ECP Community BOF Days

March 30 – April 1, 2021

Hosts: Ashley Barker (ORNL) and Osni Marques (LBNL)

The MVAPICH2 Project: Latest Status and Future Plans

Dhabaleswar K (DK) Panda

The Ohio State University

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

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

EXASCALE COMPUTING PROJECT

History of MVAPICH

- A long time ago, in a galaxy far, far away…. (actually… 21 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)

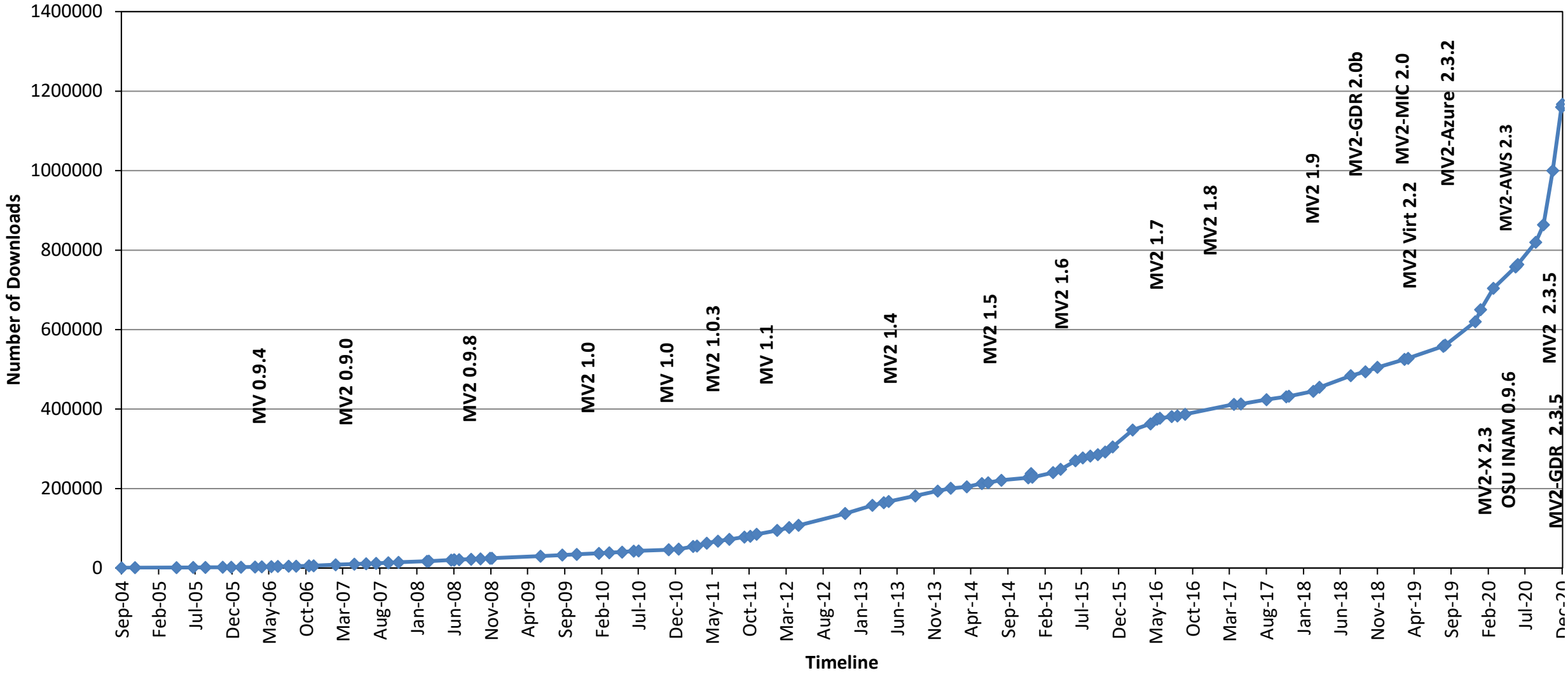
MPICH + MVICH + VAPI = MVAPICH

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 3,150 organizations in 89 countries
- More than 1.28 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '20 ranking)
  - 4th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 9th, 448, 448 cores (Frontera) at TACC
  - 14th, 391,680 cores (ABCI) in Japan
  - 21st, 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 9th ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

MVAPICH2 Release Timeline and Downloads



EXASCALE COMPUTING PROJECT

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*

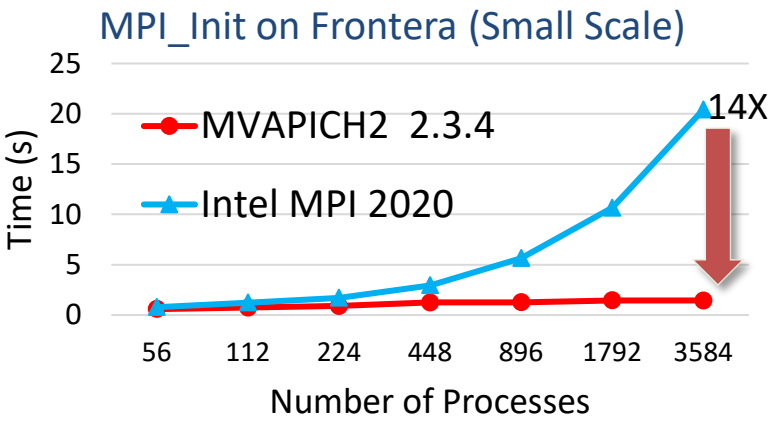
* Upcoming

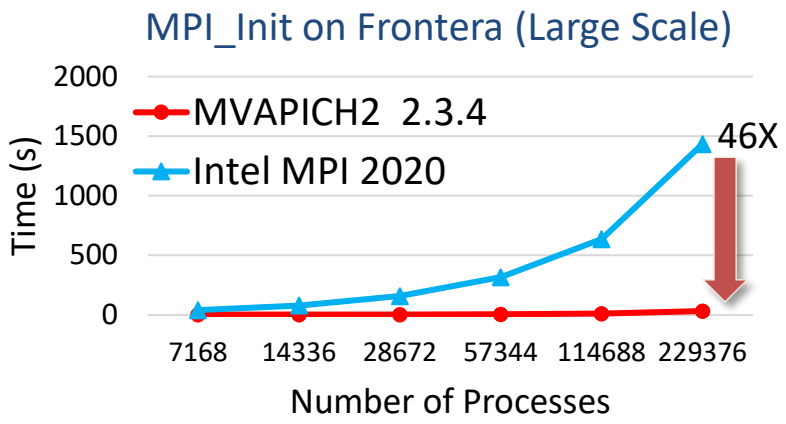
EXASCALE COMPUTING PROJECT

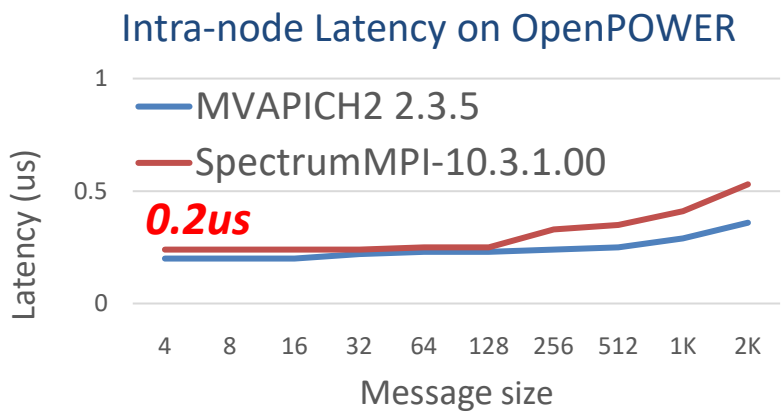
MVAPICH2 Software Family

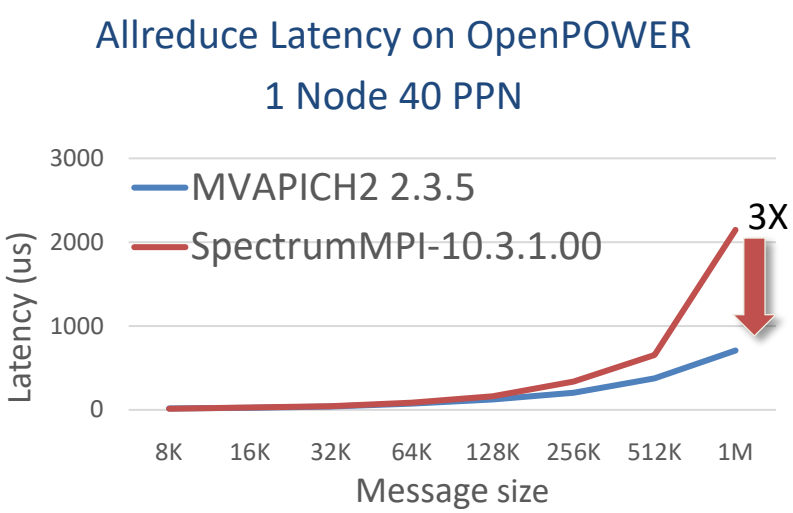
Requirements	Library
MPI with IB, iWARP, Omni-Path, and RoCE	MVAPICH2
Advanced MPI Features/Support, OSU INAM, PGAS and MPI+PGAS with IB, Omni-Path, and RoCE	MVAPICH2-X
MPI with IB, RoCE & GPU and Support for Deep/Machine Learning	MVAPICH2-GDR
HPC Cloud with MPI & IB	MVAPICH2-Virt
Energy-aware MPI with IB, iWARP and RoCE	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

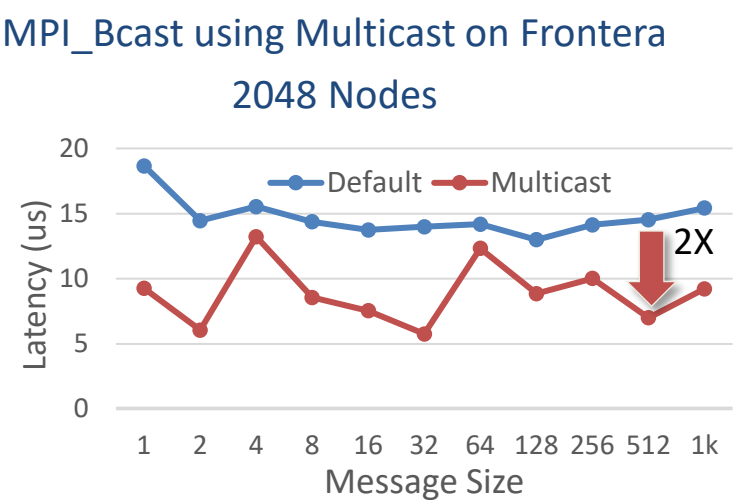
MVAPICH2 – Basic MPI

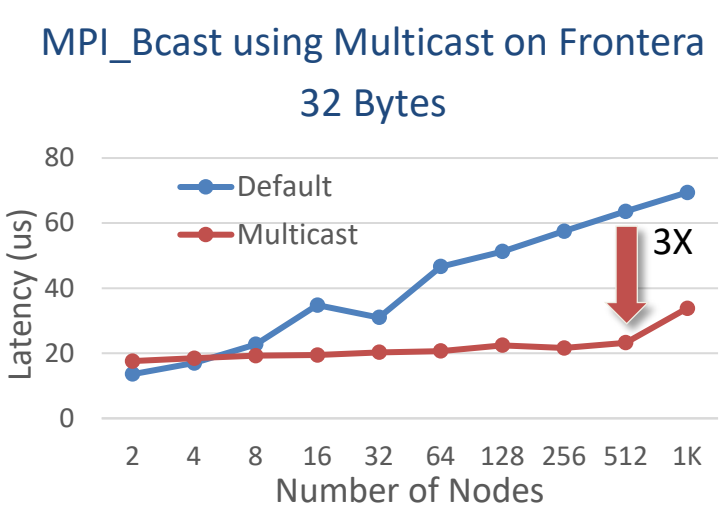






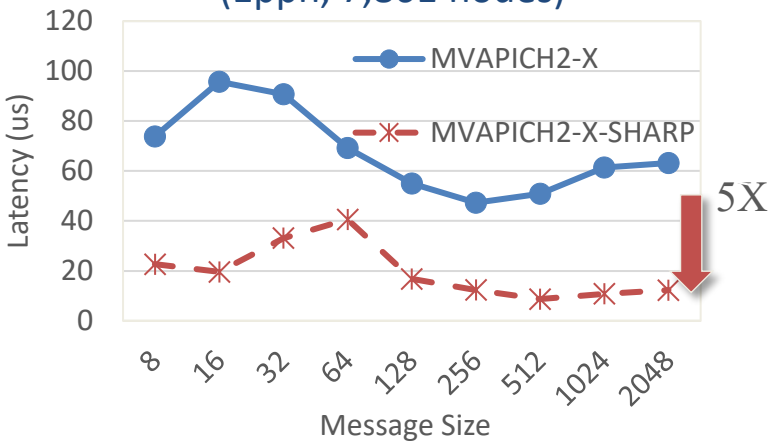




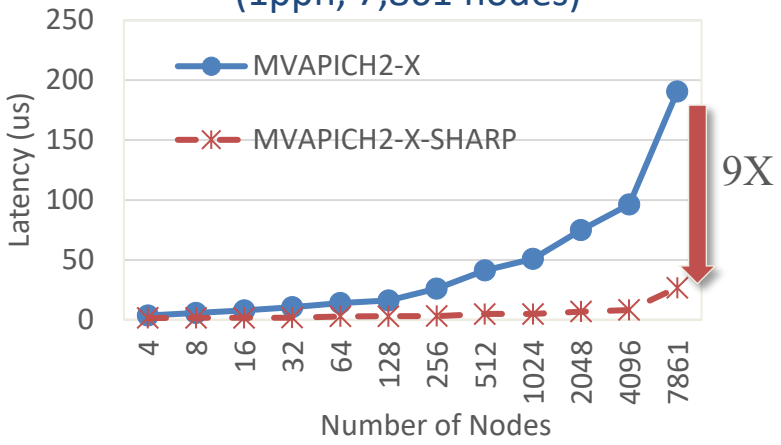


MVAPICH2-X – Advanced MPI + PGAS + Tools

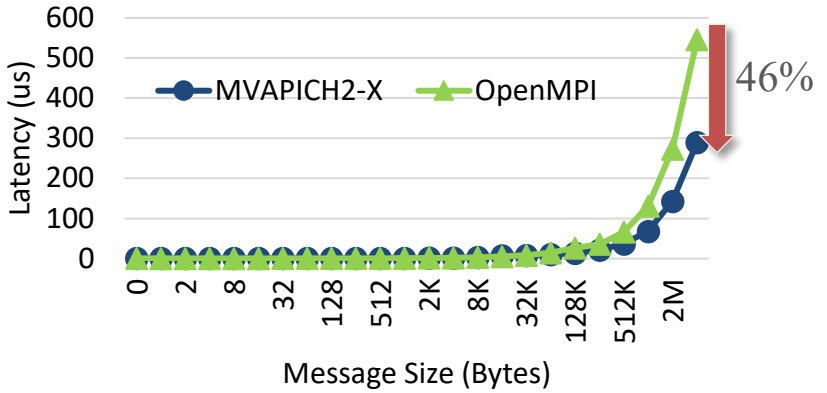
MPI_Allreduce using SHARP on Frontera (1ppn, 7,861 nodes)



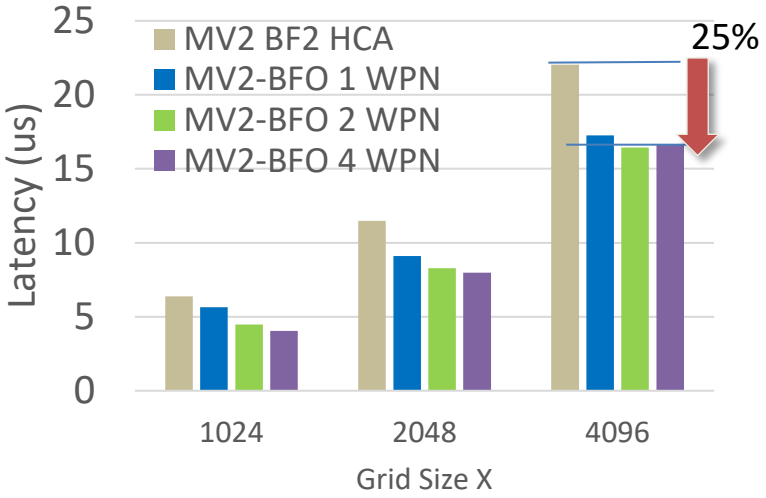
MPI_Barrier using SHARP on Frontera (1ppn, 7,861 nodes)



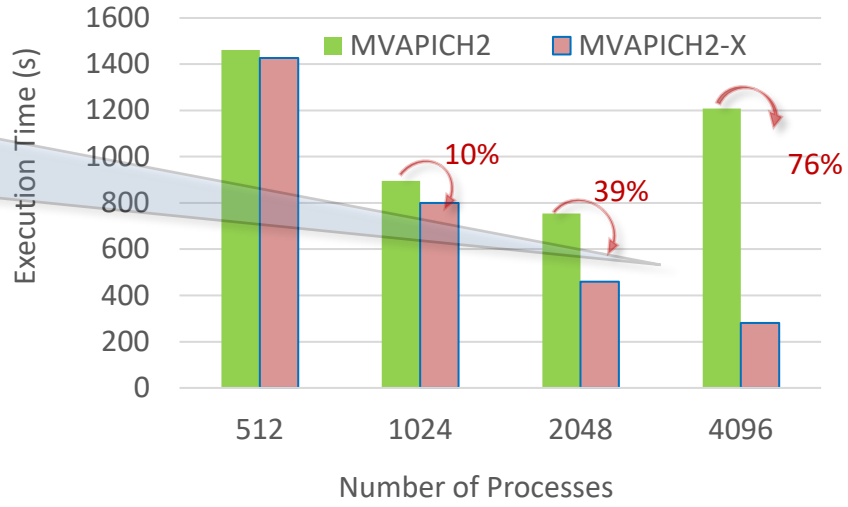
Intra-node Latency on ARM A64FX Oakami



P3DFFT using BlueField-2 DPU on HPCAC



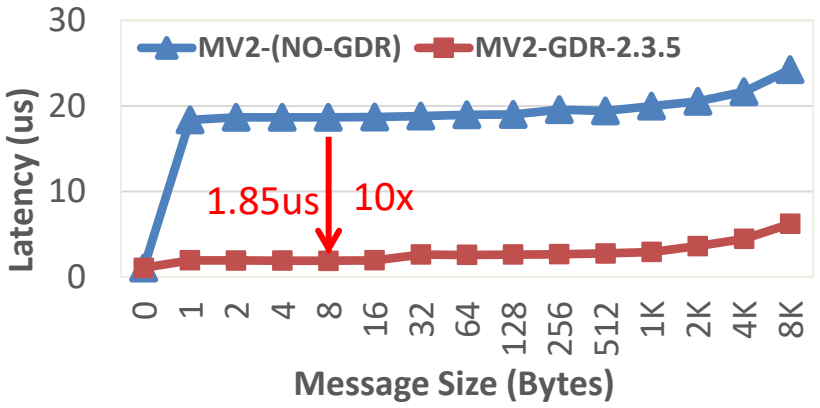
Overhead of RC protocol for connection establishment and communication



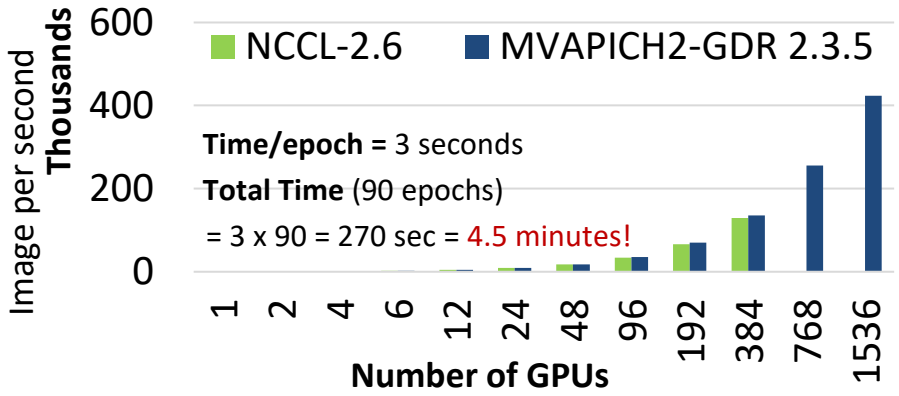
EXASCALE COMPUTING PROJECT

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

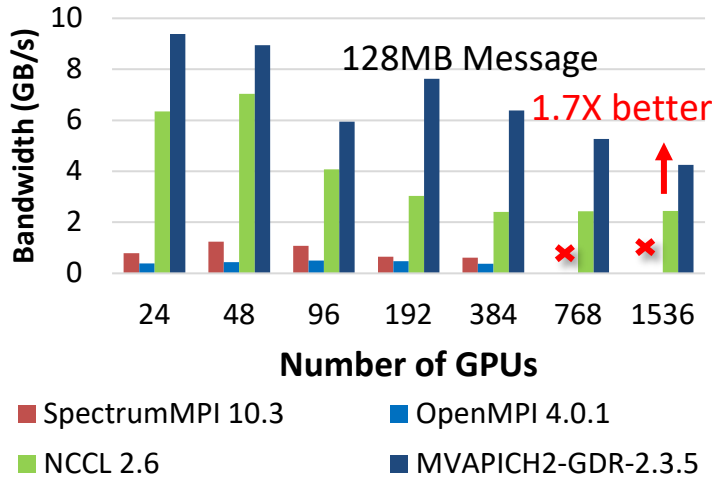
Best Performance for GPU-based Transfers



TensorFlow Training with MVAPICH2-GDR on Summit



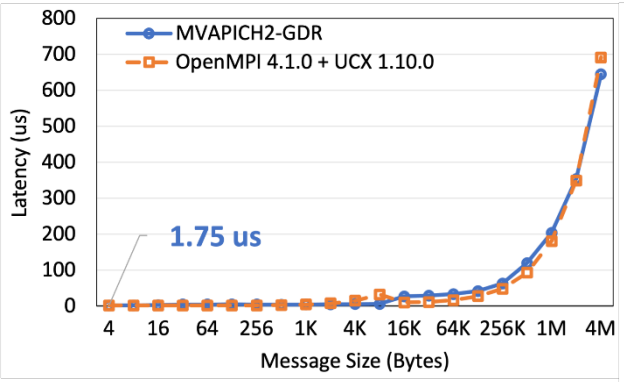
GPU-Based MPI_Allreduce on Summit



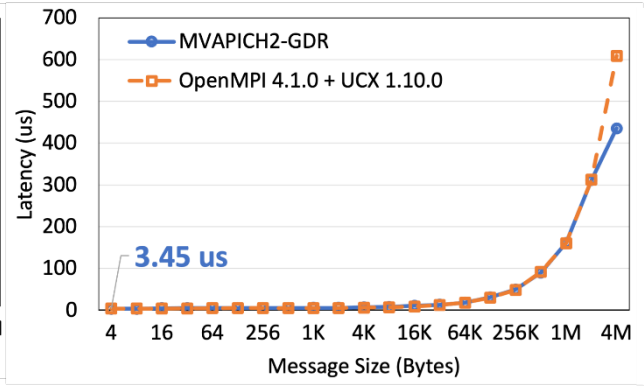
ROCm Support for AMD GPUs (Available with MVAPICH2-GDR 2.3.5)

LLNL Corona Cluster - ROCm-4.0.0 (mi50 AMD GPUs)

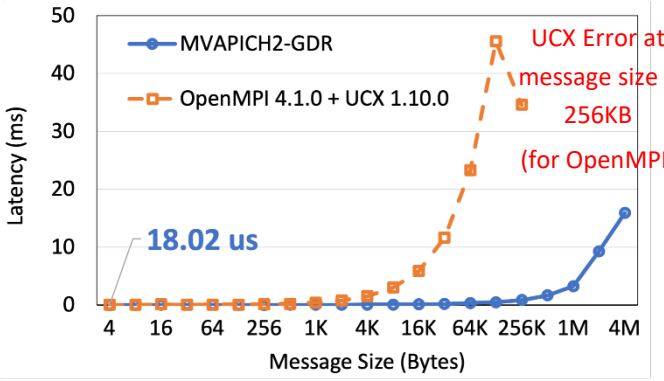
Intra-Node Point-to-Point Latency



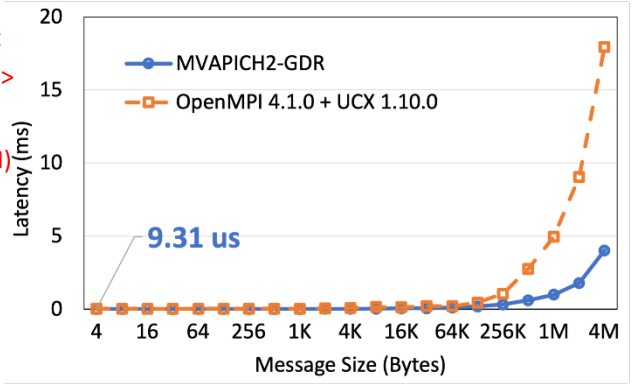
Inter-Node Point-to-Point Latency



Allreduce 128 GPUs (16 Nodes, 8 GPN)



Broadcast 128 GPUs (16 Nodes, 8 GPN)



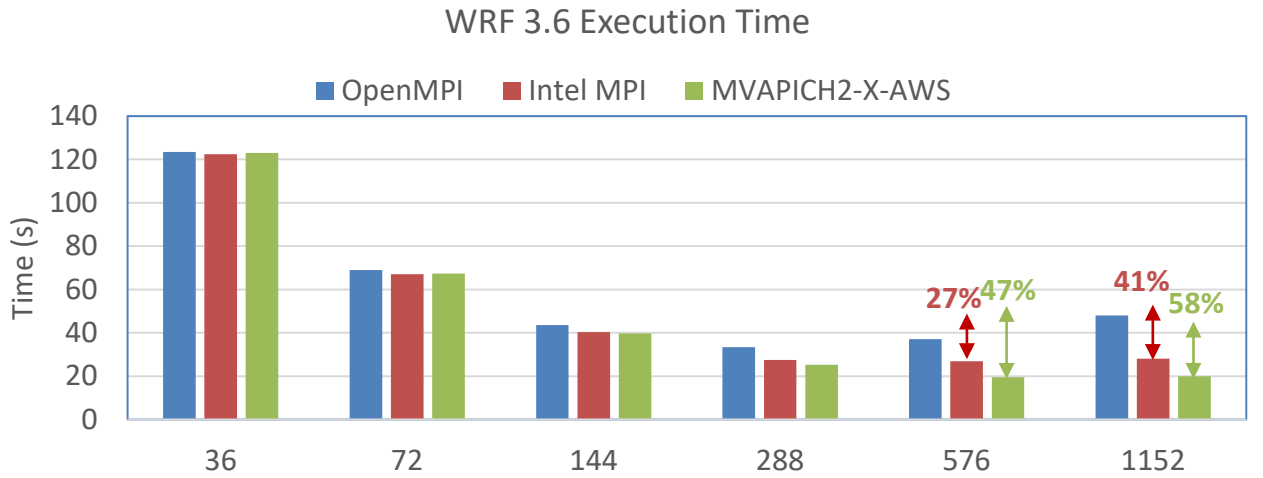
PyTorch at Scale: Training ResNet-50 on 256 GPUs:

LLNL Lassen (10,000 Images/sec faster than NCCL training)

Distributed Framework	Torch.distributed		Horovod		DeepSpeed	
Images/sec on 256 GPUs	61,794	**72,120**	74,063	**84,659**	80,217	**88,873**
Communication Backend	NCCL 2.7	MVAPICH2-GDR	NCCL 2.7	MVAPICH2-GDR	NCCL 2.7	MVAPICH2-GDR

MVAPICH2-X Advanced Support for HPC-Clouds

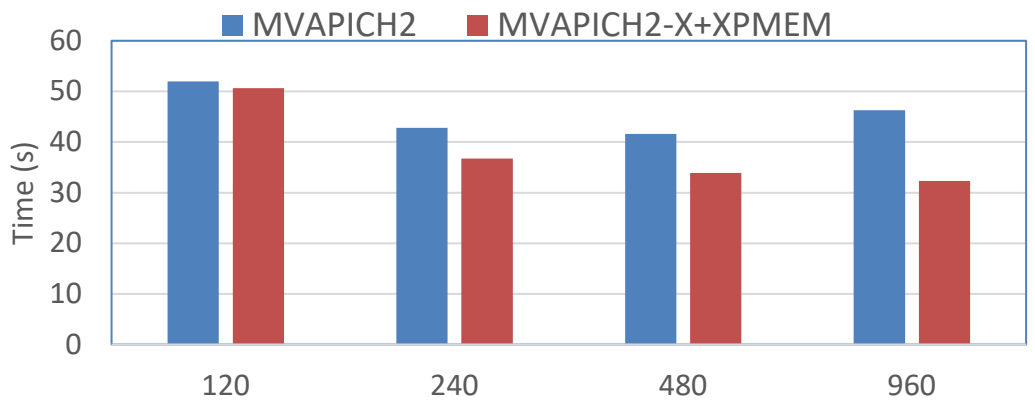
Performance on Amazon EFA



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

- MVAPICH2-X-AWS 2.3
- Released on 09/24/2020
- 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 & p3d instance types and different Operation Systems

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)

- MVAPICH2-Azure 2.3.3
  - Released on 05/20/2020
  - Integerated Azure CentOS HPC Images:

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

- MVAPICH2-X 2.3.RC3
- Major Features and Enhancements
  - **Based on MVAPICH2-2.3.3**
  - **Enhanced tuning for point-to-point and collective operations**
  - **Targeted for Azure HB & HC virtual machine instances**
  - **Tested with Azure HB & HC VM instances**
- Detailed User Guide: http://mvapich.cse.ohio-state.edu/userguide/mv2-azure/

MVAPICH2 – Future Roadmap and Plans for Exascale

- Update to MPICH 3.3 CH3 channel
  - 2021
- Initial support for the CH4 channel
  - 2021/2022
- Making CH4 channel default
  - 2022/2023
- 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
  - GenZ*
  - 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!

panda@cse.ohio-state.edu

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

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/