MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiBD
High-Performance
Big Data

HiDL
High-Performance
Deep Learning

Experiences in Designing, Developing, Packaging, and Deploying the MVAPICH2 Libraries

Talk at E4S Forum (September '19)

by

Hari Subramoni

The Ohio State University

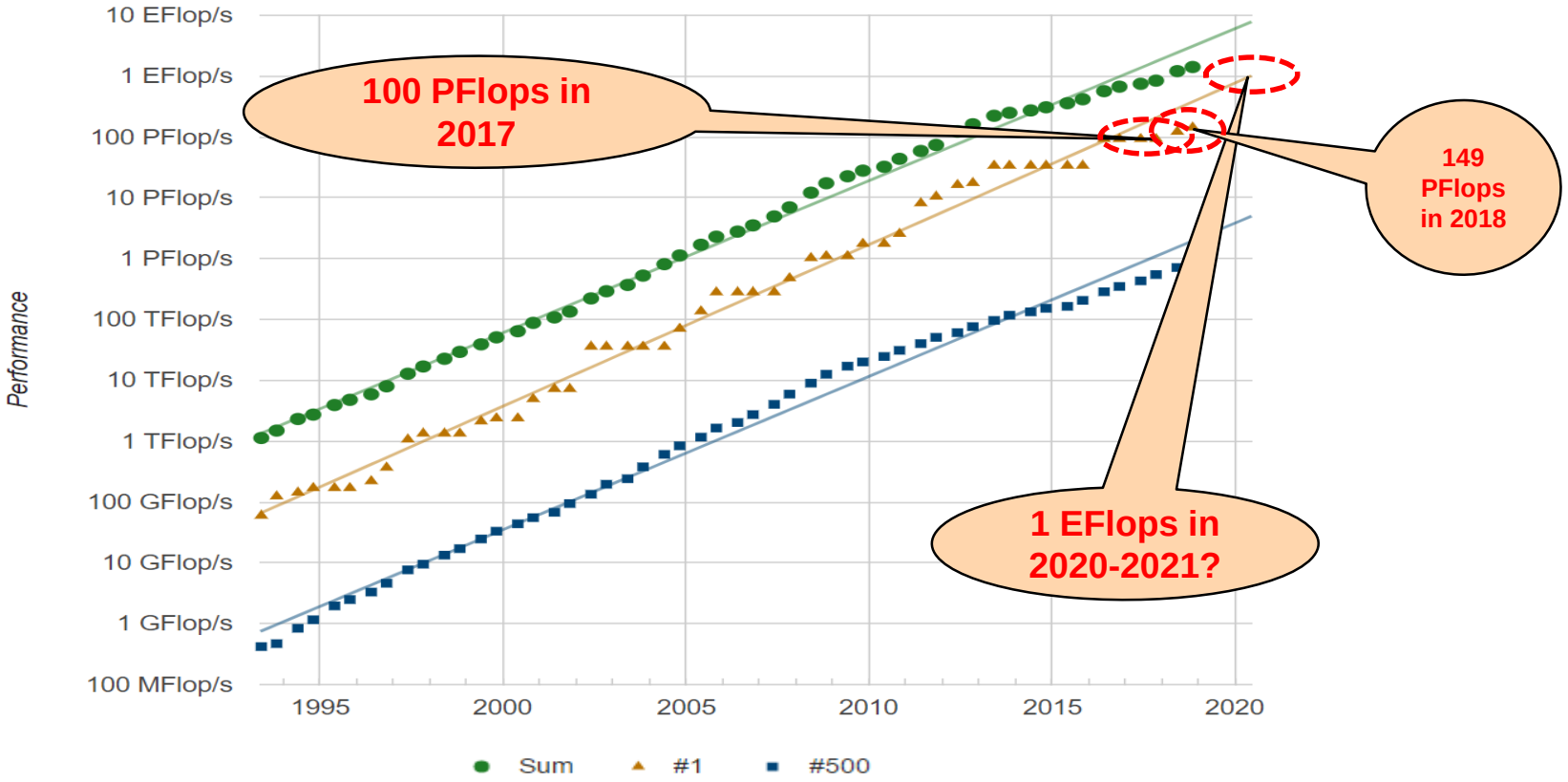
E-mail: subramon@cse.ohio-state.edu

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

Follow us on

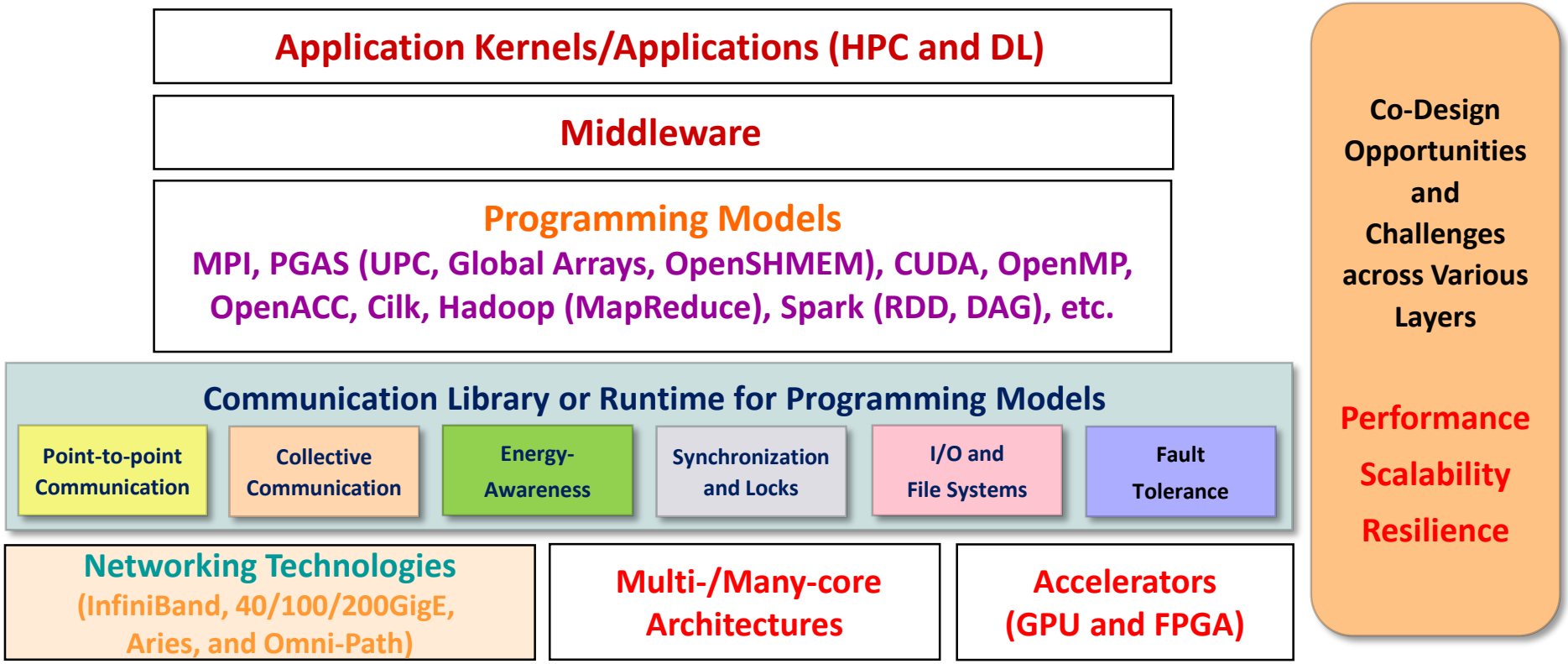
https://twitter.com/mvapich

High-End Computing (HEC): PetaFlop to ExaFlop



Expected to have an ExaFlop system in 2020-2021!

Supporting Programming Models for Multi-Petaflop and Exaflop Systems: Challenges



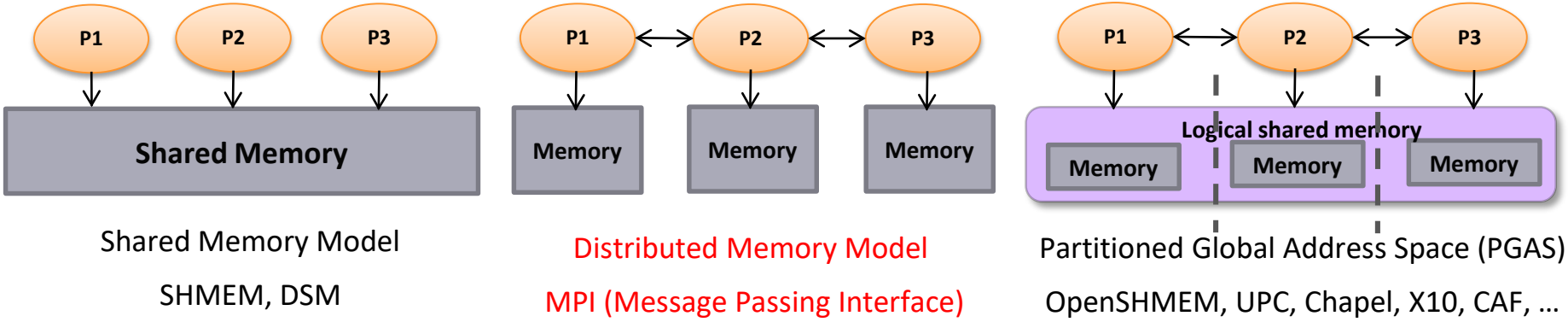
Designing (MPI+X) at Exascale

- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Scalable job start-up
  - Low memory footprint
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Balancing intra-node and inter-node communication for next generation nodes (128-1024 cores)
  - Multiple end-points per node
- Support for efficient multi-threading
- Integrated Support for Accelerators (GPGPUs and FPGAs)
- Fault-tolerance/resiliency
- QoS support for communication and I/O
- Support for Hybrid MPI+PGAS programming (MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, MPI+UPC++, CAF, …)
- Virtualization
- Energy-Awareness

Presentation Overview

- **MVAPICH Project**
  - **MPI and PGAS Library with CUDA-Awareness**
- HiDL Project
  - High-Performance Deep Learning
- Public Cloud Deployment
  - Microsoft-Azure and Amazon-AWS
- Deployment Solutions
- Conclusions

Parallel Programming Models Overview



Shared Memory Model

SHMEM, DSM

Distributed Memory Model

MPI (Message Passing Interface)

Partitioned Global Address Space (PGAS)

OpenSHMEM, UPC, Chapel, X10, CAF, …

- Programming models provide abstract machine models
- Models can be mapped on different types of systems
  - e.g. Distributed Shared Memory (DSM), MPI within a node, etc.
- PGAS models and Hybrid MPI+PGAS models are gradually receiving importance

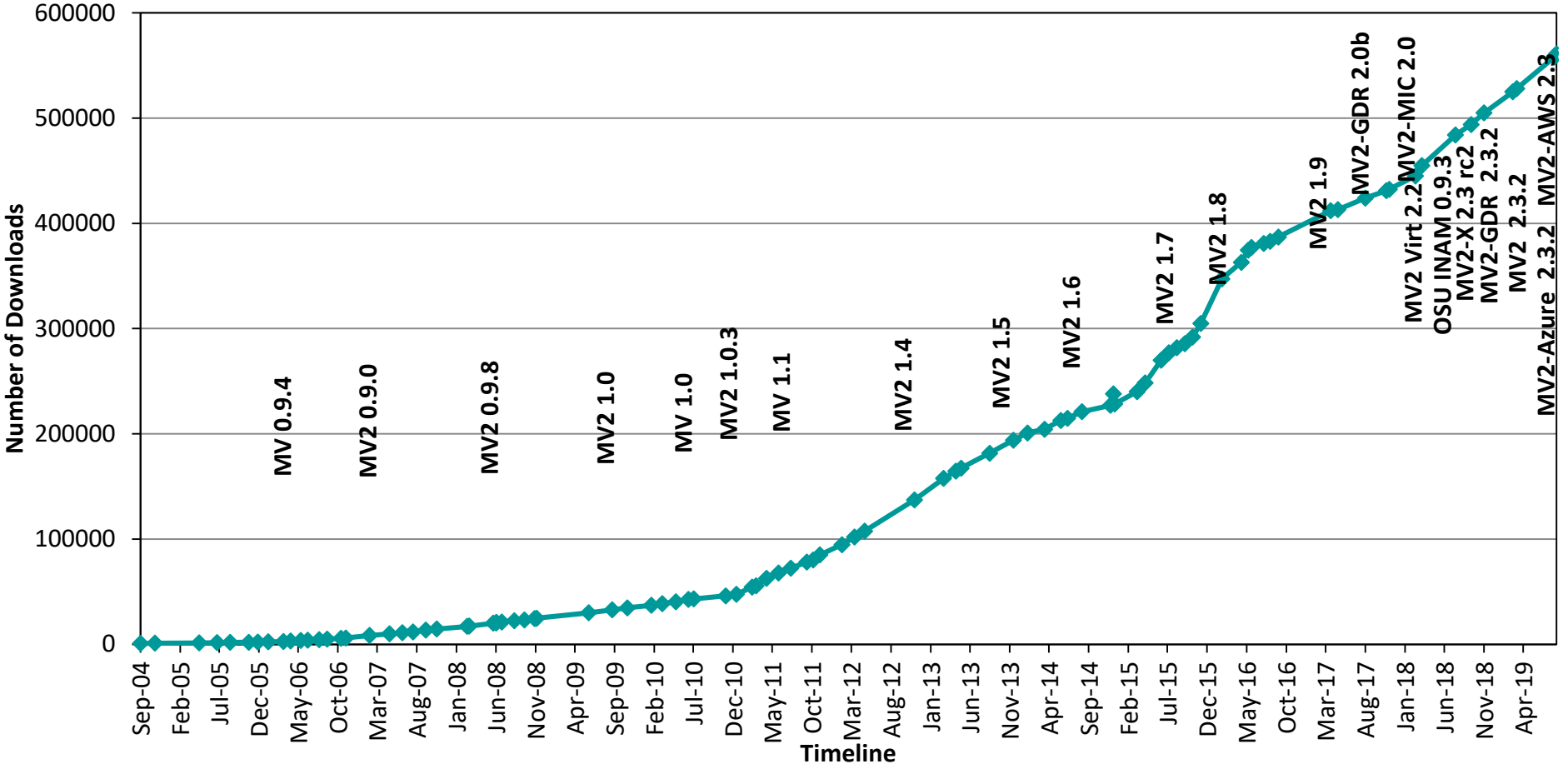
Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.1), Started in 2001, First version available in 2002
  - 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 3,025 organizations in 89 countries**
  - **More than 589,000 (> 0.5 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '18 ranking)
    - 3rd, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China
    - 5th, 448, 448 cores (Frontera) at TACC
    - 8th, 391,680 cores (ABCI) in Japan
    - 15th, 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**

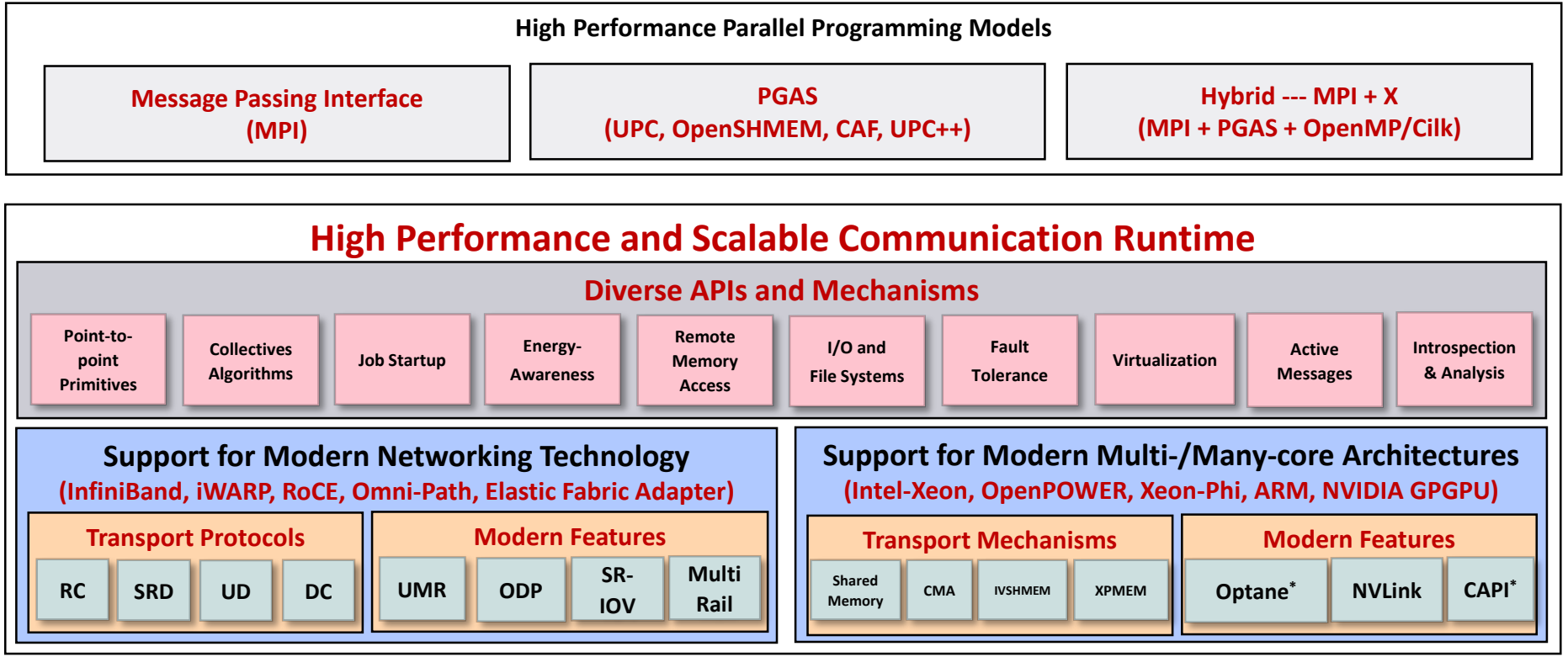
Partner in the TACC Frontera System

- Empowering Top500 systems for over a decade

MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family



* Upcoming

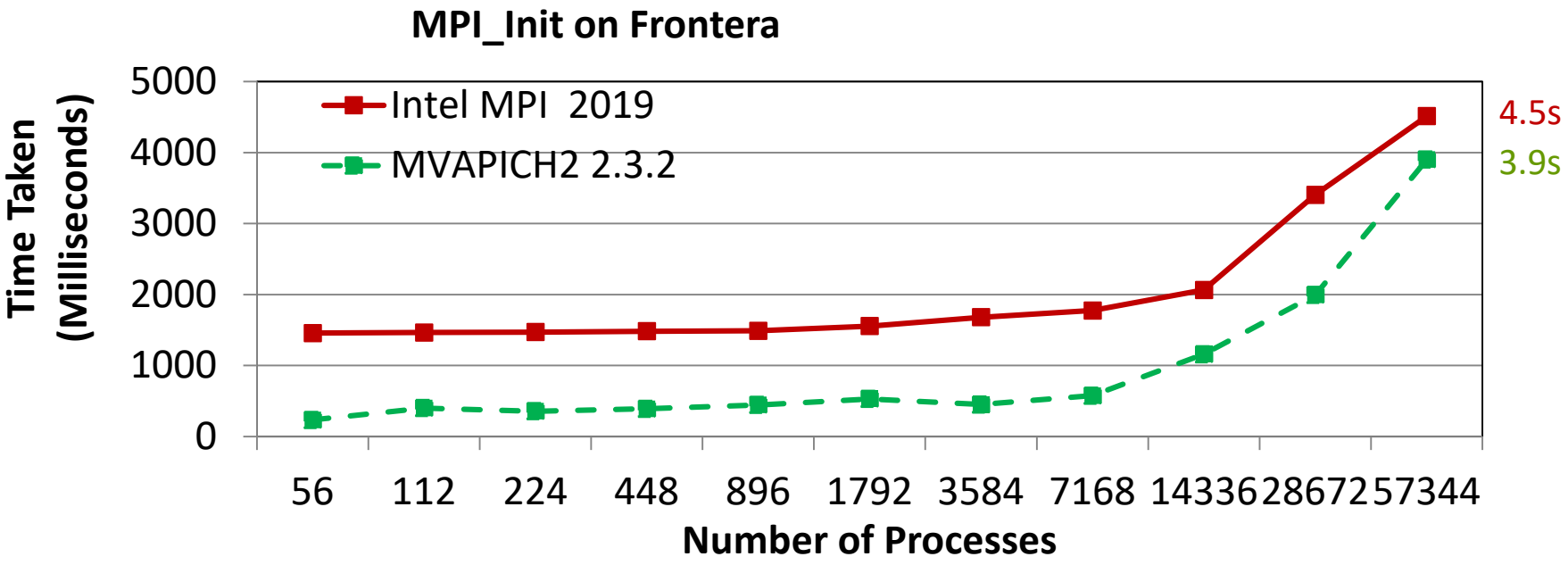
Strong Procedure for Design, Development and Release

- Research is done for exploring new designs
- Designs are first presented to conference/journal publications
- Best performing designs are incorporated into the codebase
- Rigorous Q&A procedure before making a release
  - Exhaustive unit testing
  - Various test procedures on diverse range of platforms and interconnects
  - Test 19 different benchmarks and applications including, but not limited to
    - OMB, IMB, MPICH Test Suite, Intel Test Suite, NAS, ScaLAPACK, and SPEC
  - Spend about 18,000 core hours per commit
  - Performance regression and tuning
  - Applications-based evaluation
  - Evaluation on large-scale systems (Lassen, Frontera, Summit etc)
- All versions (alpha, beta, RC1 and RC2) go through the above testing

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 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

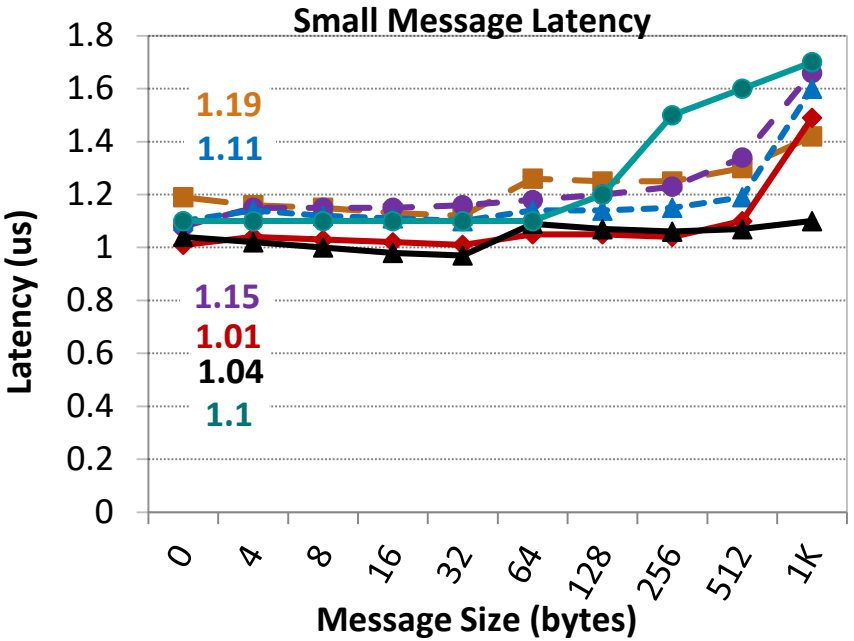
Startup Performance on TACC Frontera

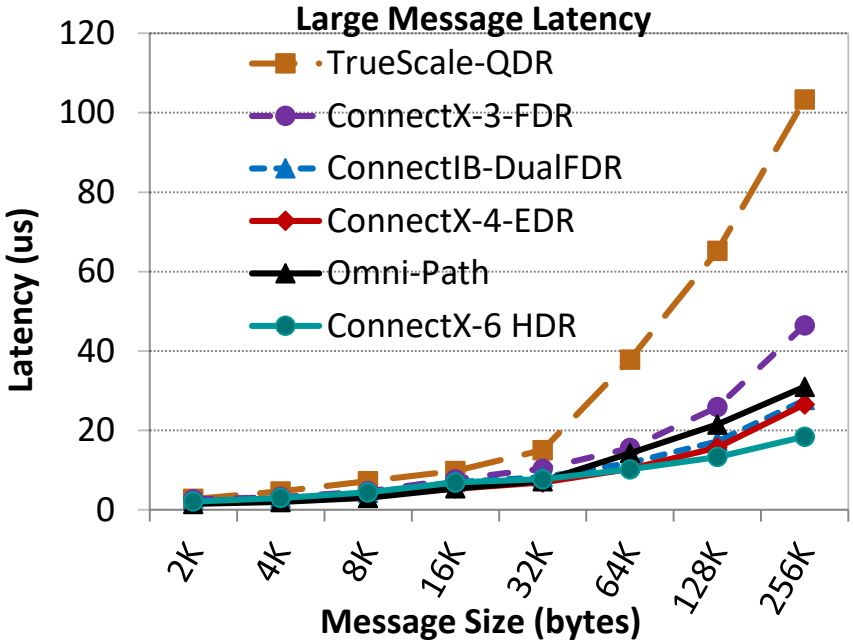


- MPI_Init takes 3.9 seconds on 57,344 processes on 1,024 nodes
- All numbers reported with 56 processes per node

New designs available in MVAPICH2-2.3.2

One-way Latency: MPI over IB with MVAPICH2





TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

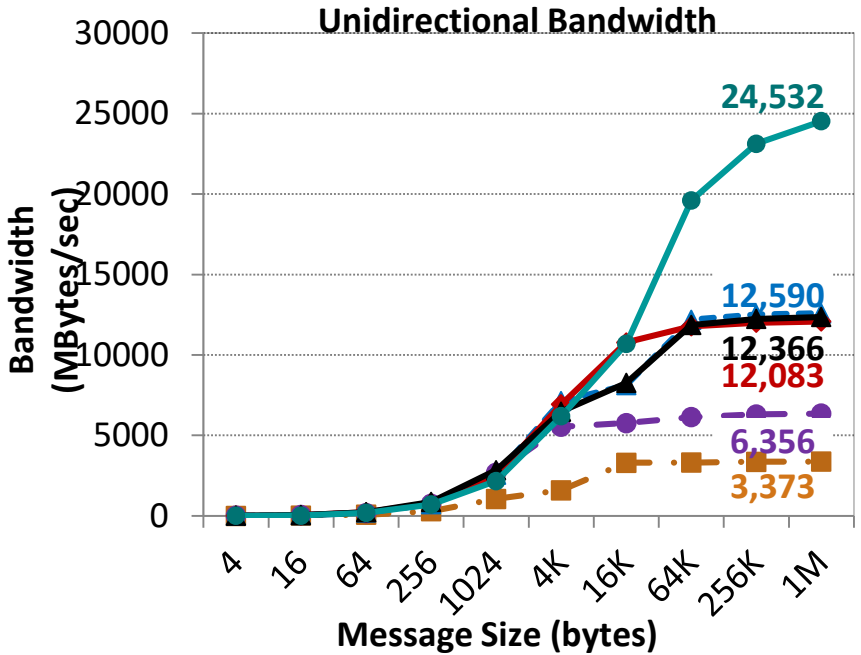
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

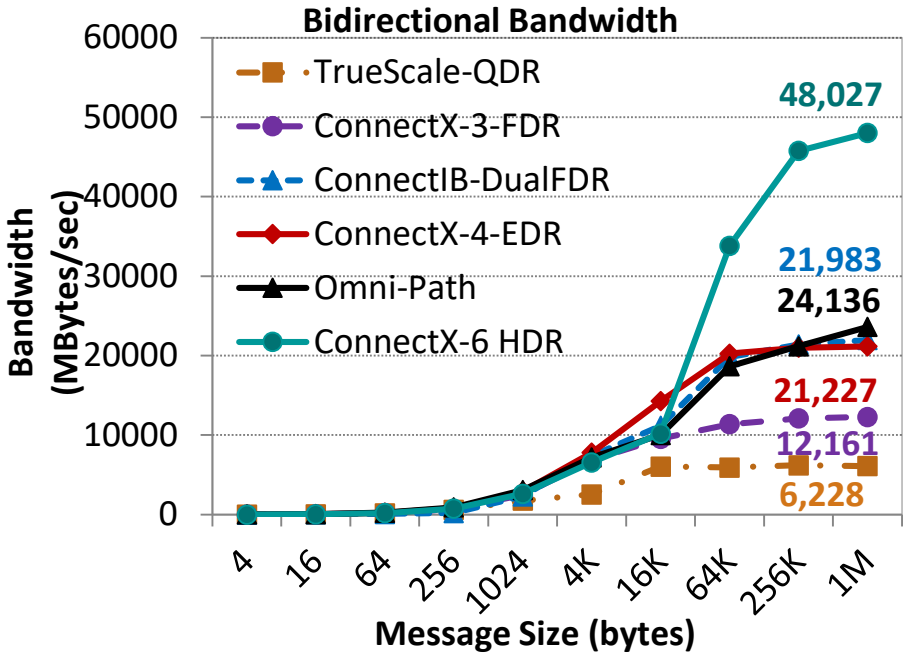
ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch

ConnectX-6-HDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

Bandwidth: MPI over IB with MVAPICH2





TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

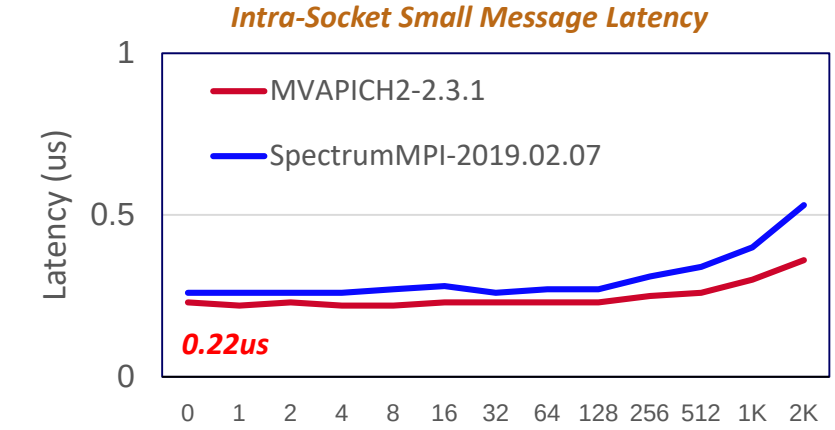
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

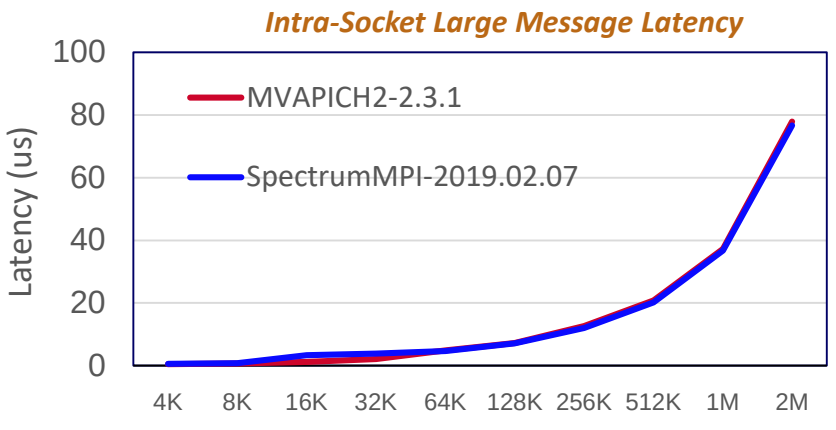
ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

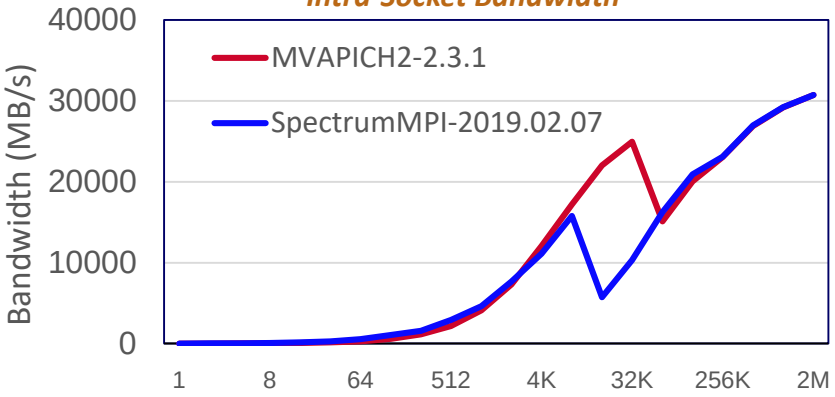
Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch

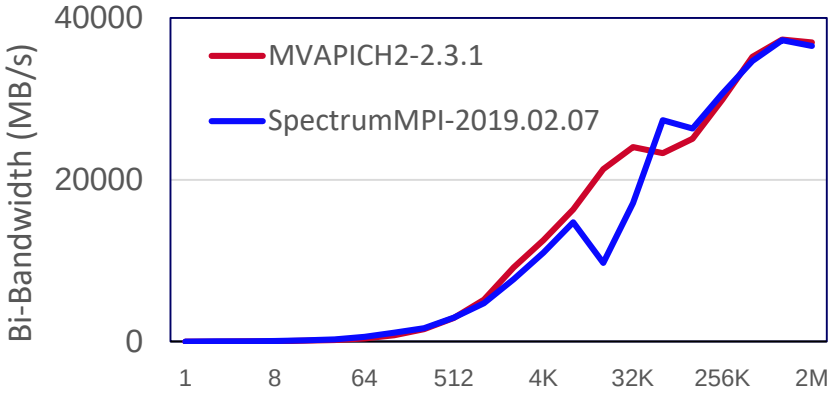
ConnectX-6-HDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

Intra-node Point-to-Point Performance on OpenPower



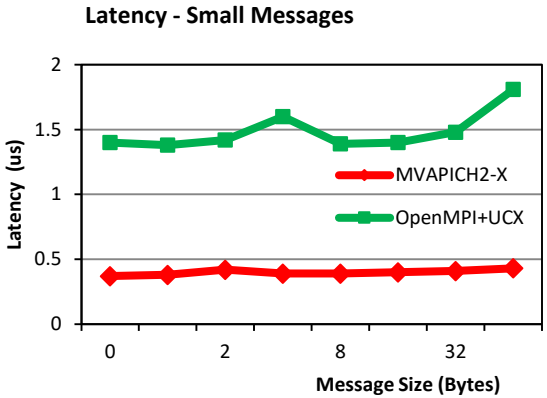


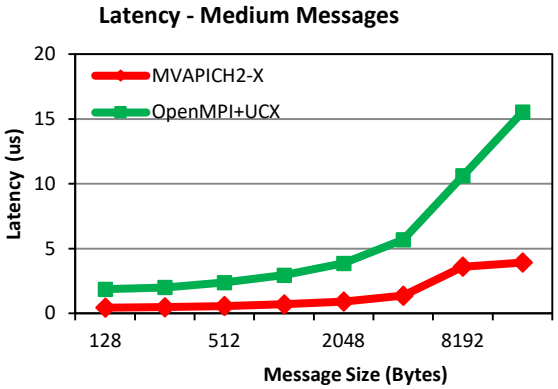


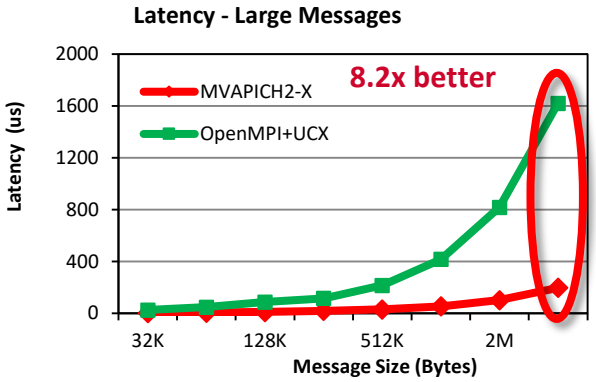


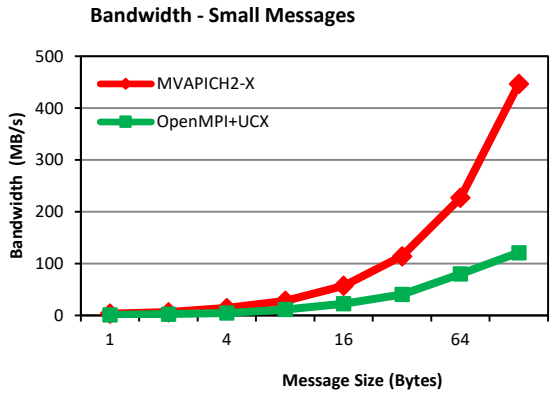
Platform: Two nodes of OpenPOWER (POWER9-ppc64le) CPU using Mellanox EDR (MT4121) HCA

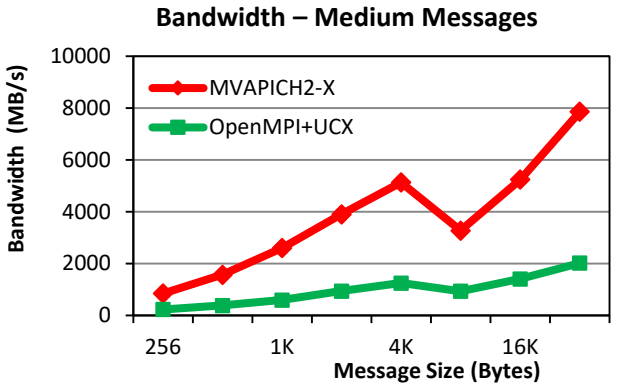
Point-to-point: Latency & Bandwidth (Intra-socket) on Mayer

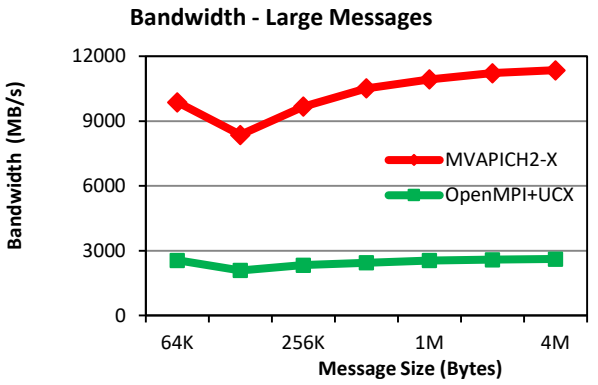




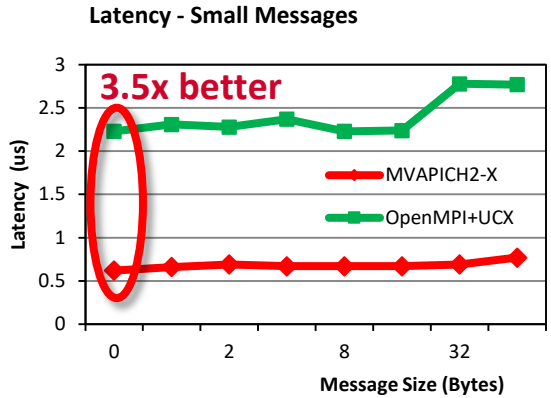


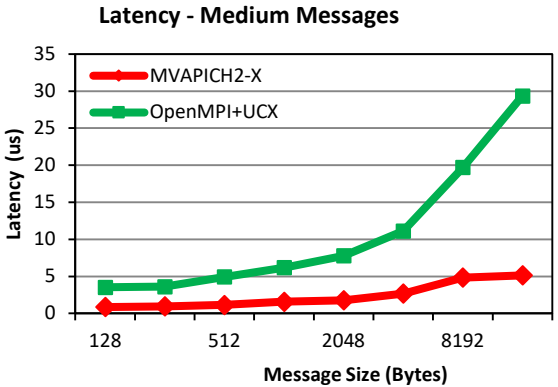


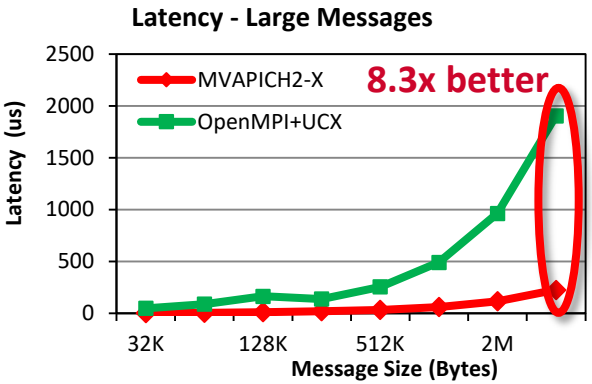


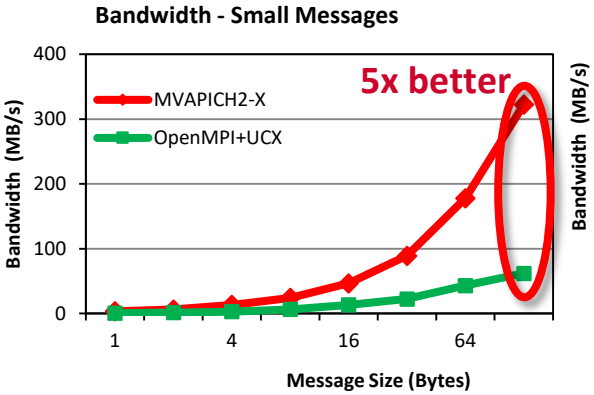


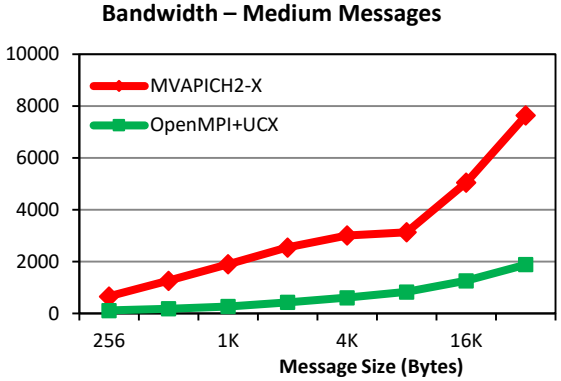
Point-to-point: Latency & Bandwidth (Inter-socket) on Mayer

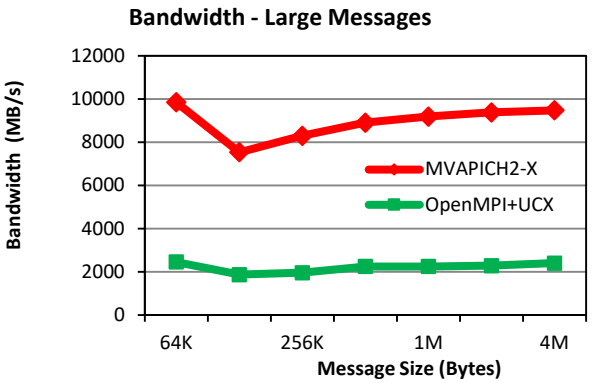




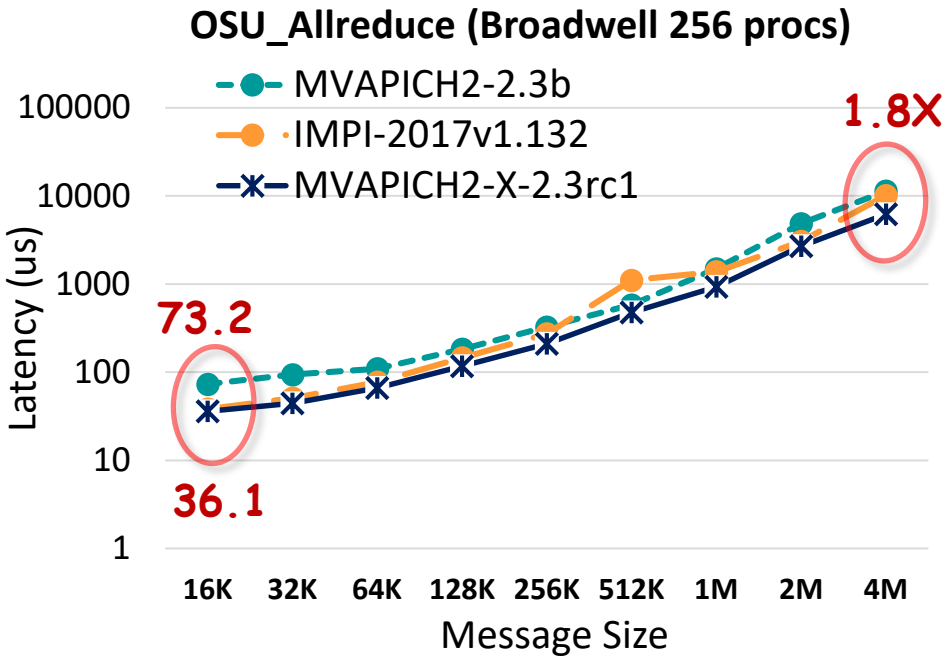


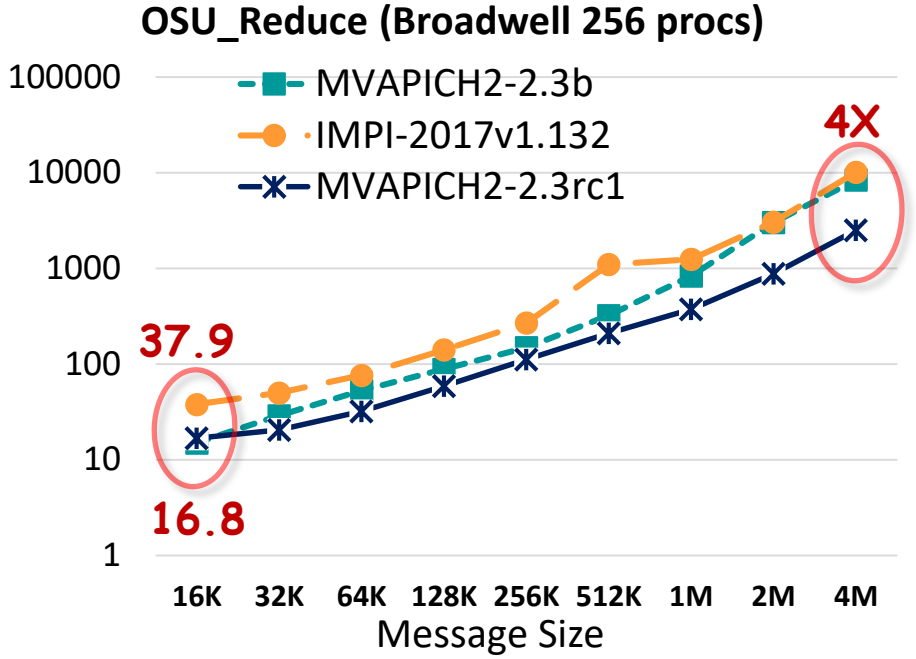






Shared Address Space (XPMEM)-based Collectives Design



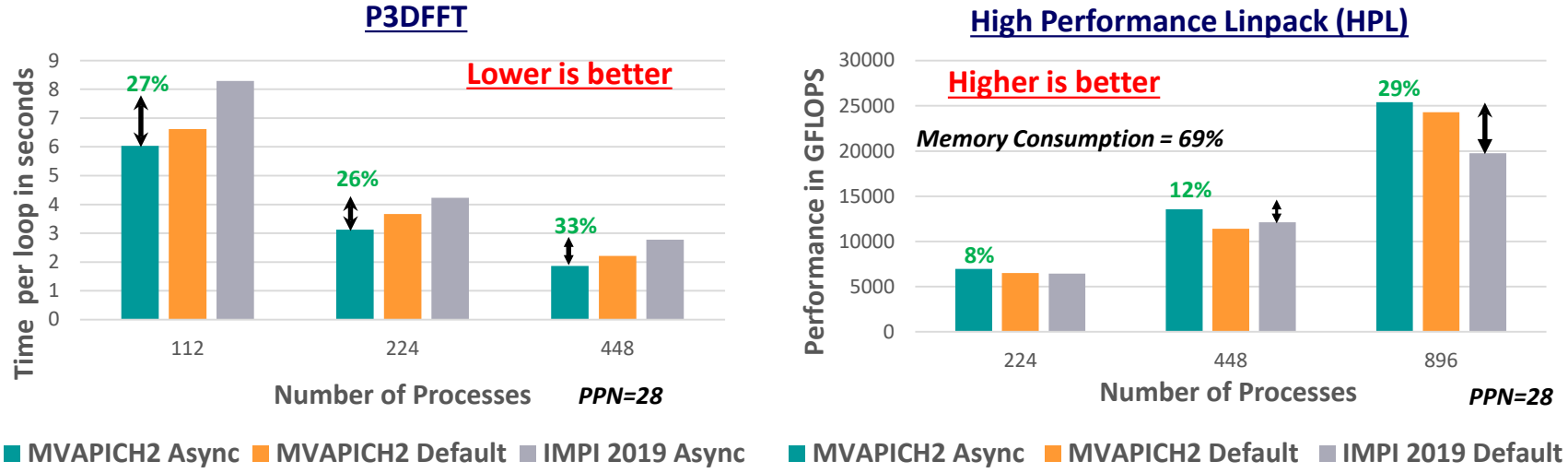


- "*Shared Address Space*"-based true *zero-copy* Reduction collective designs in MVAPICH2
- Offloaded computation/communication to peers ranks in reduction collective operation
- Up to **4X** improvement for 4MB Reduce and up to **1.8X** improvement for 4M AllReduce

J. Hashmi, S. Chakraborty, M. Bayatpour, H. Subramoni, and D. Panda, Designing Efficient Shared Address Space Reduction Collectives for Multi-/Many-cores, International Parallel & Distributed Processing Symposium (IPDPS '18), May 2018.

Available since MVAPICH2-X 2.3rc1

Benefits of the New Asynchronous Progress Design: Broadwell + InfiniBand



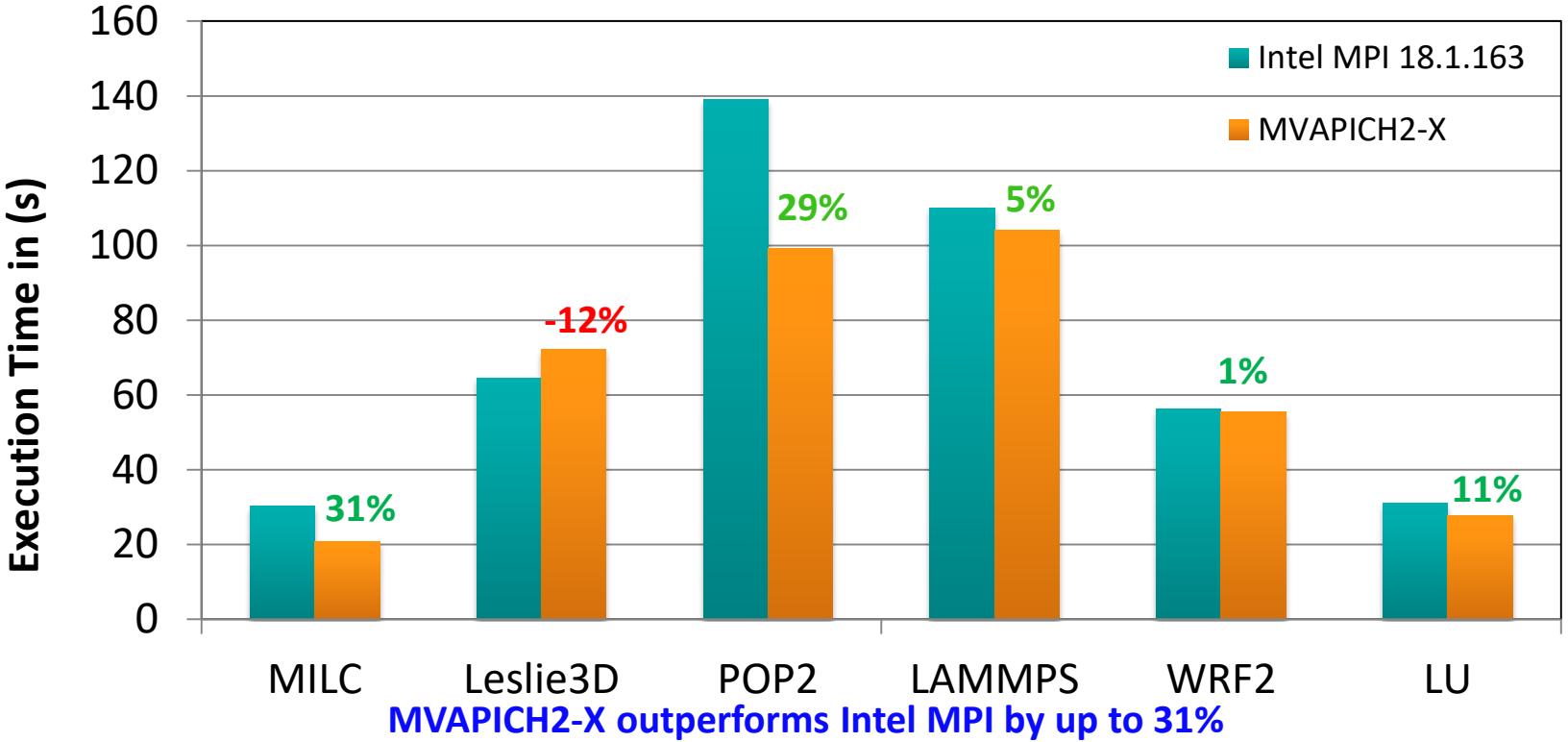
Up to 33% performance improvement in P3DFFT application with 448 processes

Up to 29% performance improvement in HPL application with 896 processes

A. Ruhela, H. Subramoni, S. Chakraborty, M. Bayatpour, P. Kousha, and D.K. Panda, Efficient Asynchronous Communication Progress for MPI without Dedicated Resources, EuroMPI 2018. Enhanced version accepted for PARCO Journal.

Available since MVAPICH2-X 2.3rc1

SPEC MPI 2007 Benchmarks: Broadwell + InfiniBand



MVAPICH2-X outperforms Intel MPI by up to 31%

Configuration: 448 processes on 16 Intel E5-2680v4 (Broadwell) nodes having 28 PPN and interconnected with 100Gbps Mellanox MT4115 EDR ConnectX-4 HCA

GPU-Aware (CUDA-Aware) MPI Library: MVAPICH2-GPU

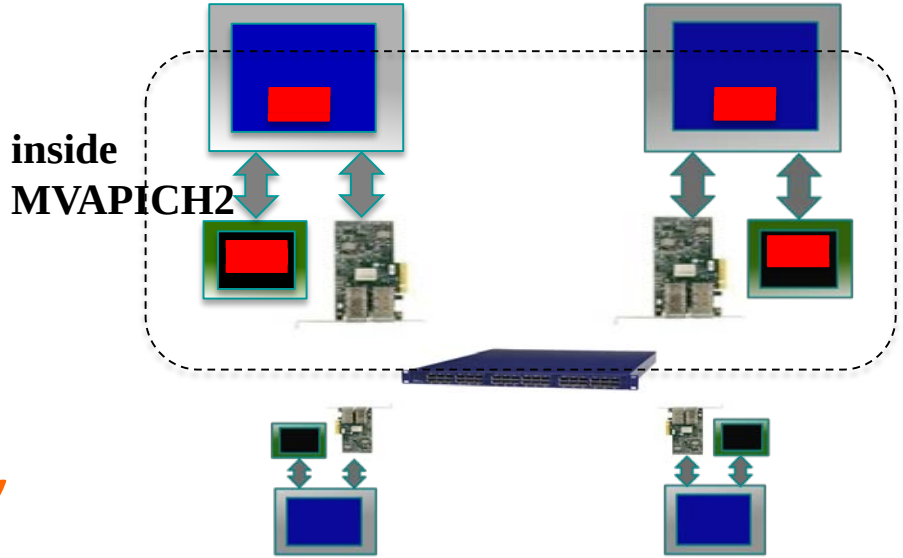
- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

At Sender:

```
MPI_Send(s_devbuf, size, …);
```

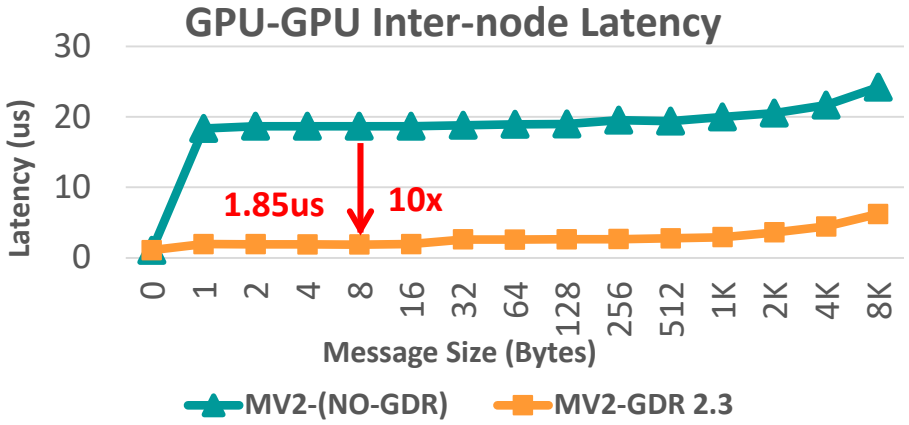
At Receiver:

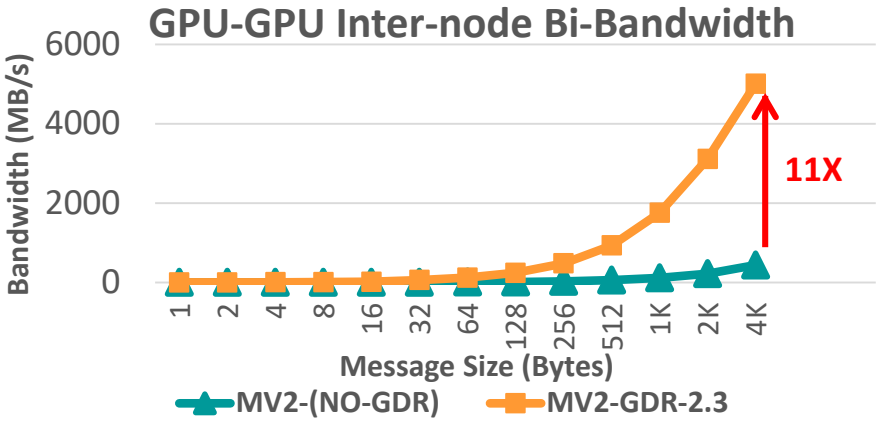
```
MPI_Recv(r_devbuf, size, …);
```

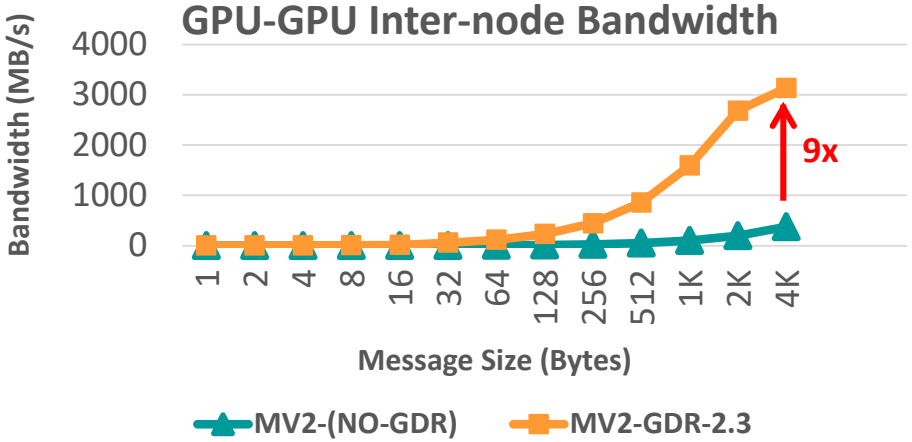


High Performance and High Productivity

Optimized MVAPICH2-GDR Design

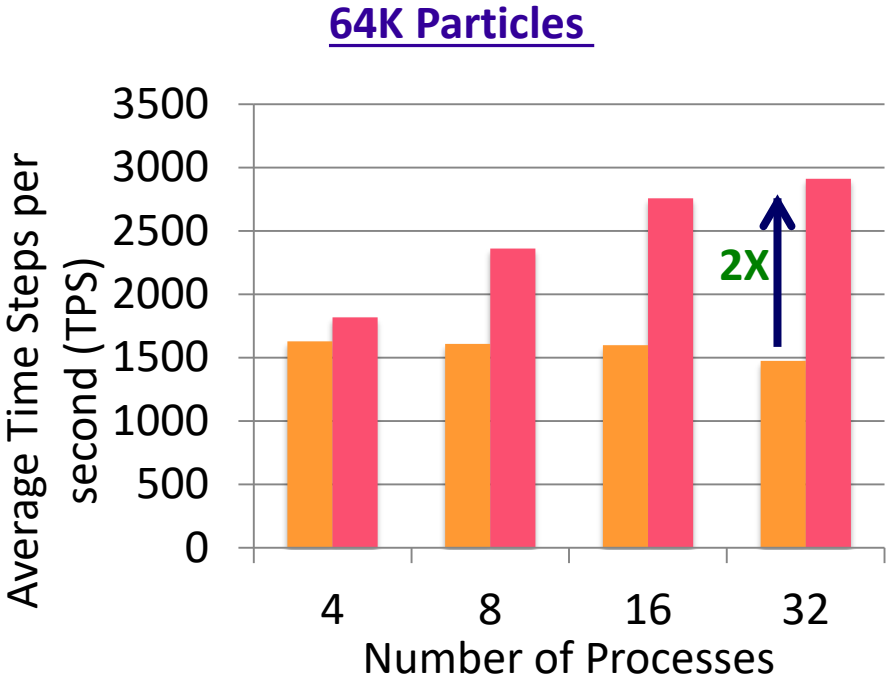


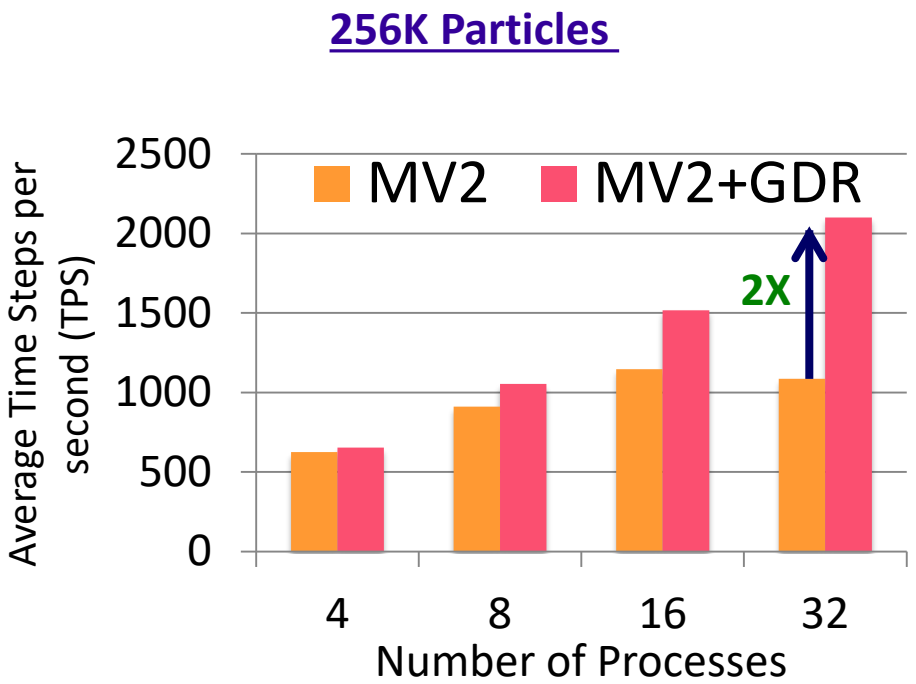




MVAPICH2-GDR-2.3
Intel Haswell (E5-2687W @ 3.10 GHz) node - 20 cores
NVIDIA Volta V100 GPU
Mellanox Connect-X4 EDR HCA
CUDA 9.0
Mellanox OFED 4.0 with GPU-Direct-RDMA

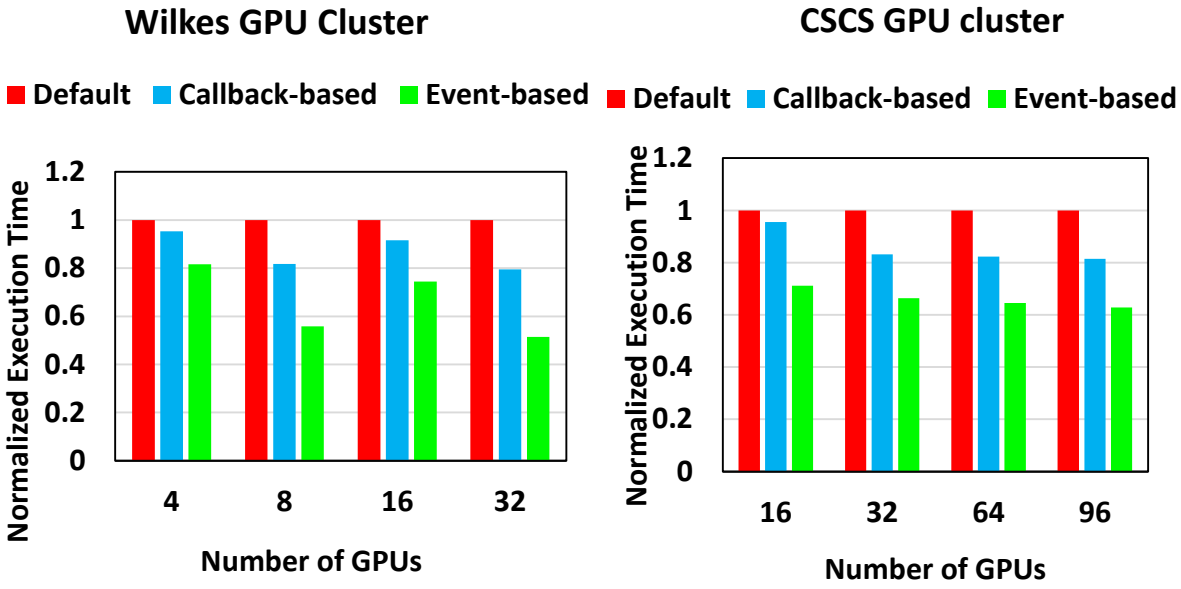
Application-Level Evaluation (HOOMD-blue)

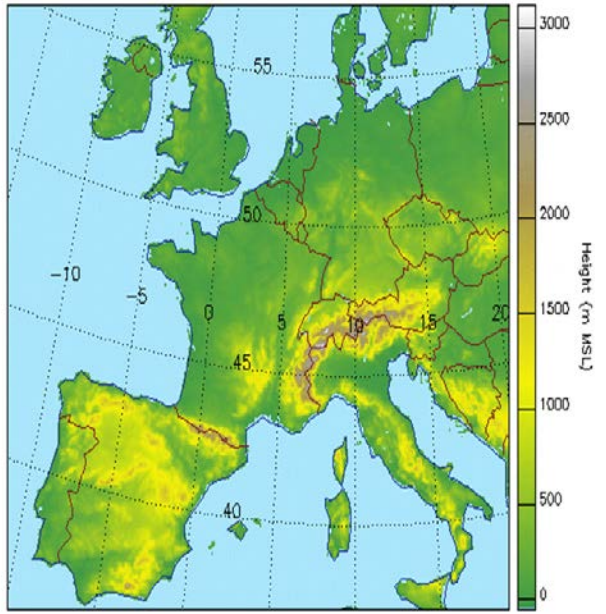




- Platform: Wilkes (Intel Ivy Bridge + NVIDIA Tesla K20c + Mellanox Connect-IB)
- HoomdBlue Version 1.0.5
  - GDRCOPY enabled: MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_0 MV2_IBA_EAGER_THRESHOLD=32768 MV2_VBUF_TOTAL_SIZE=32768 MV2_USE_GPUDIRECT_LOOPBACK_LIMIT=32768 MV2_USE_GPUDIRECT_GDRCOPY=1 MV2_USE_GPUDIRECT_GDRCOPY_LIMIT=16384

Application-Level Evaluation (Cosmo) and Weather Forecasting in Switzerland





- 2X improvement on 32 GPUs nodes
- 30% improvement on 96 GPU nodes (8 GPUs/node)

Cosmo model: http://www2.cosmo-model.org/content/tasks/operational/meteoSwiss/

On-going collaboration with CSCS and MeteoSwiss (Switzerland) in co-designing MV2-GDR and Cosmo Application

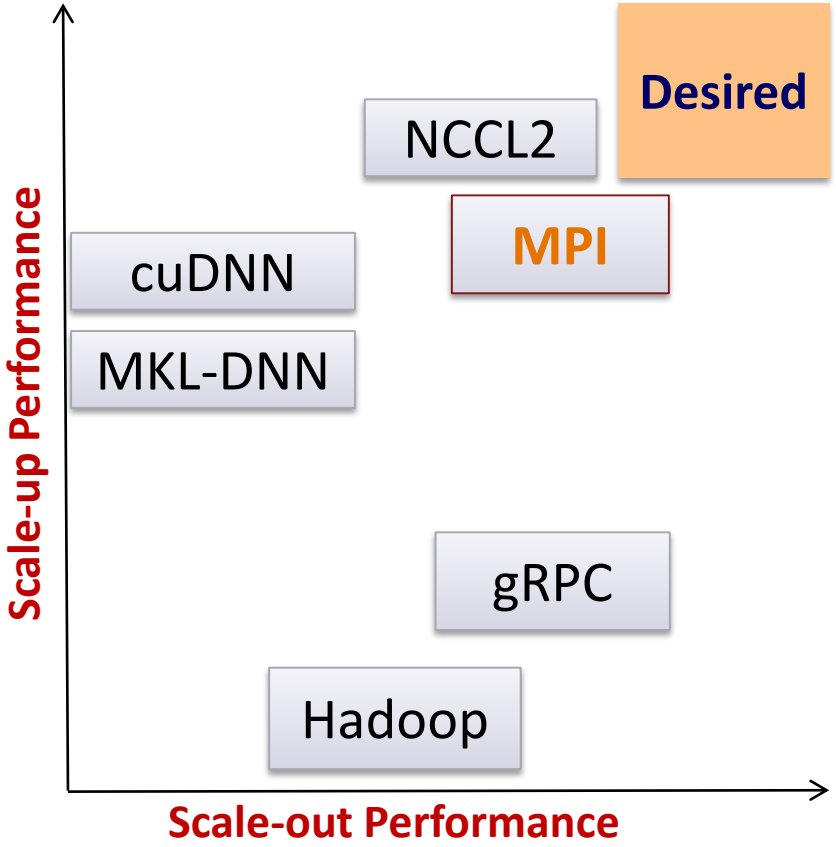
C. Chu, K. Hamidouche, A. Venkatesh, D. Banerjee , H. Subramoni, and D. K. Panda, Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, IPDPS'16

Presentation Overview

- MVAPICH Project
  - MPI and PGAS Library with CUDA-Awareness
- **HiDL Project**
  - **High-Performance Deep Learning**
- Public Cloud Deployment
  - Microsoft-Azure and Amazon-AWS
- Deployment Solutions
- Conclusions

Deep Learning: New Challenges for MPI Runtimes

- Deep Learning frameworks are a different game altogether
  - Unusually large message sizes (order of megabytes)
  - Most communication based on GPU buffers
- Existing State-of-the-art
  - cuDNN, cuBLAS, NCCL --> **scale-up** performance
  - NCCL2, CUDA-Aware MPI --> **scale-out** performance
    - For small and medium message sizes only!
- Proposed: Can we **co-design** the MPI runtime (**MVAPICH2-GDR**) and the DL framework (**Caffe**) to achieve both?
  - Efficient **Overlap** of Computation and Communication
  - Efficient **Large-Message** Communication (Reductions)
  - What **application co-designs** are needed to exploit **communication-runtime co-designs**?



A. A. Awan, K. Hamidouche, J. M. Hashmi, and D. K. Panda, S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP '17)

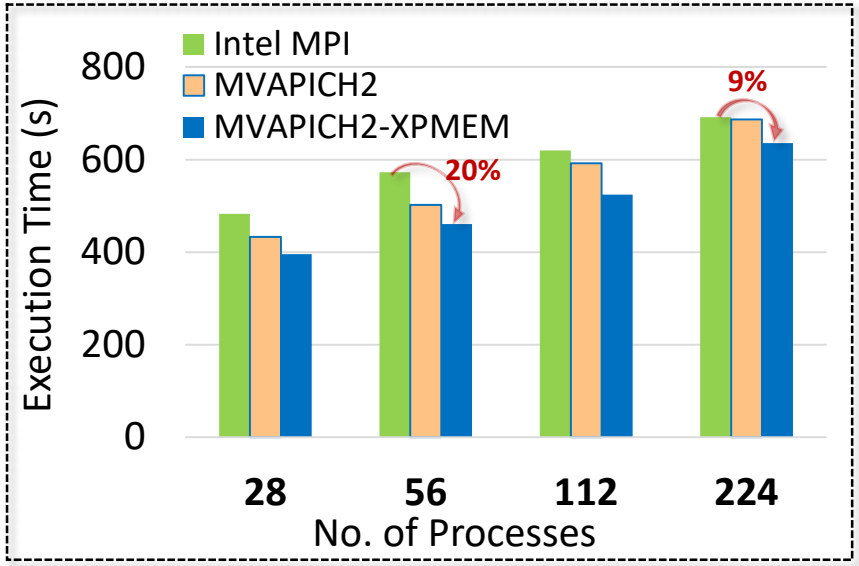
High-Performance Deep Learning

- CPU-based Deep Learning
- GPU-based Deep Learning

Performance of CNTK with MVAPICH2-X on CPU-based Deep Learning

- CPU-based training of AlexNet neural network using ImageNet ILSVRC2012 dataset
- Advanced XPMEM-based designs show up to 20% benefits over Intel MPI (IMPI) for CNTK DNN training using All_Reduce
- The proposed designs show good scalability with increasing system size

CNTK AlexNet Training
(B.S=default, iteration=50, ppn=28)

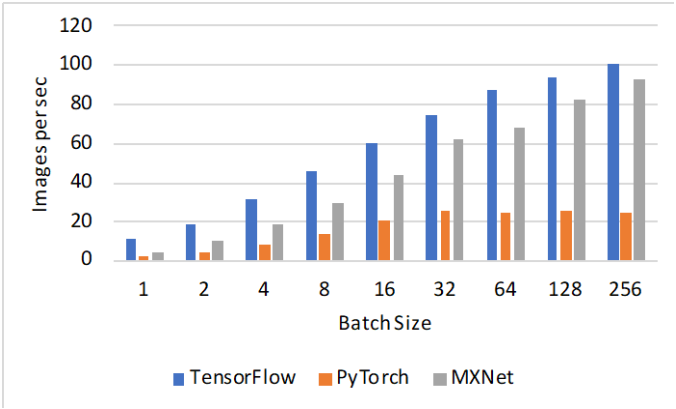


Available since MVAPICH2-X 2.3rc1 release

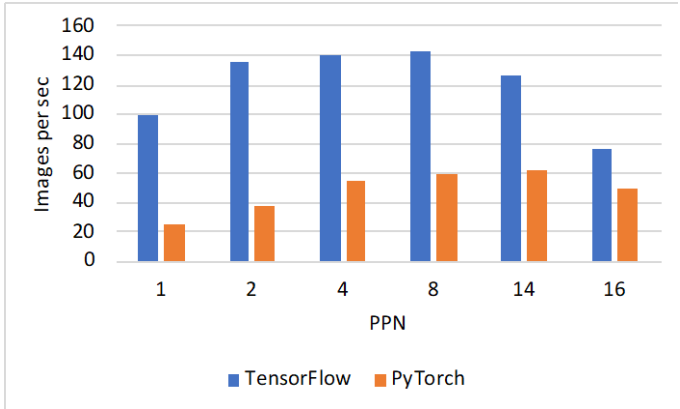
Designing Efficient Shared Address Space Reduction Collectives for Multi-/Many-cores, J. Hashmi, S. Chakraborty, M. Bayatpour, H. Subramoni, and DK Panda, 32nd IEEE International Parallel & Distributed Processing Symposium (IPDPS '18), May 2018

Deep Learning on Frontera

- TensorFlow, PyTorch, and MXNet are widely used Deep Learning Frameworks
- Optimized by Intel using Math Kernel Library for DNN (MKL-DNN) for Intel processors
- Single Node performance can be improved by running Multiple MPI processes



Impact of Batch Size on Performance for ResNet-50

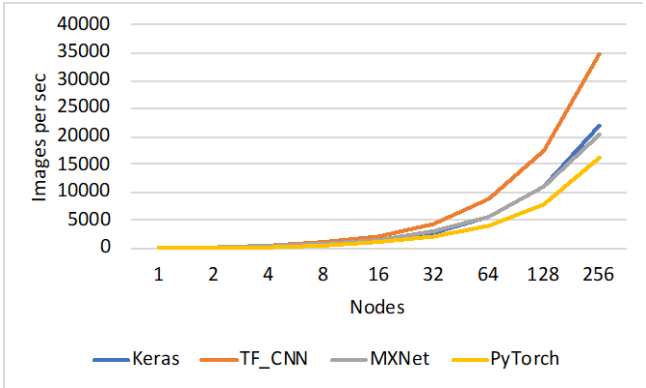


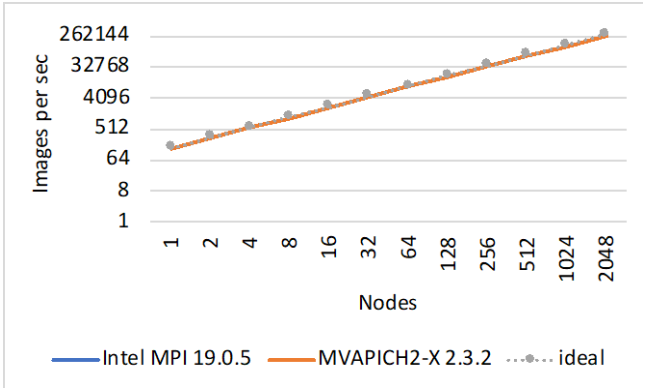
Performance Improvement using Multiple MPI processes

A. Jain et al., Scaling Deep Learning Frameworks on Frontera using MVAPICH2 MPI, under review

Deep Learning on Frontera

- Observed 260K images per sec for ResNet-50 on 2,048 Nodes
- Scaled MVAPICH2-X on 2,048 nodes on Frontera for Distributed Training using TensorFlow
- ResNet-50 can be trained in 7 minutes on 2048 nodes (114,688 cores)

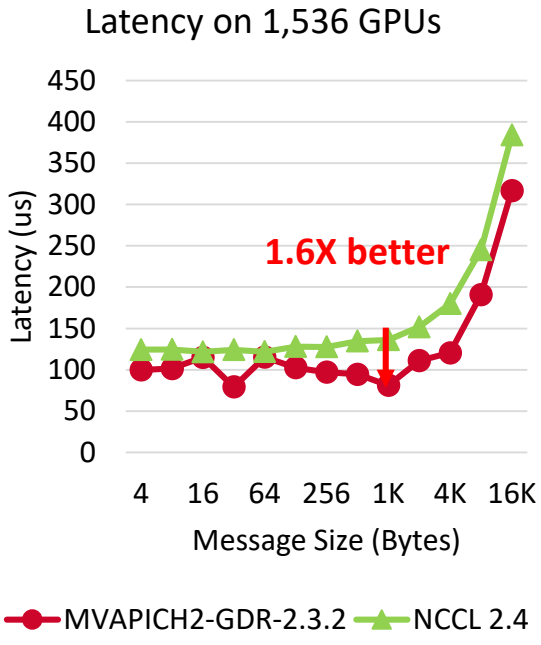


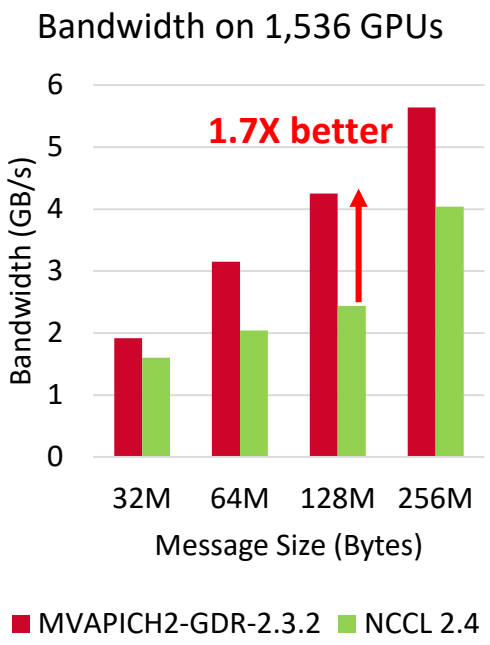


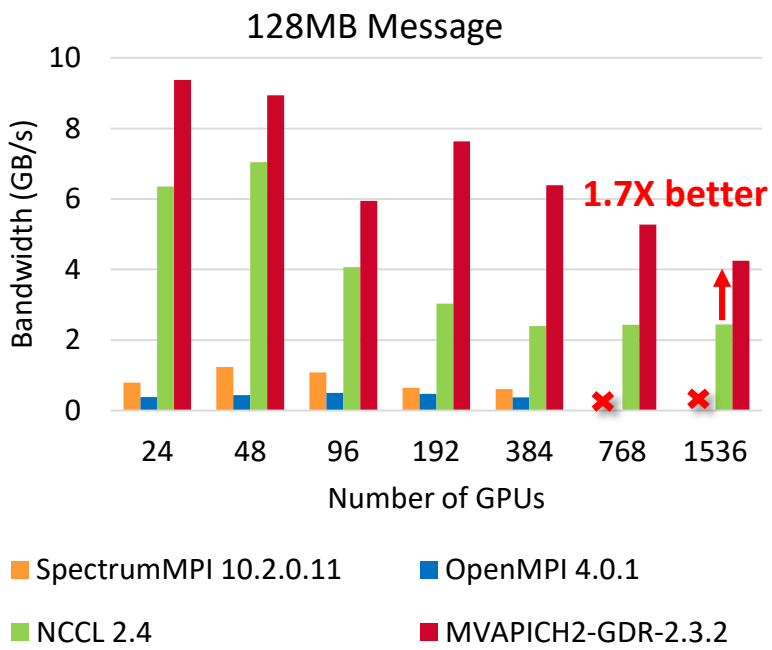
A. Jain et al., Scaling Deep Learning Frameworks on Frontera using MVAPICH2 MPI, under review

MVAPICH2-GDR: Enhanced MPI_Allreduce at Scale

- **Optimized designs in upcoming MVAPICH2-GDR offer better performance for most cases**
- **MPI_Allreduce (MVAPICH2-GDR) vs. ncclAllreduce (NCCL2) up to 1,536 GPUs**



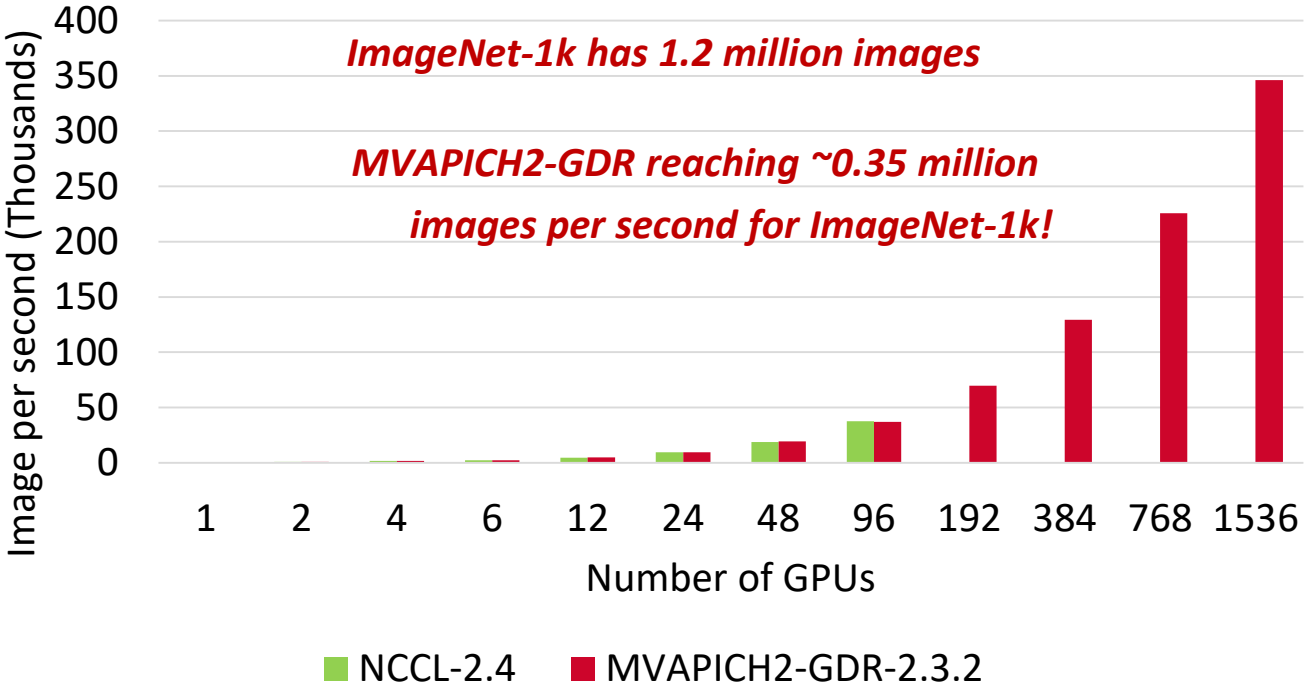




Platform: Dual-socket IBM POWER9 CPU, 6 NVIDIA Volta V100 GPUs, and 2-port InfiniBand EDR Interconnect

Distributed Training with TensorFlow and MVAPICH2-GDR

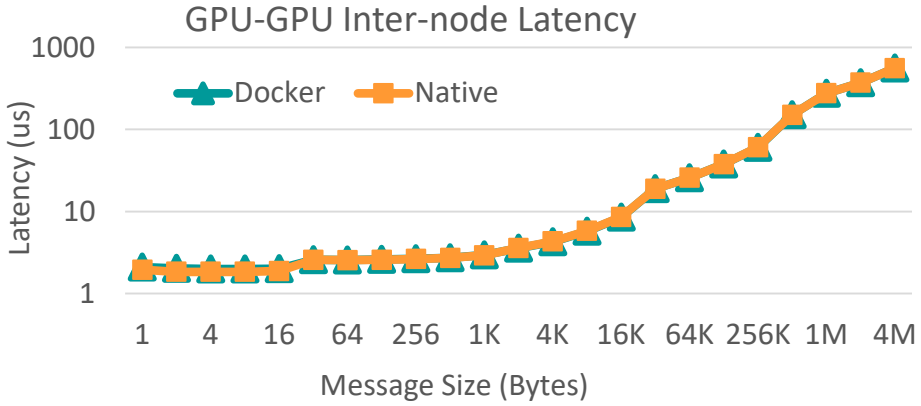
- ResNet-50 Training using TensorFlow benchmark on SUMMIT -- 1536 Volta GPUs!
- 1,281,167 (1.2 mil.) images
- Time/epoch = 3.6 seconds
- Total Time (90 epochs) = 3.6 x 90 = 332 seconds = **5.5 minutes!**

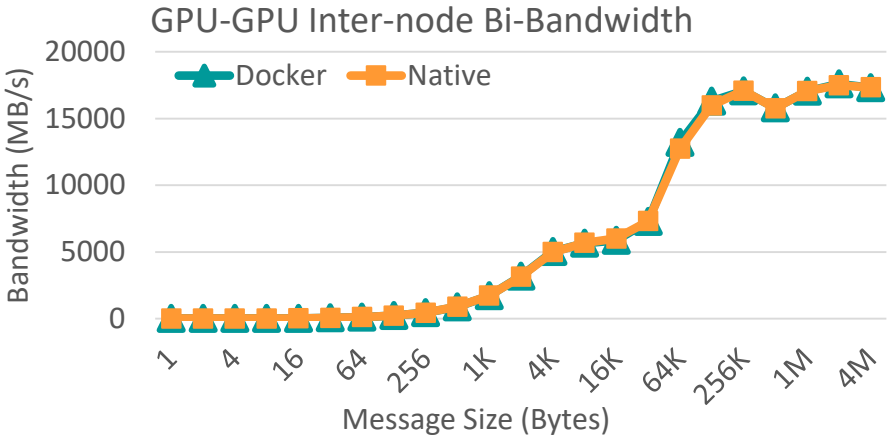


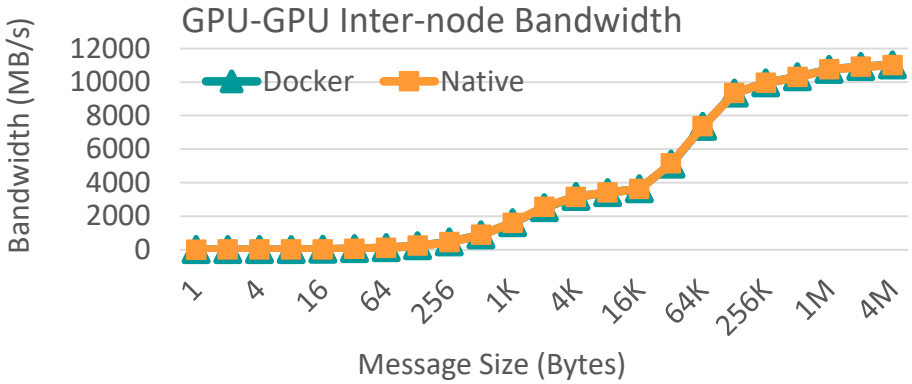
*We observed errors for NCCL2 beyond 96 GPUs

Platform: The Summit Supercomputer (#1 on Top500.org) – 6 NVIDIA Volta GPUs per node connected with NVLink, CUDA 9.2

MVAPICH2-GDR on Container with Negligible Overhead







MVAPICH2-GDR-2.3.2
Intel Haswell (E5-2687W @ 3.10 GHz) node - 20 cores
NVIDIA Volta V100 GPU
Mellanox Connect-X4 EDR HCA
CUDA 9.0
Mellanox OFED 4.0 with GPU-Direct-RDMA

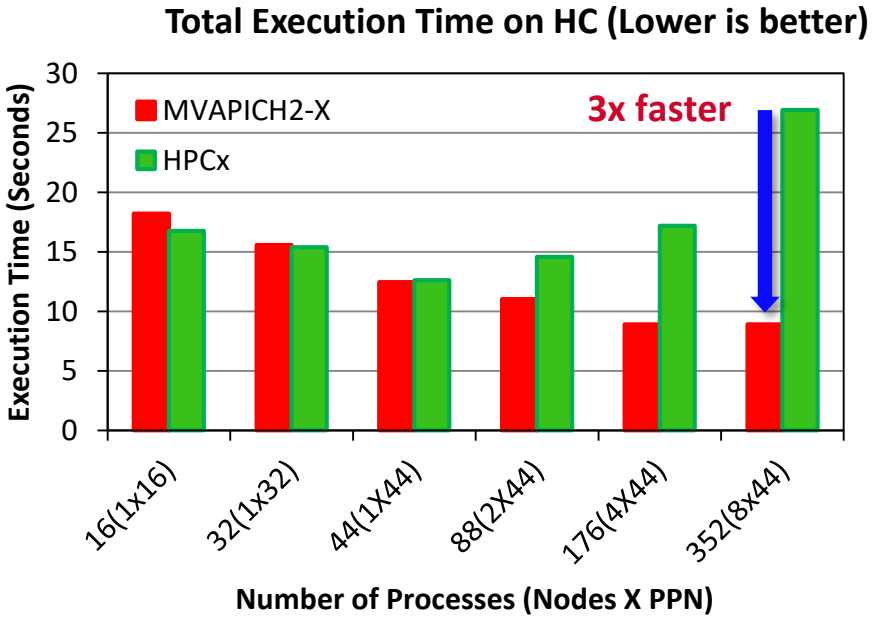
Presentation Overview

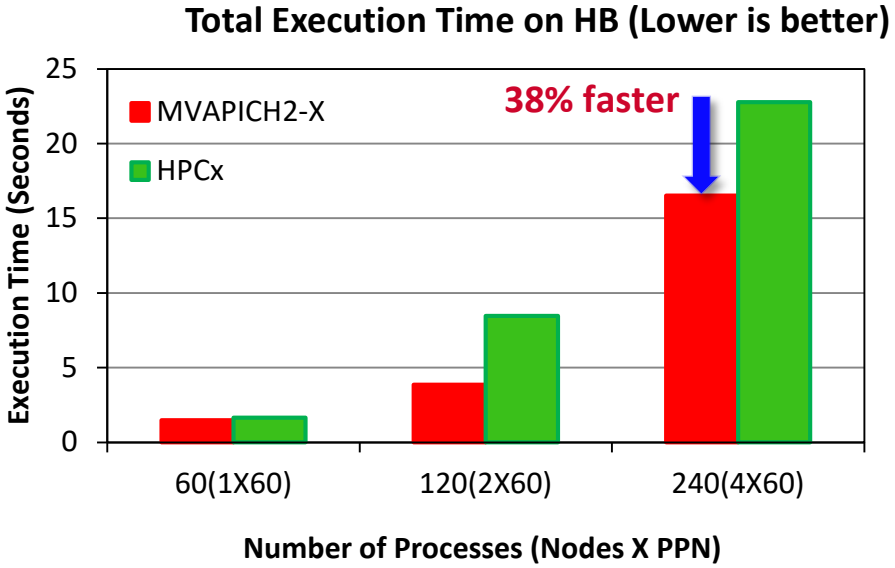
- MVAPICH Project
  - MPI and PGAS Library with CUDA-Awareness
- HiDL Project
  - High-Performance Deep Learning
- **Public Cloud Deployment**
  - **Microsoft-Azure and Amazon-AWS**
- Deployment Solutions
- Conclusions

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**

Performance of Radix

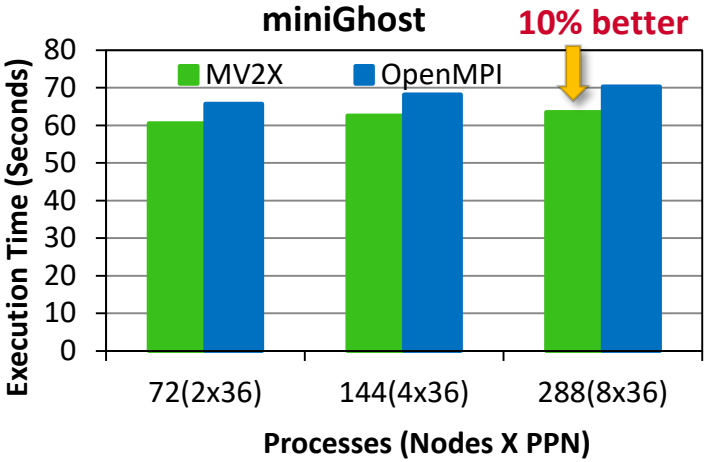


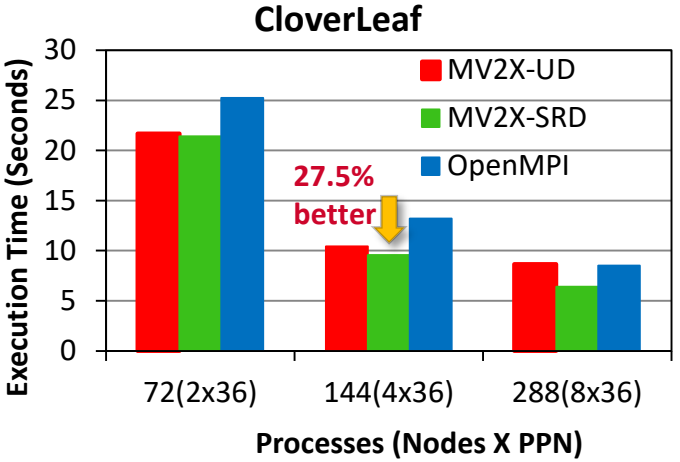


MVAPICH2-X-AWS 2.3

- **Released on 08/12/2019**
- Major Features and Enhancements
  - **Based on MVAPICH2-X 2.3**
  - **New design based on Amazon EFA adapter's Scalable Reliable Datagram (SRD) transport protocol**
  - **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**

Application Performance





- Up to 10% performance improvement for MiniGhost on 8 nodes
- Up to 27% better performance with CloverLeaf on 8 nodes

S. Chakraborty, S. Xu, H. Subramoni and D. K. Panda, Designing Scalable and High-Performance MPI Libraries on Amazon Elastic Adapter, Hot Interconnect, 2019

Presentation Overview

- MVAPICH Project
  - MPI and PGAS Library with CUDA-Awareness
- HiBD Project
  - High-Performance Big Data Analytics Library
- HiDL Project
  - High-Performance Deep Learning
- Public Cloud Deployment
  - Microsoft-Azure and Amazon-AWS
- **Deployment Solutions**
  - **RPM, OpenHPC, Spack**
- Conclusions

RPM and Debian Deployments

- Provide customized RPMs for different system requirements
  - ARM, Power8, Power9, x86 (Intel and AMD)
  - Different versions of Compilers (ICC, PGI, GCC, XLC, ARM), CUDA, OFED/Intel IFS

Deploy to Azure

MVAPICH2-GDR 2.3.2 Library

- The MVAPICH2-GDR library is distributed under the BSD License.
- OSU MVAPICH2-GDR 2.3.2 (08/08/2019), ABI compatible with MPICH-3.2.1
  - CHANGELOG for MVAPICH2-GDR 2.3.2.
- These RPMs contain the MVAPICH2-GDR software on the corresponding distro. **Please note that the RHEL RPMs are compatible with CentOS as well. For Debian/Ubuntu users, please follow the instructions in the install section in the userguide.**

OpenPOWER RPMs

	GNU 4.9.3	GNU 4.9.3 (w/ jsrun)	GNU 7.3.1	GNU 7.3.1 (w/ jsrun)	PGI 18.7	PGI 18.7 (w/ jsrun)	PGI 19.4	PGI 19.4 (w/ jsrun)
MLNX-OFED 4.3(Lassen/Sierra)	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 10.1]	[CUDA 10.1]
	GNU 4.8.5	**GNU 6.4.0**	**GNU 7.4.0**	**PGI 18.7**	**PGI 19.4**			
MLNX-OFED 4.5(Summit)	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 10.1]			

RHEL/CENTOS 7 RPMs

	GNU 4.8.5 (w/o SLURM)	GNU 4.8.5 (w/ SLURM)	GNU 4.8.5 (w/ PBS)	PGI (w/o SLURM)	PGI (w/ SLURM)	PGI (w/ PBS)
MLNX-OFED 4.X*	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2]	[CUDA 9.2] [CUDA 10.1]	[CUDA 9.2] [CUDA 10.1]	

*Note that the MOFED 3.X RPMs were built against MOFED 3.4 and the MOFED 4.X RPMs were built against MOFED 4.5
- However, these RPMs should work against the other MOFEDs with the same major MOFED version number
- e.g MOFED 4.X RPMs should work if you have MOFED 4.0, MOFED 4.2, MOFED 4.4, or MOFED 4.5
- Please email mvapich-help@cse.ohio-state.edu if you encounter any issues

MVAPICH2-X 2.3rc2 Library and User Guide

- The MVAPICH2-X 2.3rc2 library is distributed under the BSD License.
- OSU MVAPICH2-X 2.3rc2 (04/02/19), ABI compatible with MPICH-3.2.
  - CHANGELOG for MVAPICH2-X 2.3rc2
  - Patch to add PMI Extensions with SLURM 15
  - Patch to add PMI Extensions with SLURM 16
  - Patch to add PMI Extensions with SLURM 17
- MVAPICH2-X User Guide: A detailed user guide with instructions to install MVAPICH2-X and execute MPI/UPC/UPC++/OpenSHMEM/CAF/Hybrid programs is available. (HTML, PDF)
- **Installation Guide**
  - These tarballs contain the MVAPICH2-X software for Redhat and Debian based systems combined together in one combined package.
  - Running the install.sh script in the tarball will install the libraries.
  - These RPMs are relocatable and advanced users may skip the install.sh script to directly use alternate commands to install the desired RPMs.
- **Which RPM should I install?**
  - InfiniBand / RoCE System
  - Omni-Path System
- **Advanced Install Options**
  - Install library using a prefix other than the default of /opt/mvapich2/.

```
$ rpm --prefix /custom/install/prefix -Uvh --nodeps mvapich2-x-basic-mofed3.4-gnu4.8.5-2.3rc2-1.el7.centos.x86_64.rpm
```

  - If you do not have root permission or are on a system that does not use RPMs you can use rpm2cpio to extract the library.

```
$ rpm2cpio mvapich2-x-basic-mofed3.4-gnu4.8.5-2.3rc2-1.el7.centos.x86_64.rpm | cpio -id
```

  When using the rpm2cpio method, you will need to update the MPI compiler scripts, such as mpicc, in order to point to the correct path of where you place the library.
  - Tip: If you are using a Debian based system such as Ubuntu you can convert the rpm to a deb using a tool such as alien or follow the rpm2cpio instructions above.

Combined Tarballs

	x86-64	OpenPOWER	ARM
Stock OFED	[GNU 4.8.5]	Coming Soon!	Coming Soon!
MLNX-OFED 3.X*	[GNU 4.8.5]	Coming Soon!	Coming Soon!
MLNX-OFED 4.X*	[GNU 4.8.5]	Coming Soon!	Coming Soon!
Intel IFS 10.6	[GNU 4.8.5]	N/A	N/A
Intel IFS 10.9	[GNU 4.8.5]	N/A	N/A

SPACK : Solving complexities in Software installation

- A flexible package manager that supports multiple versions, configurations, platforms, and compilers.
- Automates all package-related processes :
  - Source Download
  - Checksum Verification
  - Configuration
  - Build
  - Installation
  - Testing (Very Basic)

How to install Spack:
$ git clone https://github.com/spack/spack
$. spack/share/spack/setup-env.sh

How to install a package: MVAPICH2
$ spack install mvapich2 fabrics=mrail

Running applications through SPACK : AMG
$ spack install amg %gcc ^mvapich2@2.3.1 fabrics=mrail
$ spack cd -i amg # to go in AMG install folder
$ spack find –p mvapich2@2.3.1 # Set your MPI_HOME
$ $MPI_HOME/bin/mpirun_rsh -np 56 -hostfile hosts ./bin/amg

SPACK : Solving complexities in Software installation

How to install Spack:
$ git clone https://github.com/spack/spack
$. spack/share/spack/setup-env.sh

How to install a package: MVAPICH2
$ spack install mvapich2

Test with MVAPICH2 OMB

Concluding Remarks

- Upcoming Exascale systems need to be designed with a holistic view of HPC, Deep Learning, and Cloud
- Presented an overview of designing convergent software stacks
- Presented solutions enable HPC and Deep Learning communities to take advantage of current and next-generation systems
- Presented solutions to deploy these solutions on traditional HPC and Cloud systems

Commercial Support for MVAPICH2, HiBD, and HiDL Libraries

- Supported through X-ScaleSolutions (http://x-scalesolutions.com)
- Benefits:
  - Help and guidance with installation of the library
  - Platform-specific optimizations and tuning
  - Timely support for operational issues encountered with the library
  - Web portal interface to submit issues and tracking their progress
  - Advanced debugging techniques
  - Application-specific optimizations and tuning
  - Obtaining guidelines on best practices
  - Periodic information on major fixes and updates
  - Information on major releases
  - Help with upgrading to the latest release
  - Flexible Service Level Agreements
- Support provided to Lawrence Livermore National Laboratory (LLNL) for the last two years

X-ScaleSolutions

Silver ISV Member for the OpenPOWER Consortium + Products

- Has joined the OpenPOWER Consortium as a silver ISV member
- Provides flexibility:
  - To have MVAPICH2, HiDL and HiBD libraries getting integrated into the OpenPOWER software stack
  - A part of the OpenPOWER ecosystem
  - Can participate with different vendors for bidding, installation and deployment process
- Introduced two new integrated products with support for OpenPOWER systems (Presented at the OpenPOWER North America Summit)
  - X-ScaleHPC
  - X-ScaleAI
  - Send an e-mail to contactus@x-scalesolutions.com for free trial!!

X-ScaleSolutions

Funding Acknowledgments

Funding Support by

Equipment Support by

Personnel Acknowledgments

Current Students (Graduate)

- A. Awan (Ph.D.)
- M. Bayatpour (Ph.D.)
- C.-H. Chu (Ph.D.)
- J. Hashmi (Ph.D.)
- A. Jain (Ph.D.)
- K. S. Kandadi (M.S.)
- Kamal Raj (M.S.)
- K. S. Khorassani (Ph.D.)
- P. Kousha (Ph.D.)
- A. Quentin (Ph.D.)
- B. Ramesh (M. S.)
- S. Xu (M.S.)
- Q. Zhou (Ph.D.)

Current Research Scientist

- H. Subramoni

Current Post-doc

- M. S. Ghazimeersaeed
- A. Ruhela
- K. Manian

Current Students (Undergraduate)

- V. Gangal (B.S.)
- N. Sarkauskas (B.S.)

Current Research Specialist

- J. Smith

Past Students

- A. Augustine (M.S.)
- P. Balaji (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.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- 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.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- A. Singh (Ph.D.)
- J. Sridhar (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 Scientist

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

Past Programmers

- D. Bureddy
- J. Perkins

Past Research Specialist

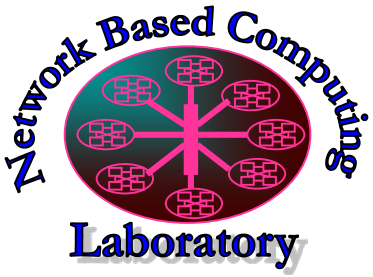
- M. Arnold

Past Post-Docs

- D. Banerjee
- X. Besseron
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- S. Marcarelli
- J. Vienne
- H. Wang

Thank You!

panda@cse.ohio-state.edu



Follow us on

https://twitter.com/mvapich

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 Big Data Project

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

The High-Performance Deep Learning Project

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