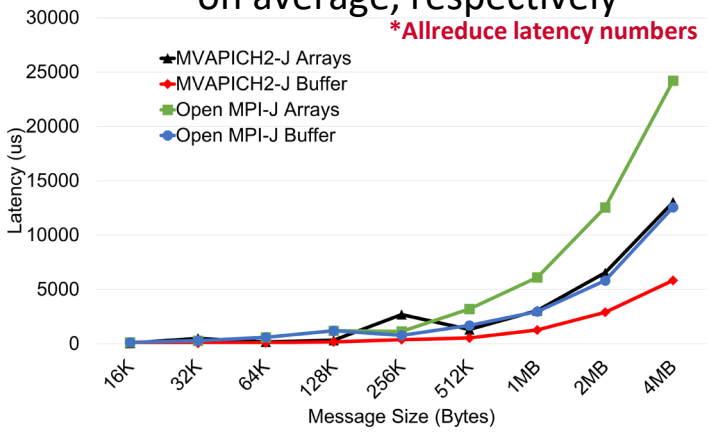
AMD MI250-X GPUs, Slingshot-11 Networking, ROCm-5.3.0

MVAPICH2 Java Bindings (MVAPICH2-J)

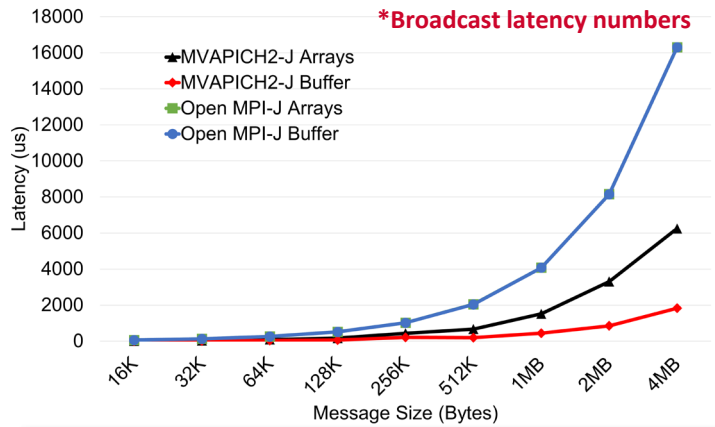
- MVAPICH2-J is an effort to produce Java bindings for the MVAPICH2 library
- Features
  - Provides Java bindings to the MVAPICH2 family of libraries
  - Support for communication of basic Java datatypes and Java new I/O (NIO) package direct ByteBuffers
  - Support for blocking and non-blocking point-point communication protocols
  - Support for blocking collective and strided collective communication protocols
- Results (Performance is evaluated against Open MPI's Java bindings)
  - Broadcast Performance
    - For both buffer and Java arrays, MVAPICH2-J outperforms by 6.2x and 2.2x on average, respectively
  - AllReduce Performance
    - For both buffer and Java arrays, MVAPICH2-J outperforms by 2.8x and 1.6x on average, respectively



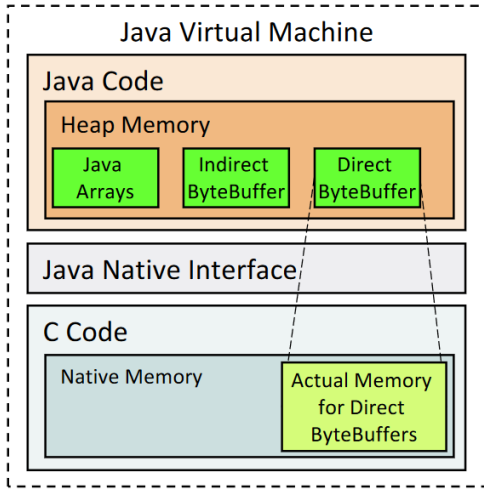
*Layered Architecture of the Java Bindings for the MVAPICH2 Library



On avg, MVAPICH2-J outperforms by 2.8x and 1.6x



On avg, MVAPICH2-J outperforms by 6.2x and 2.2x

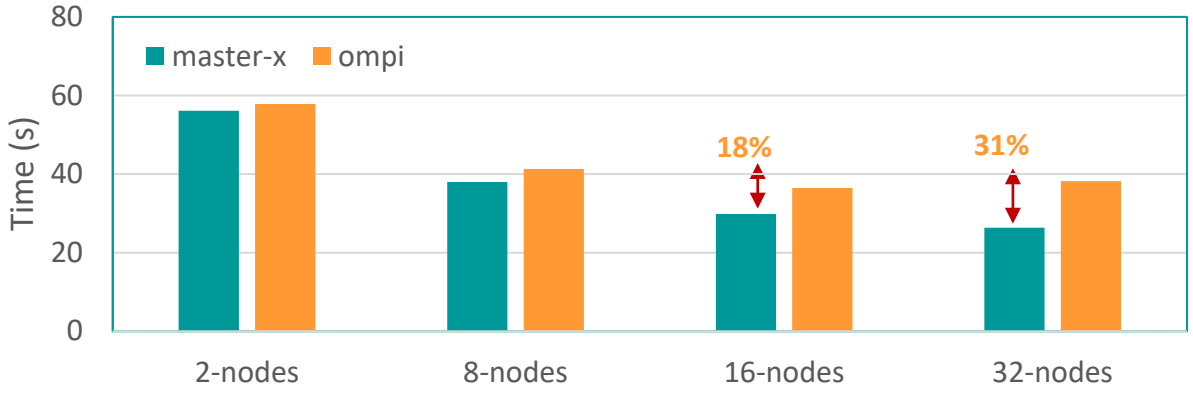


*The layout of Direct/Non-direct ByteBuffers and Java Arrays in the JVM

MVAPICH2-X Advanced Support for HPC-Clouds

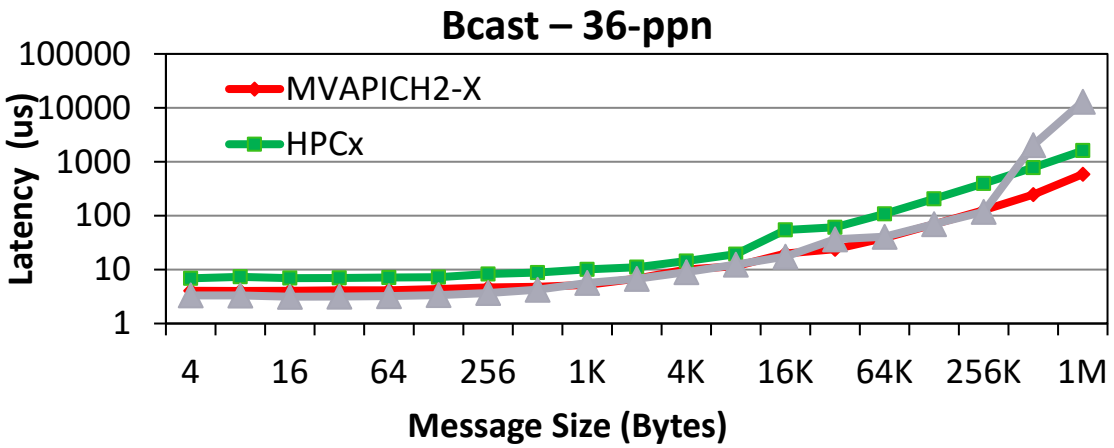
Performance on Amazon EFA

WRF 3.6 Execution time



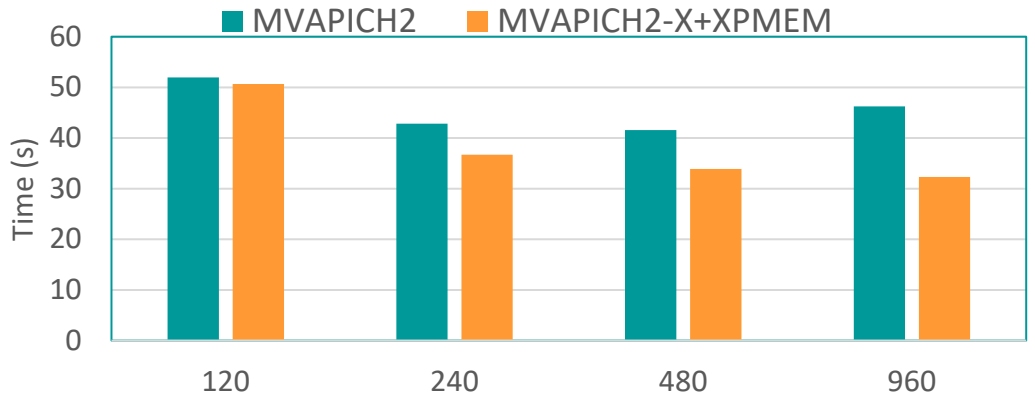
Instance type: c6gn.16xlarge
CPU: Amazon Graviton 2 @ 2.50GHz (64 cores per node)
MVAPICH2 version: MVAPICH2-X-AWS-2.3.7 (aarch64)
OpenMPI version: Open MPI v4.1.0 with libfabric 1.13.2

Performance on Oracle HPC Shapes



Performance of WRF on Microsoft Azure

WRF 3.6 Execution time



VM type: HBv2
CPU: AMD EPYC 7V12 @ 2.45GHz
MVAPICH2 version: MVAPICH2-Azure 2.3.3
MVAPICH2-X version: MVAPICH2-X (2.3rc3)

- Releases
  - MVAPICH2-X-AWS 2.3.7
  - MVAPICH2-Azure 2.3.6
  - Integrated Azure CentOS HPC Images:

https://github.com/Azure/azhpc-images/releases/tag/centos-hpc-20220112

MVAPICH2 – Future Roadmap and Plans for Exascale

- Making CH4 channel default
  - Early 2023
- Performance and Memory scalability toward 1M-10M cores
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - MPI + Task*
- Enhanced Optimization for GPUs and FPGAs*
- Taking advantage of advanced features of Mellanox InfiniBand
  - Tag Matching*
  - Adapter Memory*
- Enhanced communication schemes for upcoming architectures
  - NVLINK*
  - CAPI*
  - Bluefield2*
- Extended topology-aware collectives
- Extended Energy-aware designs and Virtualization Support
- Extended Support for MPI Tools Interface (as in MPI 3.0)
- Extended FT support
- Support for * features will be available in future MVAPICH2 Releases

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.)
- A. Jain (Ph.D.)
- K. S. Khorassani (Ph.D.)
- P. Kousha (Ph.D.)
- B. Michalowicz (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 (M.S.)
- L. Xu (Ph.D.)
- H. Ahn (Ph.D.)
- G. Kuncham (Ph.D.)
- R. Vaidya (Ph.D.)
- J. Yao (Ph.D.)
- M. Han (M.S.)
- A. Guptha (M.S.)

Current Research Scientists

- M. Abduljabbar
- A. Shafi

Current Students (Undergrads)

- V. Shah
- T. Chen

Current Faculty

- H. Subramoni

Current Software Engineers

- B. Seeds
- N. Pavuk
- N. Shineman
- M. Lieber

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.)
- J. Hashmi (Ph.D.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (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 Post-Docs

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

Past Research Scientists

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

Past Senior Research Associate

- J. Hashmi

Past Programmers

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

Past Research Specialist

- M. Arnold
- J. Smith

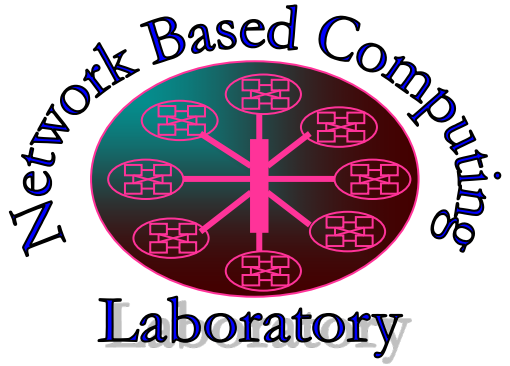
Join us for Multiple Events at SC '22

- Presentations at OSU and X-Scale Booth (#4035)
  - 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 (Mellanox, Intel, Microsoft, and AWS) and satellite events
- Complete details available at

  http://mvapich.cse.ohio-state.edu/conference/904/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/