MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC '21)

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)

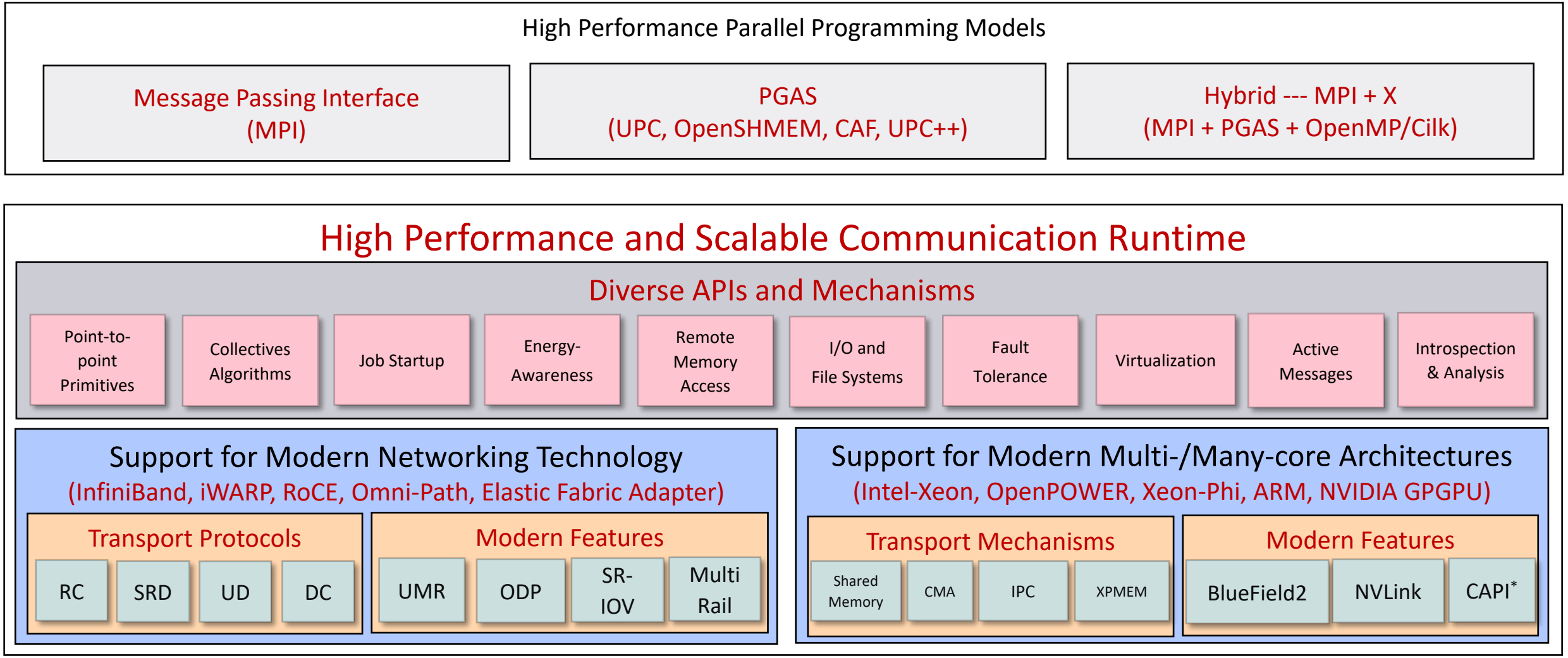
MPICH + MVICH + VAPI = MVAPICH

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), and AWS EFA
- 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 GPGPUs, since 2014
  - 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

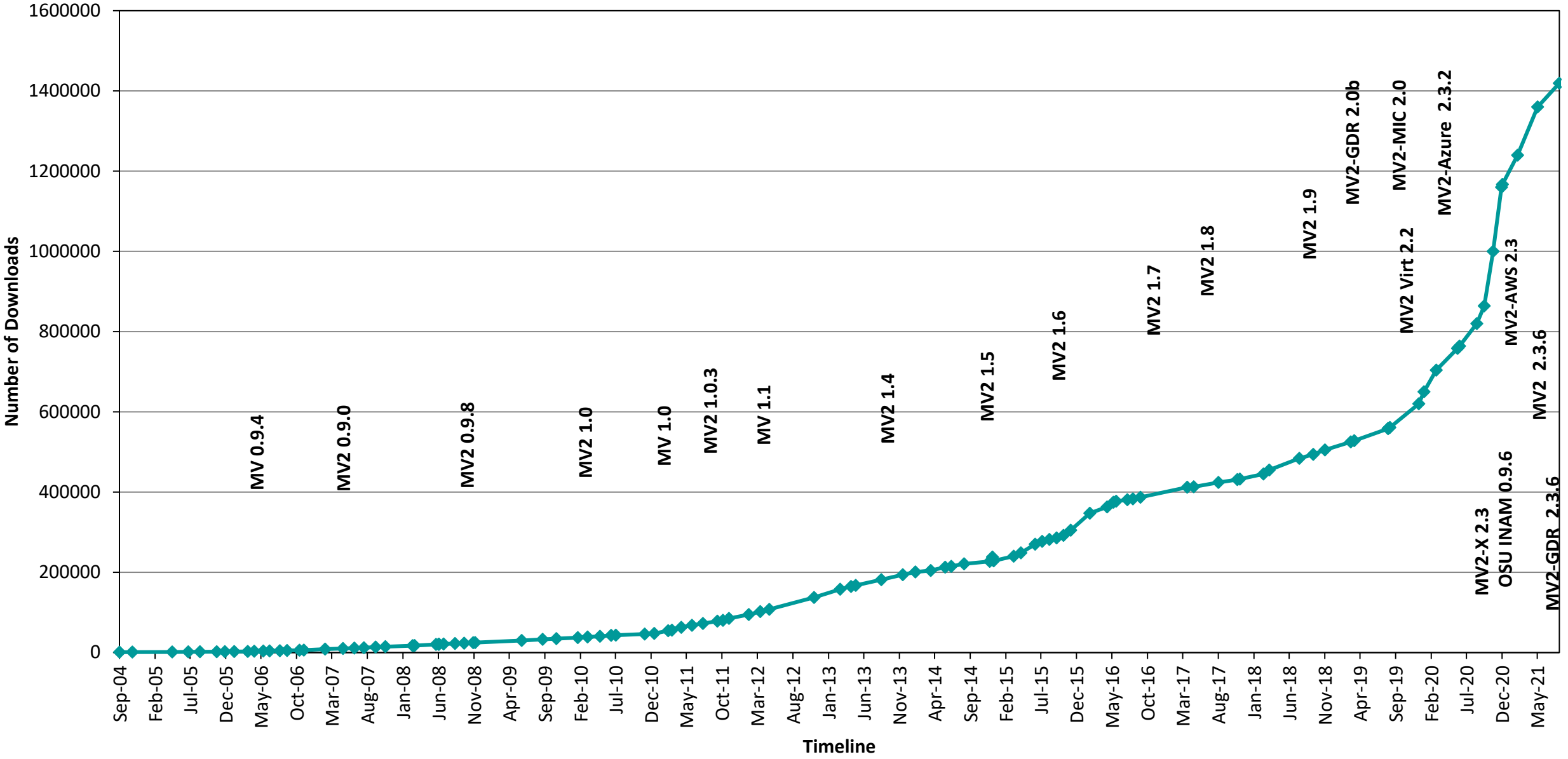
- Used by more than 3,200 organizations in 89 countries
- More than 1.52 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov. '21 ranking)
  - 4th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 13th, 448, 448 cores (Frontera) at TACC
  - 26th, 288,288 cores (Lassen) at LLNL
  - 38th, 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 13th ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

Architecture of MVAPICH2 Software Family for HPC, DL/ML, and Data Science



* Upcoming

MVAPICH2 Release Timeline and Downloads

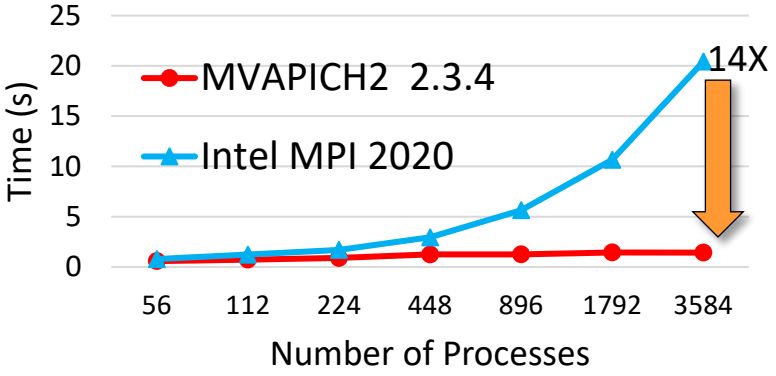


MVAPICH2 Software Family

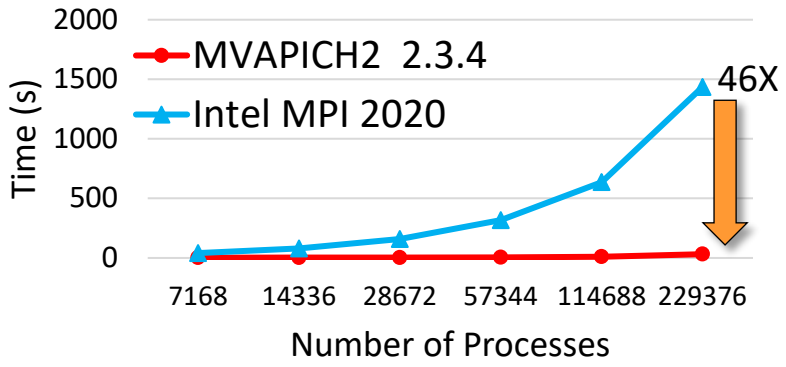
High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters with NVIDIA GPUs
MVAPICH2-Virt	High-performance and scalable MPI for hypervisor and container based HPC cloud
MVAPICH2-EA	Energy aware and High-performance MPI
MVAPICH2-MIC	Optimized MPI for clusters with Intel KNC
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 – Basic MPI

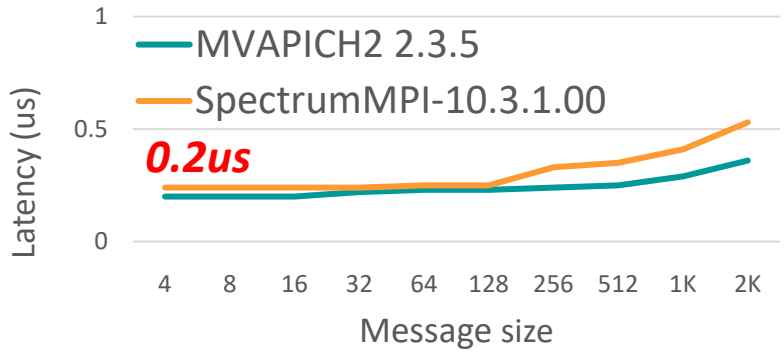
MPI_Init on Frontera (Small Scale)



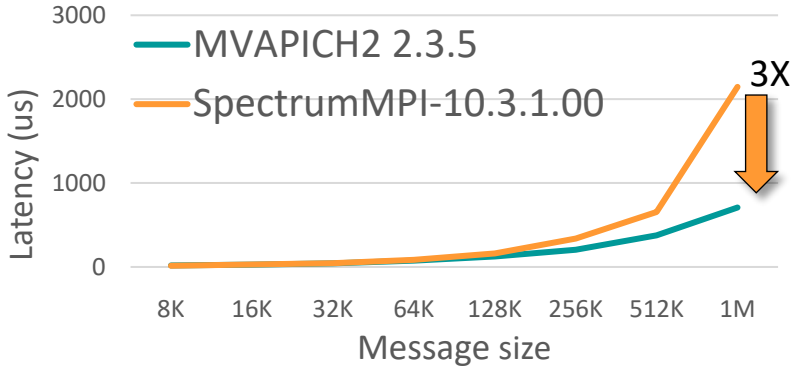
MPI_Init on Frontera (Large Scale)



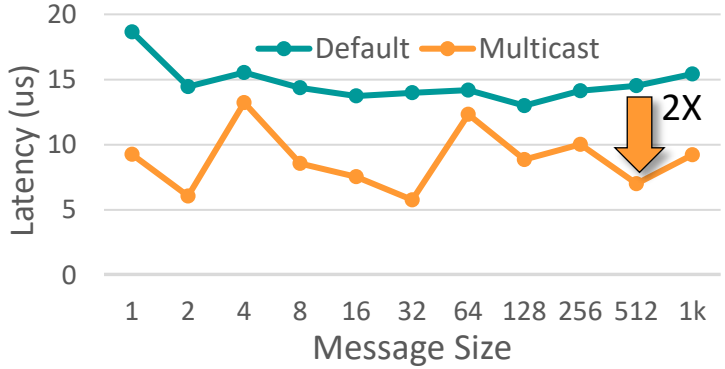
Intra-node Latency on OpenPOWER



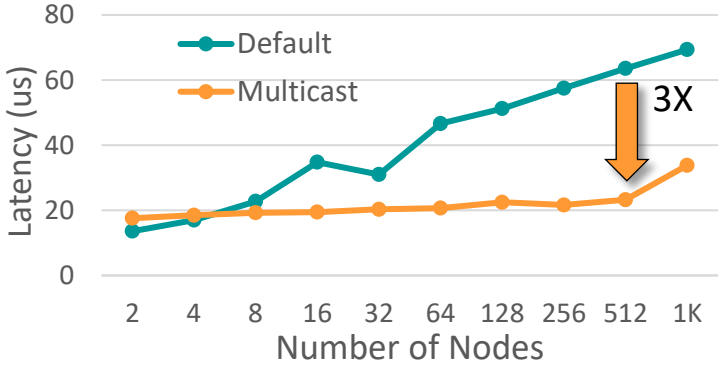
Allreduce Latency on OpenPOWER
1 Node 40 PPN



MPI_Bcast using Multicast on Frontera
2048 Nodes

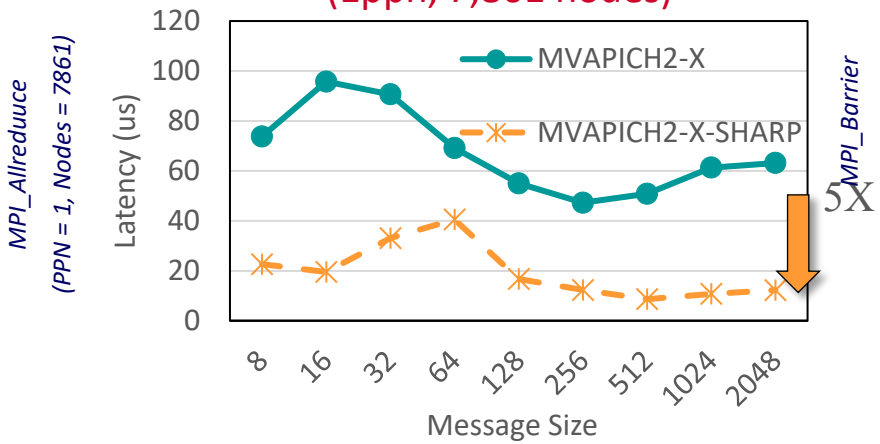


MPI_Bcast using Multicast on Frontera
32 Bytes

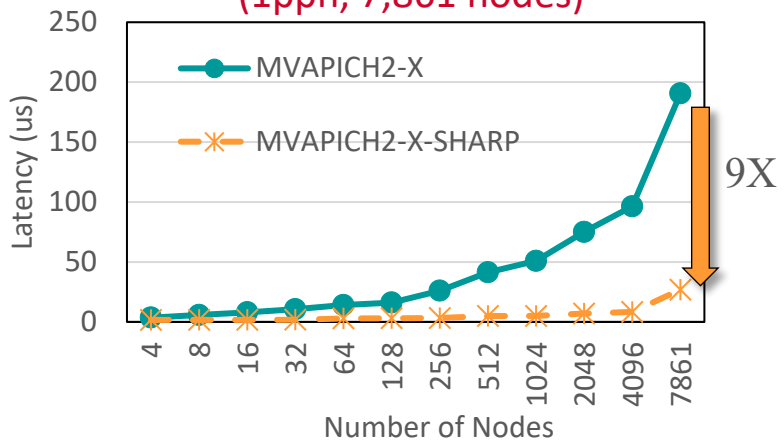


MVAPICH2-X – Advanced MPI + PGAS + Tools

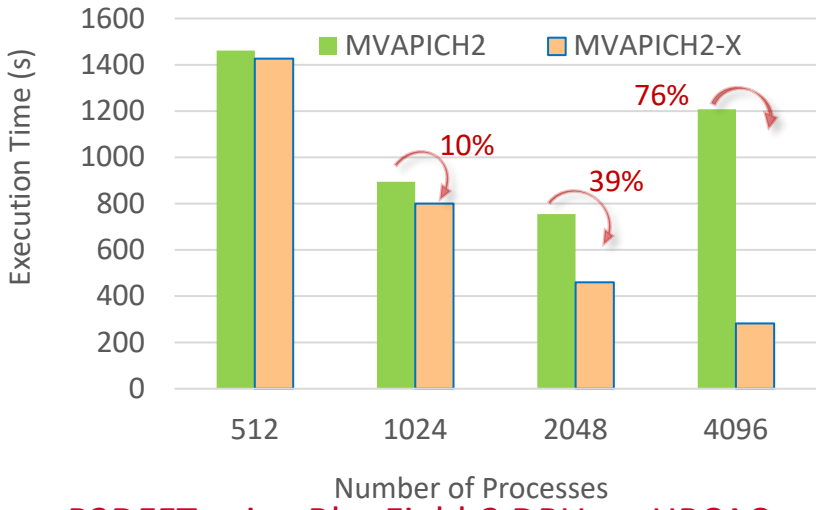
MPI_Allreduce using SHARP on Frontera (1ppn, 7,861 nodes)



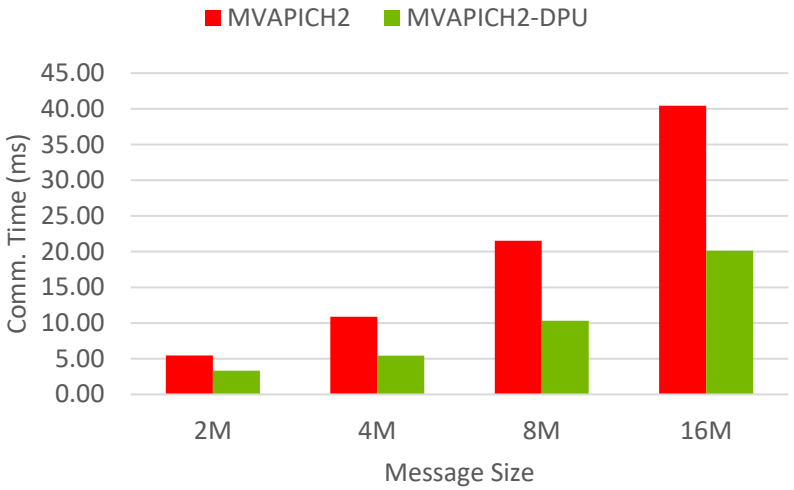
MPI_Barrier using SHARP on Frontera (1ppn, 7,861 nodes)



Impact of Transport Protocol Selection

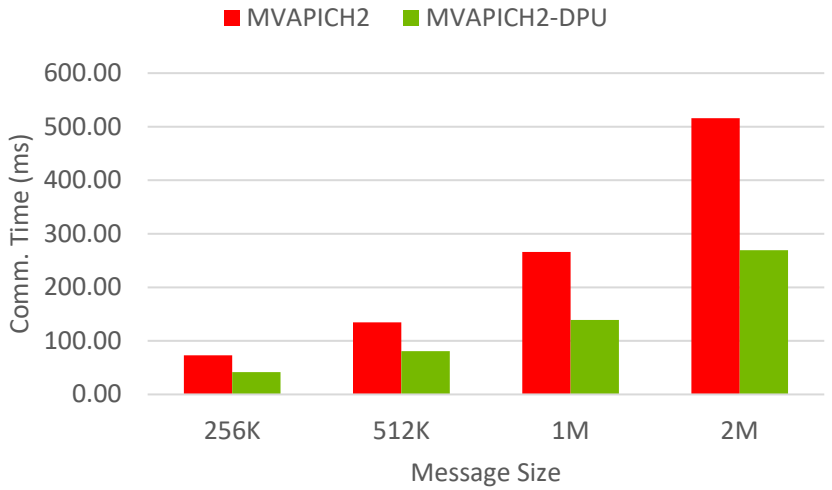


Total Execution Time, BF-2 (osu_ibcast)



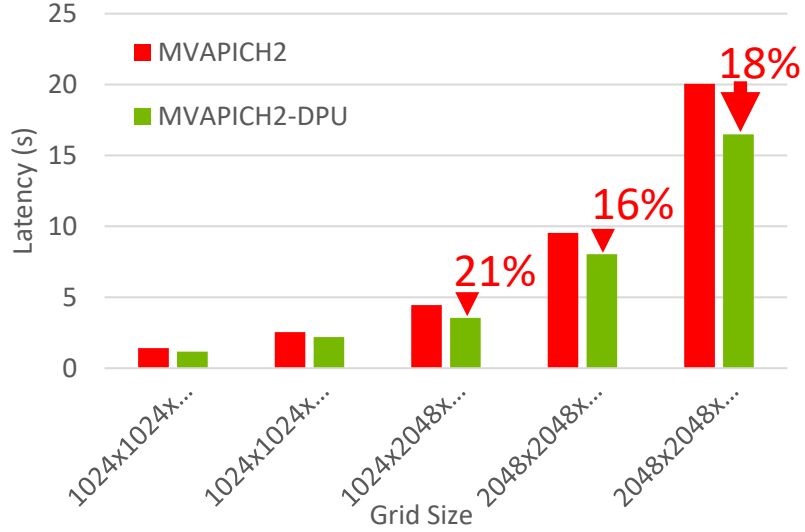
16 Nodes, 32 PPN

Total Execution Time, BF-2 (osu_iallgather)



16 Nodes, 15 PPN

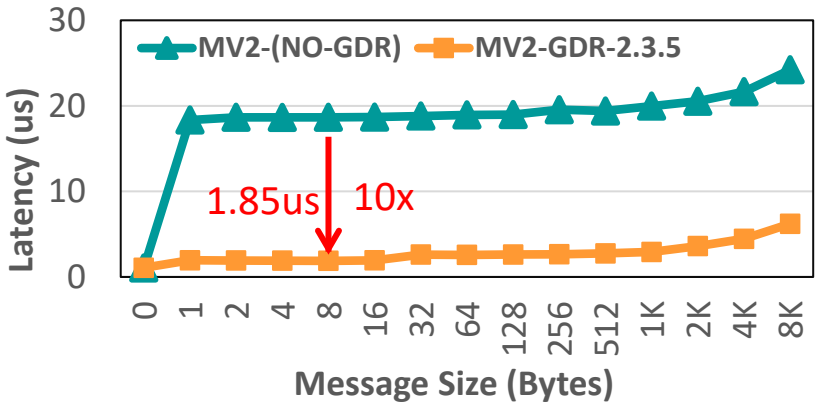
P3DFFT using BlueField-2 DPU on HPCAC



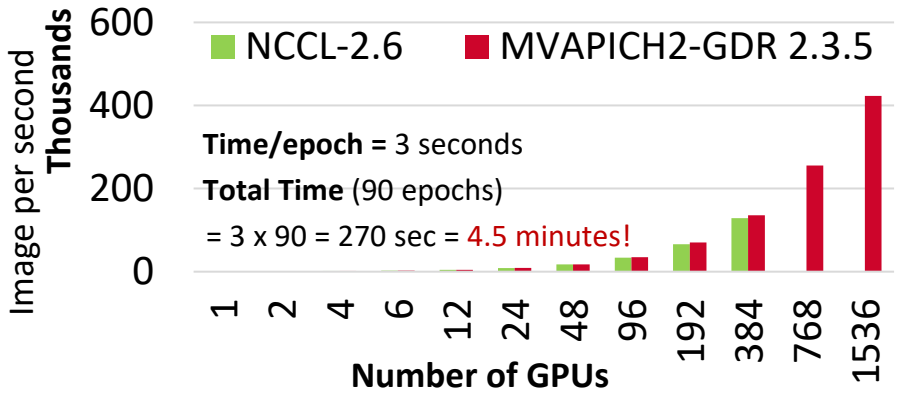
32 Nodes, 32 PPN

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

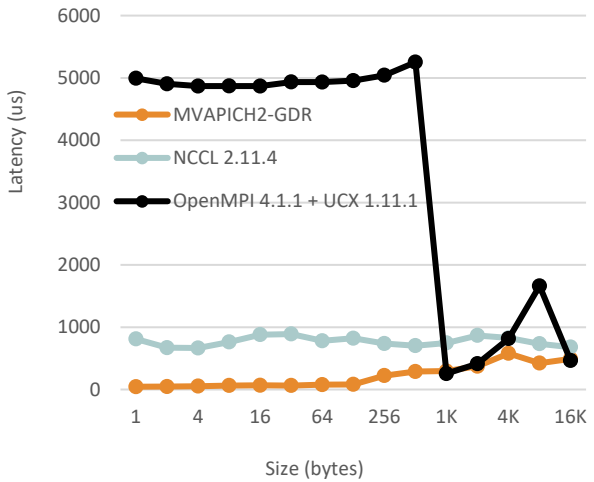
Best Performance for GPU-based Transfers



TensorFlow Training with MVAPICH2-GDR on Summit



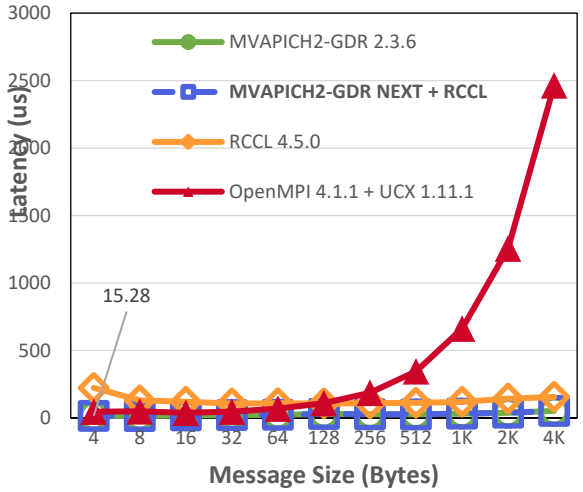
Enhanced Alltoall on DGX2-A100



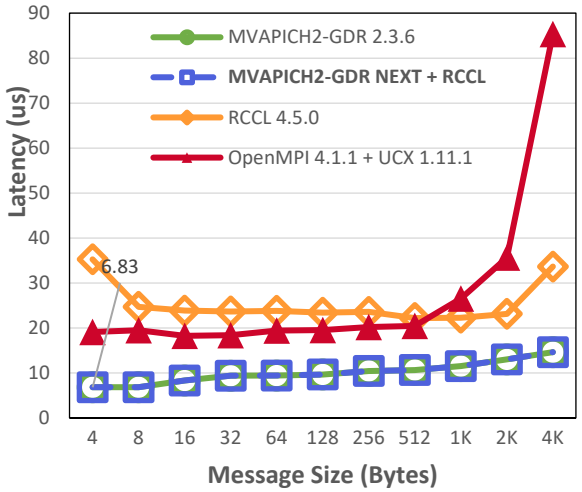
ROCm Support for AMD GPUs (Available with MVAPICH2-GDR 2.3.6)

LLNL Corona Cluster - ROCm-4.3.0 (mi50 AMD GPUs)

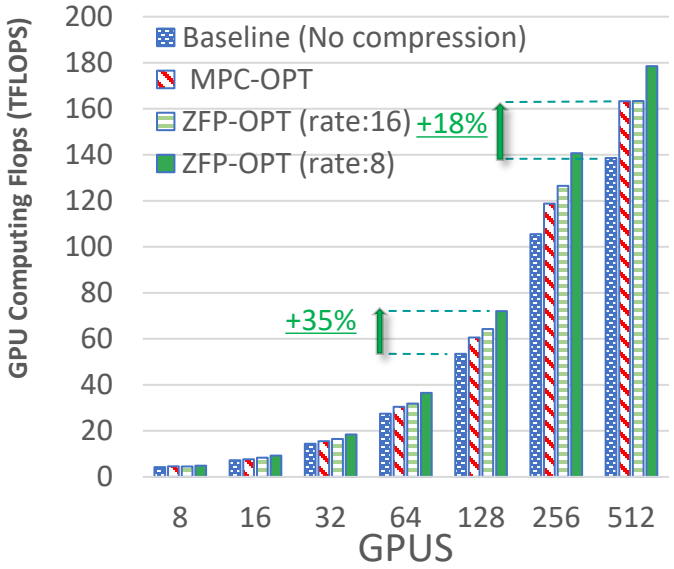
Allreduce 128 GPUs (16 Nodes, 8 GPN)



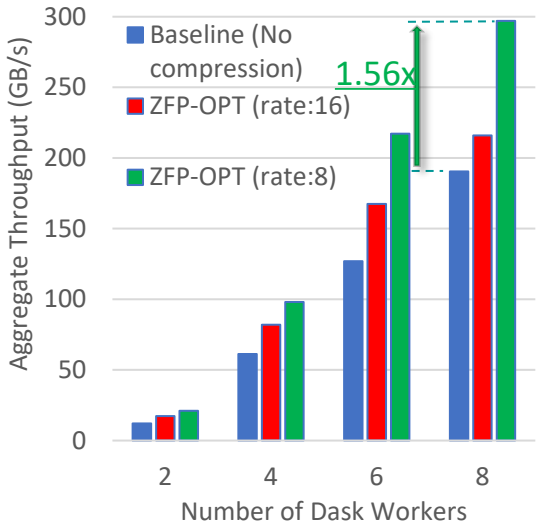
Broadcast 128 GPUs (16 Nodes, 8 GPN)



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



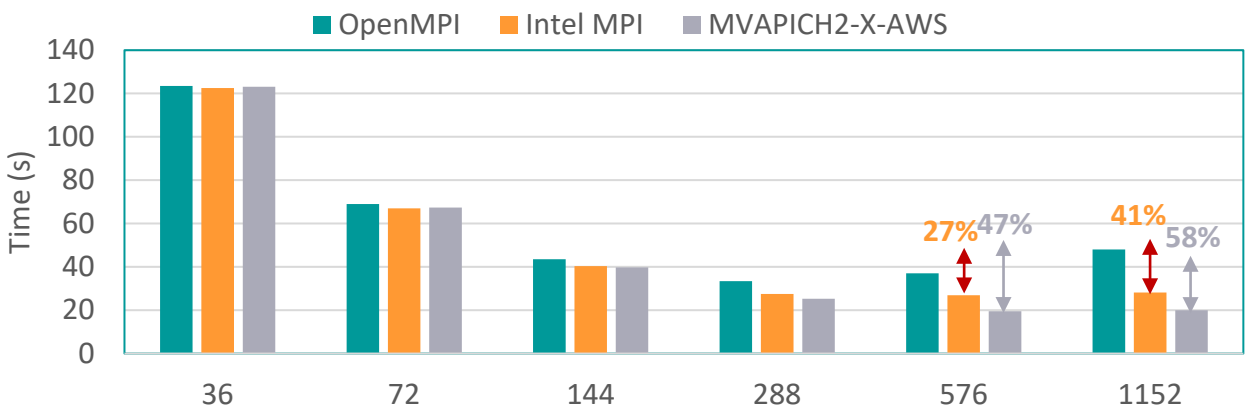
"On-the-fly" Compression Support (DASK for Data Science)



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

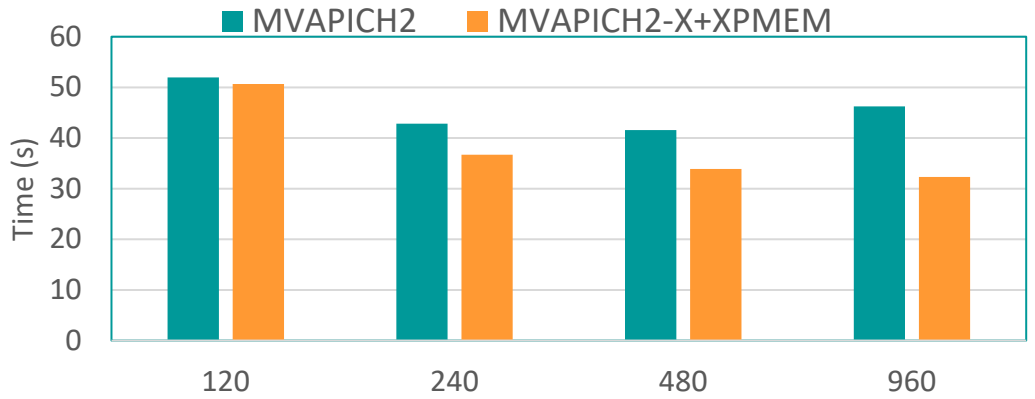
MVAPICH2-X Advanced Support for HPC-Clouds

Performance on Amazon EFA



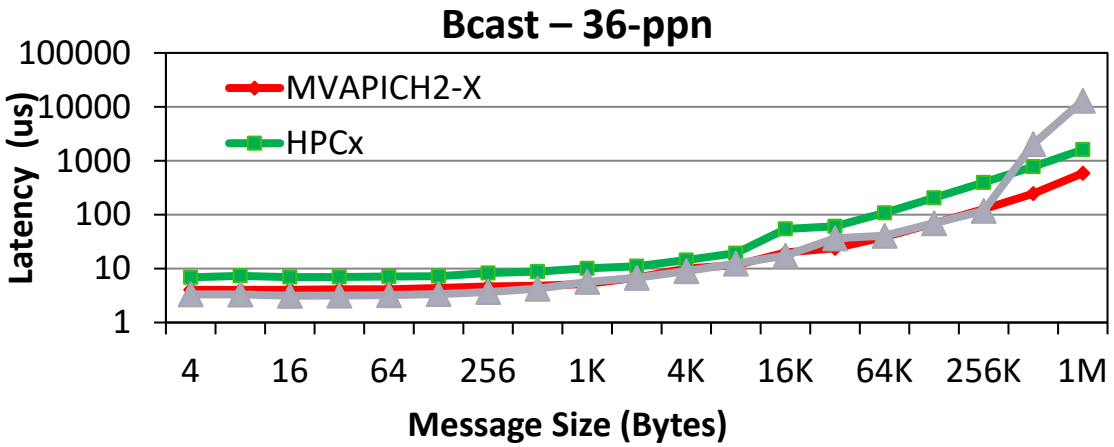
Instance type: c5n.18xlarge
CPU: Intel Xeon Platinum 8124M @ 3.00GHz
MVAPICH2 version: MVAPICH2-X-aws v2.3
OpenMPI version: Open MPI v4.0.3 with libfabric 1.9
IntelMPI version: Intel MPI 2019.7.217

Performance of WRF on Microsoft Azure



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

Performance on Oracle HPC Shapes



- Releases
  - MVAPICH2-X-AWS 2.3
  - MVAPICH2-Azure 2.3.3
  - Integrated Azure CentOS HPC Images:

https://github.com/Azure/azhpc-images/releases/tag/centos-7.6-hpc-20200417

MVAPICH2 – Future Roadmap and Plans for Exascale

- Update to MPICH 3.4.2 CH3 channel
  - 2021
- Initial support for the CH4 channel
  - Mid 2022
- Making CH4 channel default
  - Late 2022 / 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.)
- N. Sarkauskas (M.S.)

Current Research Scientists

- A. Shafi
- H. Subramoni

Current Students (Undergrads)

- M. Lieber
- L. Xu

Current Software Engineers

- B. Seeds
- N. Shineman

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.)
- 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

Past Research Specialist

- M. Arnold
- J. Smith

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

Thank You!

subramoni.1@osu.edu

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

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

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

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