The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC '19)

by

Hari Subramoni

The Ohio State University

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

http://www.cse.ohio-state.edu/~subramon

History of MVAPICH

- A long time ago, in a galaxy far, far away…. (actually 20 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)

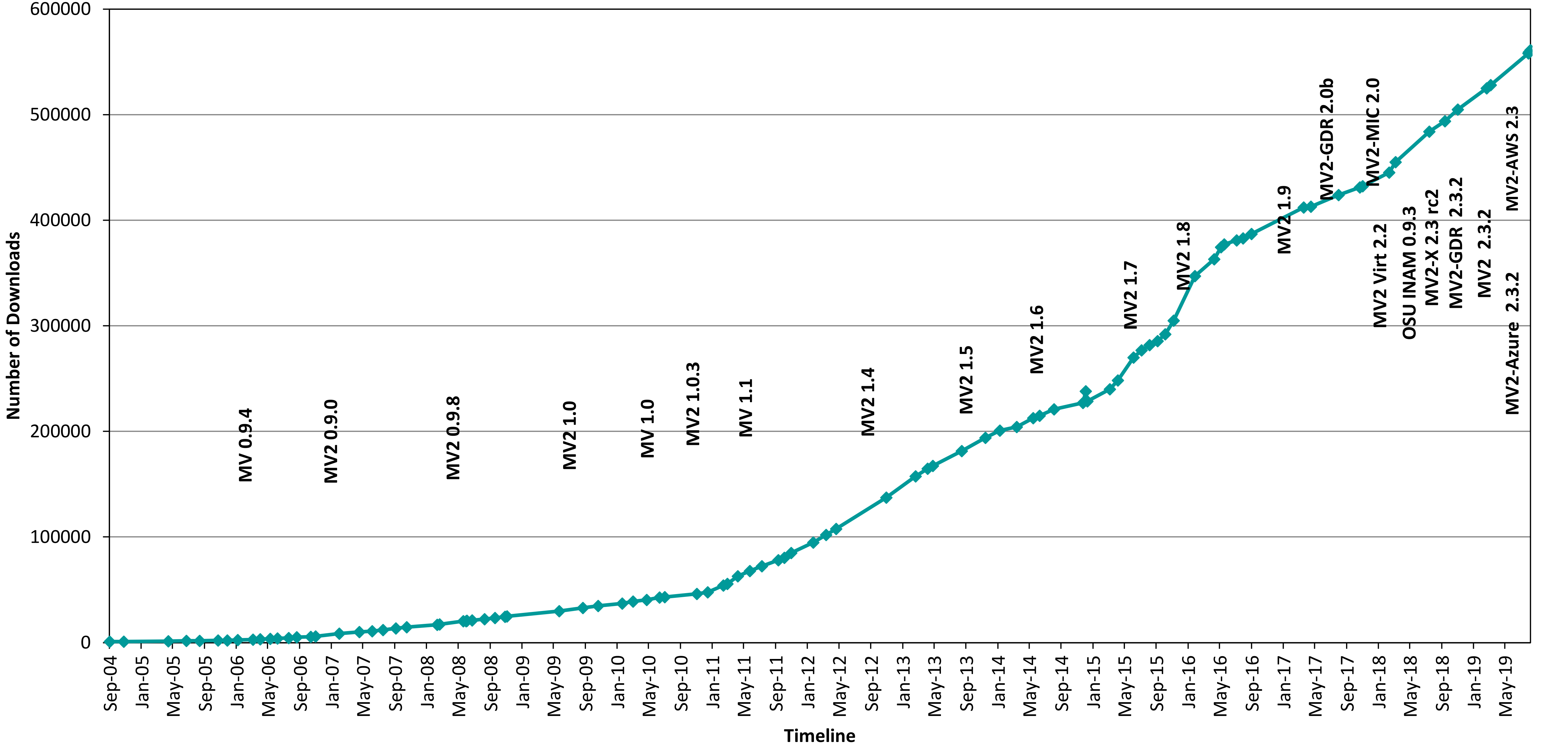
MPICH + MVICH + VAPI = MVAPICH

Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - **Based on MPICH2 3.2.1**
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.1), Started in 2001, First version available in 2002 (SC '02)
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  - **Used by more than 3,050 organizations in 89 countries**
  - **More than 615,000 (> 0.6 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '19 ranking)
    - 3^{rd}, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China
    - 5^{th}, 448, 448 cores (Frontera) at TACC
    - 8^{th}, 391,680 cores (ABCI) in Japan
    - 14^{th}, 570,020 cores (Neurion) in South Korea and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)
  - http://mvapich.cse.ohio-state.edu
- Empowering Top500 systems for over a decade

Partner in the #5^{th} TACC Frontera System

MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family (HPC and DL)

High Performance Parallel Programming Models

- Message Passing Interface (MPI)
- PGAS (UPC, OpenSHMEM, CAF, UPC++)
- Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)

High Performance and Scalable Communication Runtime

Diverse APIs and Mechanisms

- Point-to-point Primitives
- Collectives Algorithms
- Job Startup
- Energy-Awareness
- Remote Memory Access
- I/O and File Systems
- Fault Tolerance
- Virtualization
- Active Messages
- Introspection & Analysis

Support for Modern Networking Technology

(InfiniBand, iWARP, RoCE, Omni-Path, Elastic Fabric Adapter)

Transport Protocols: RC, SRD, UD, DC

Modern Features: UMR, ODP, SR-IOV, Multi Rail

Support for Modern Multi-/Many-core Architectures

(Intel-Xeon, OpenPOWER, Xeon-Phi, ARM, NVIDIA GPGPU)

Transport Mechanisms: Shared Memory, CMA, IVSHMEM, XPMEM

Modern Features: Optane*, NVLink, CAPI*

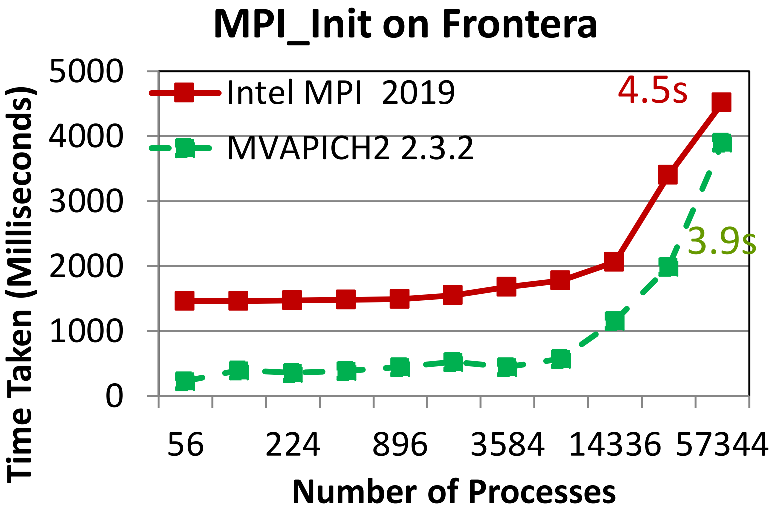
* Upcoming

MVAPICH2 Software Family

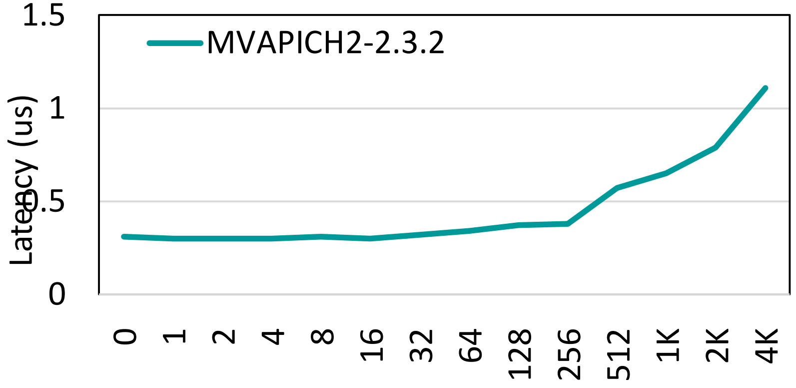
High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters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

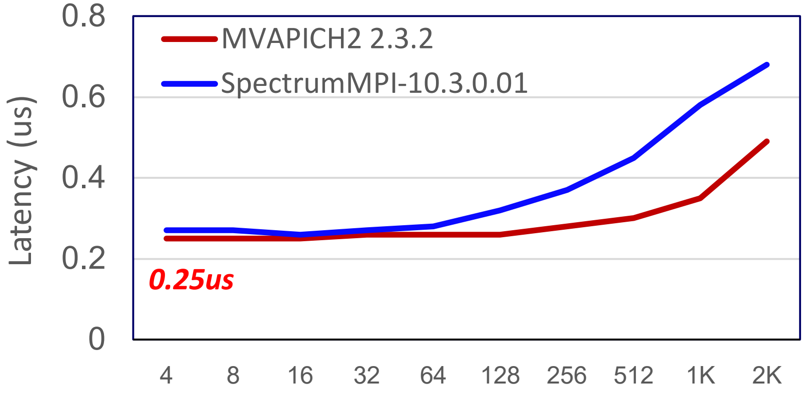
Fast Startup on Emerging Many-Cores



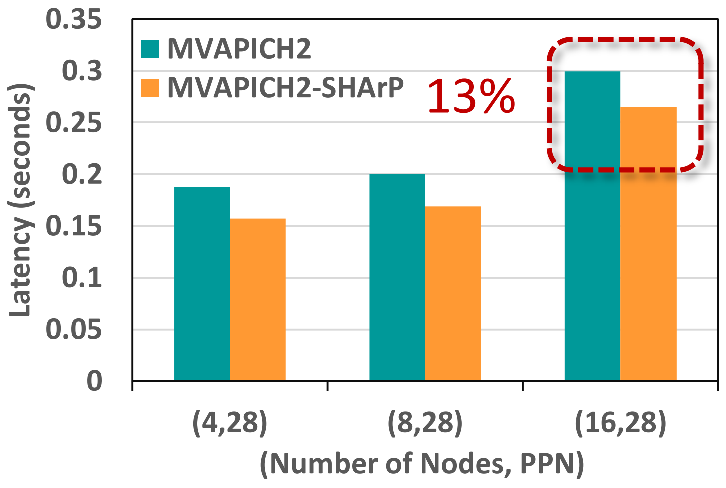
Enhanced Intra-node Performance for ARM



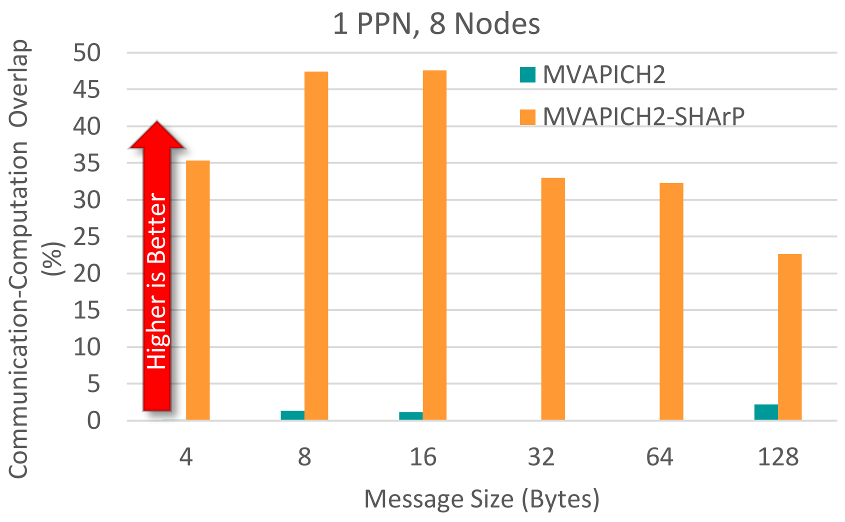
Enhanced Intra-node Performance for OpenPOWER



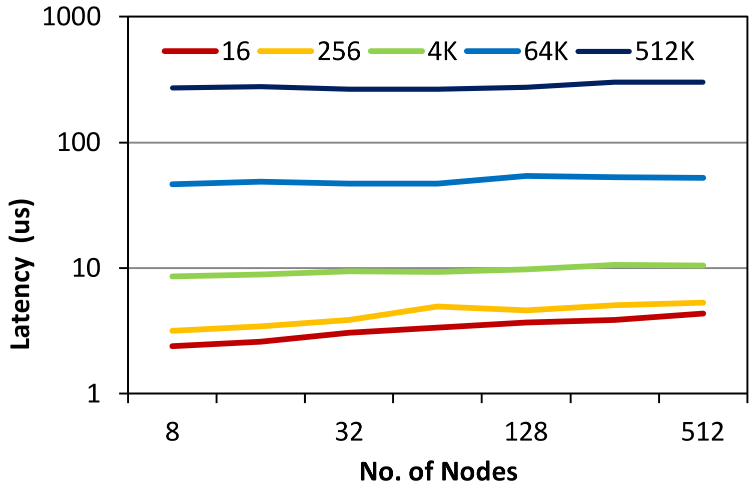
Advanced Allreduce with SHARP



Advanced Non-Blocking Allreduce with SHARP

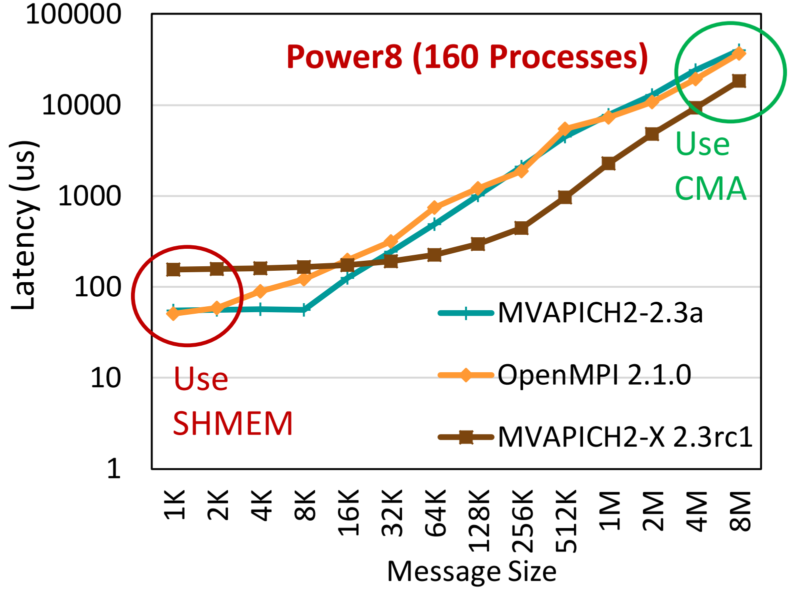


MPI_Bcast using RDMA_CM-based Multicast

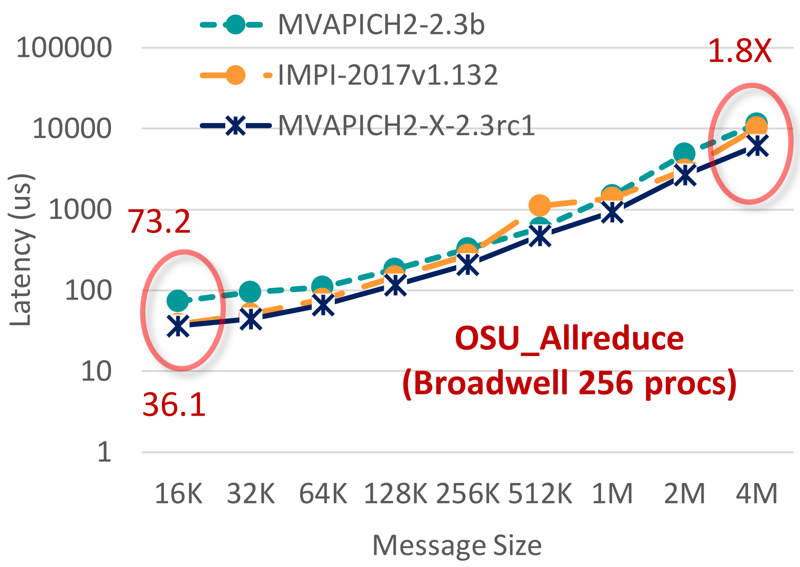


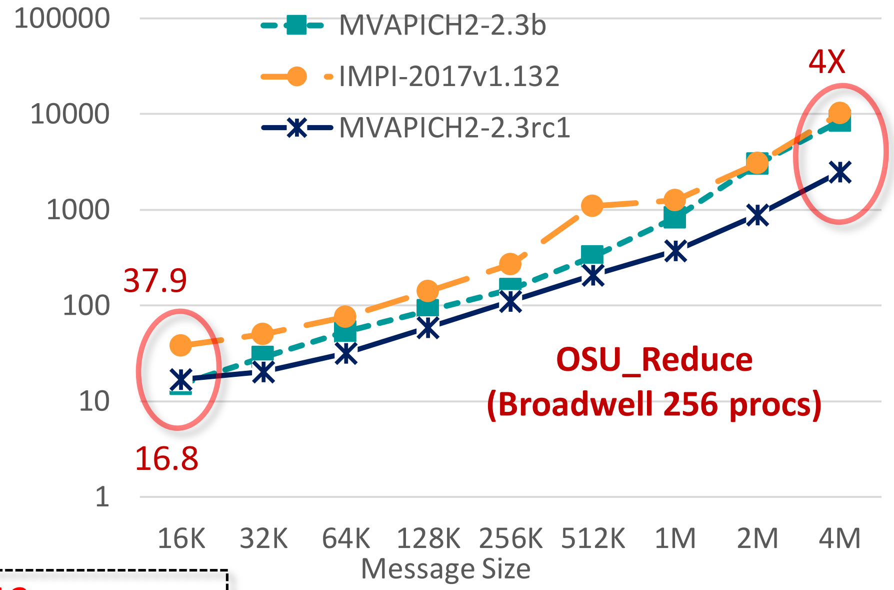
MVAPICH2-X – Advanced MPI + PGAS + Tools

CMA-Aware MPI_Bcast

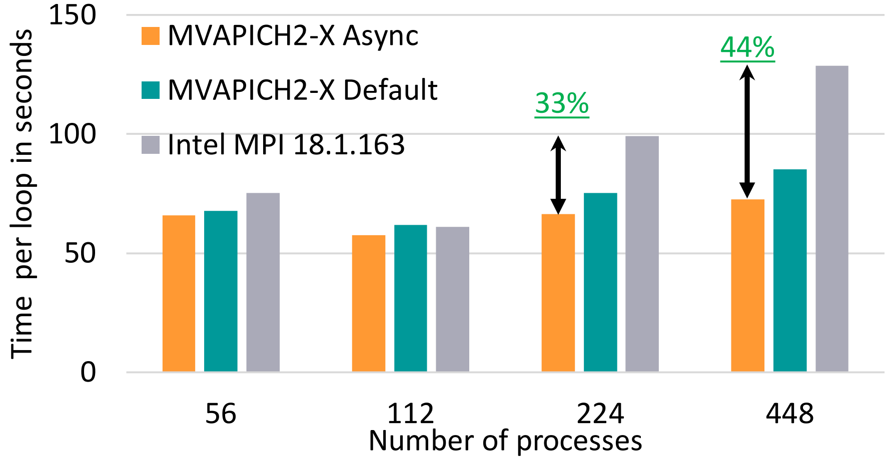


Shared Address Space (XPMEM)-based Collectives Design

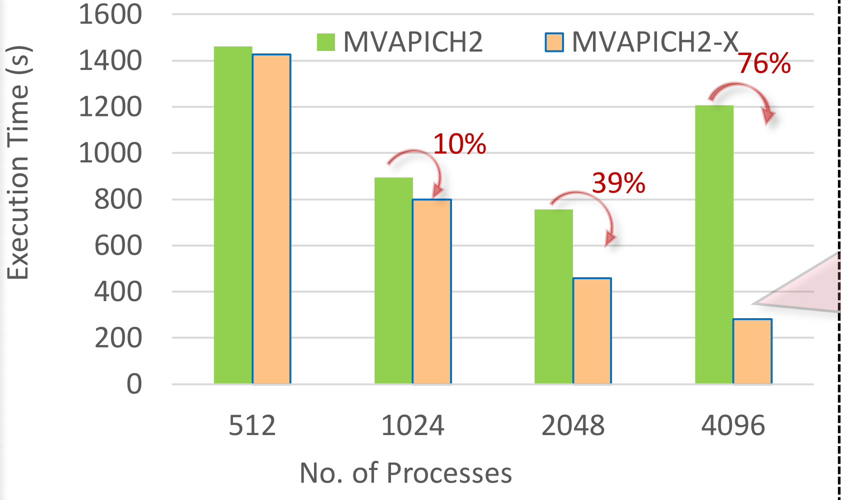




Performance of P3DFFT Optimized Async Progress



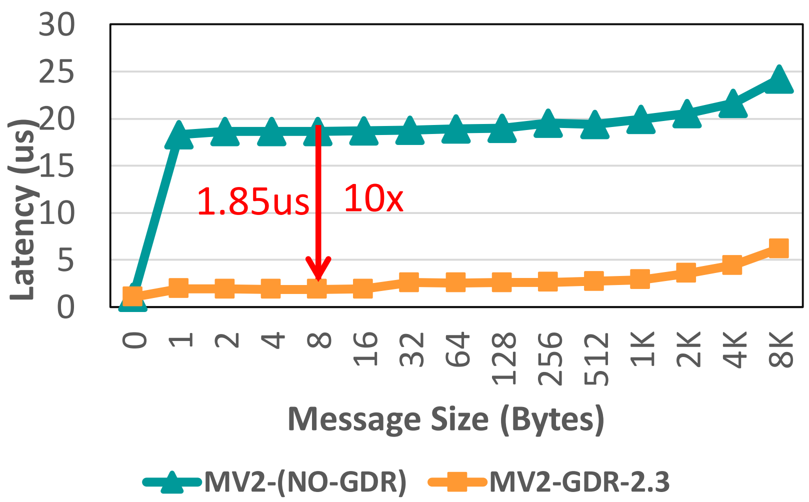
Neuron with YuEtAl2012



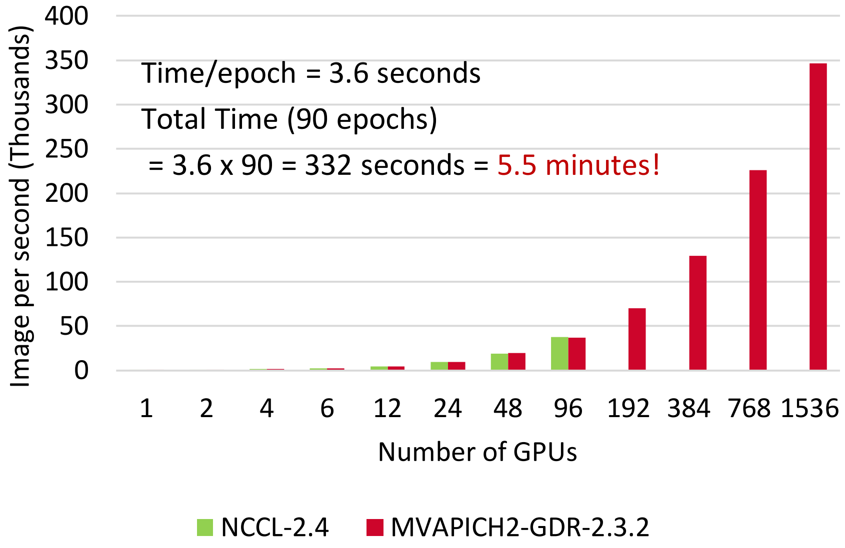
Overhead of RC protocol for connection establishment and communication

MVAPICH2-GDR – Optimized MPI for clusters with NVIDIA GPUs

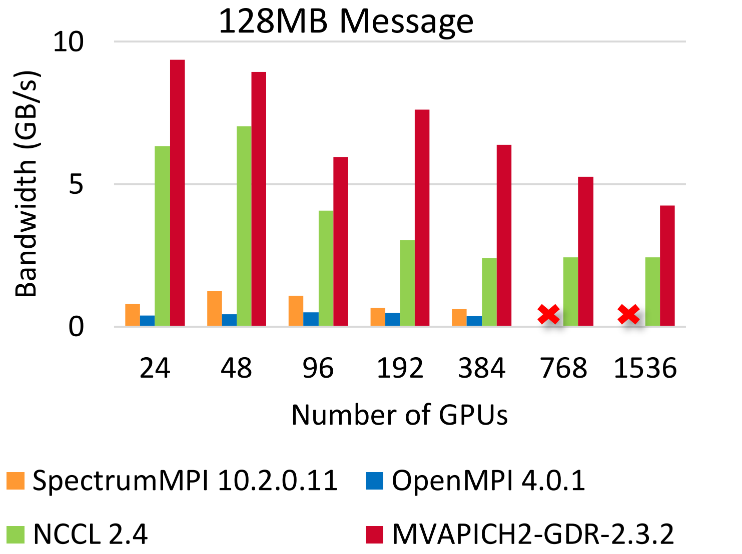
Best Performance for GPU-based Transfers



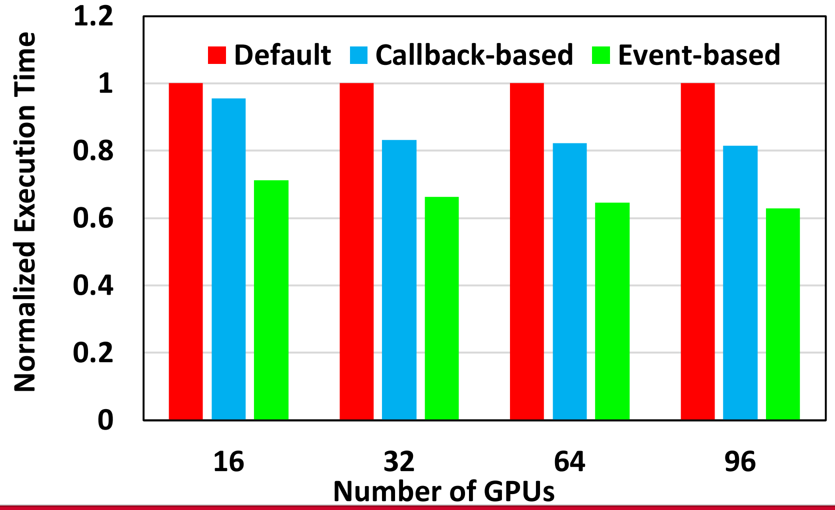
TensorFlow Training with MVAPICH2-GDR on Summit



GPU-Based MPI_Allreduce on Summit



Enhanced Kernel-based Datatype Processing for Cosmo Weather Prediction Model on x86



Enhanced Kernel-based Datatype Processing for COMB Application Kernel on POWER9

16 GPUs on POWER9 system (test Comm mpi Mesh cuda Device Buffers mpi_type)

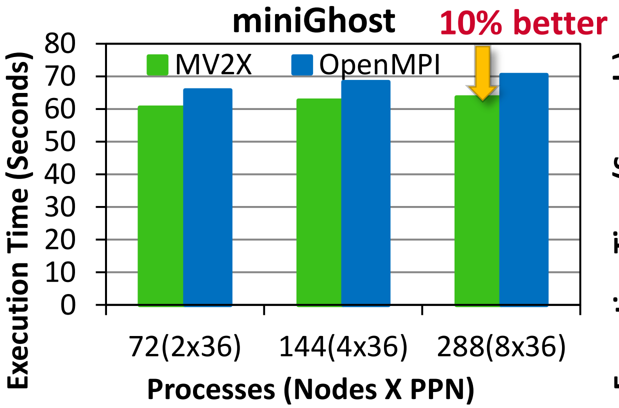
	pre-comm	post-recv	post-send	wait-recv	wait-send	post-comm	start-up	test-comm	bench-comm
Spectrum MPI 10.3	0.0001	0.0000	1.6021	1.7204	0.0112	0.0001	0.0004	7.7383	**83.6229**
MVAPICH2-GDR 2.3.2	0.0001	0.0000	0.0862	0.0871	0.0018	0.0001	0.0009	0.3558	**4.4396**
MVAPICH2-GDR 2.3.3 (Upcoming)	0.0001	0.0000	0.0030	0.0032	0.0001	0.0001	0.0009	0.0133	**0.1602**

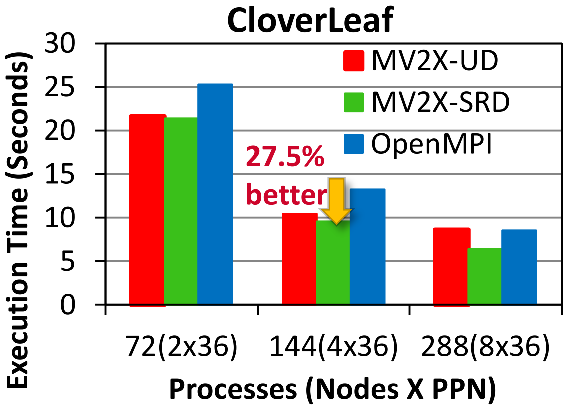
18x

27x

MVAPICH2-X Advanced Support for HPC-Clouds

Performance on Amazon EFA

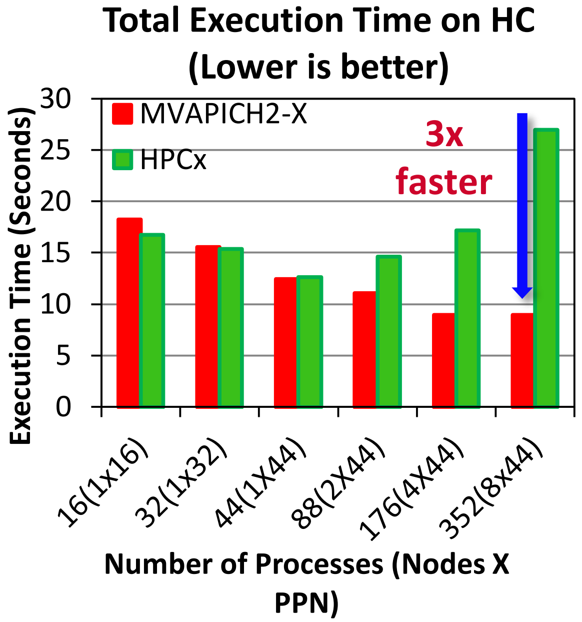


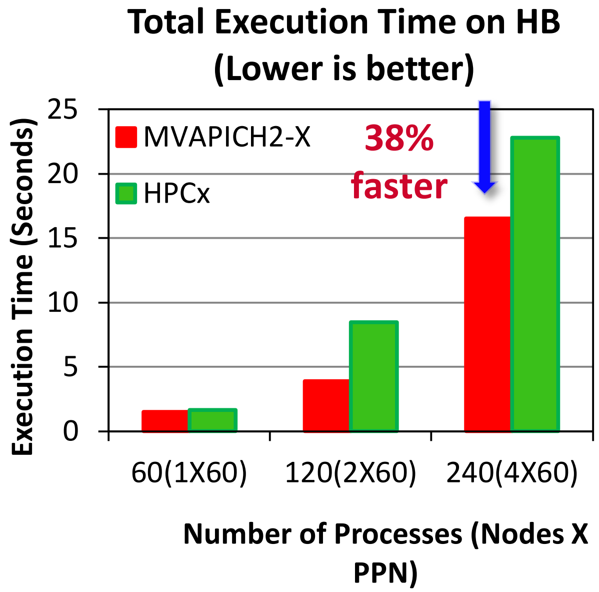


Instance type: c5n.18xlarge
CPU: Intel Xeon Platinum 8124M @ 3.00GHz
MVAPICH2 version: MVAPICH2-X 2.3rc2 + SRD support
OpenMPI version: Open MPI v3.1.3 with libfabric 1.7

- MVAPICH2-X-AWS 2.3
- Released on 08/12/2019
- Major Features and Enhancements
  - Based on MVAPICH2-X 2.3
  - Support for on Amazon EFA adapter's Scalable Reliable Datagram (SRD)
  - Support for XPMEM based intra-node communication for point-to-point and collectives
  - Enhanced tuning for point-to-point and collective operations
  - Targeted for AWS instances with Amazon Linux 2 AMI and EFA support
  - Tested with c5n.18xlarge instance

Performance of Radix on Microsoft Azure





Deploy to Azure

- MVAPICH2-Azure 2.3.2
- Released on 08/16/2019
- Major Features and Enhancements
  - **Based on MVAPICH2-2.3.2**
  - **Enhanced tuning for point-to-point and collective operations**
  - **Targeted for Azure HB & HC virtual machine instances**
  - **Flexibility for 'one-click' deployment**
  - **Tested with Azure HB & HC VM instances**
- Available for download from http://mvapich.cse.ohio-state.edu/downloads/
- Detailed User Guide: http://mvapich.cse.ohio-state.edu/userguide/mv2-azure/

MVAPICH2 – Future Roadmap and Plans for Exascale

- Update to MPICH 3.3.2 CH3 channel
  - 2020
- Initial support for the CH4 channel
  - 2020/2021
- Making CH4 channel default
  - 2021/2022
- 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*
- 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

Thank You!

subramoni.1@osu.edu

http://web.cse.ohio-state.edu/~subramon

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

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

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

HiDL
High-Performance Deep Learning

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