MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

Overview of the MVAPICH Project: Latest Status and Future Roadmap

MVAPICH2 User Group (MUG) Meeting

by

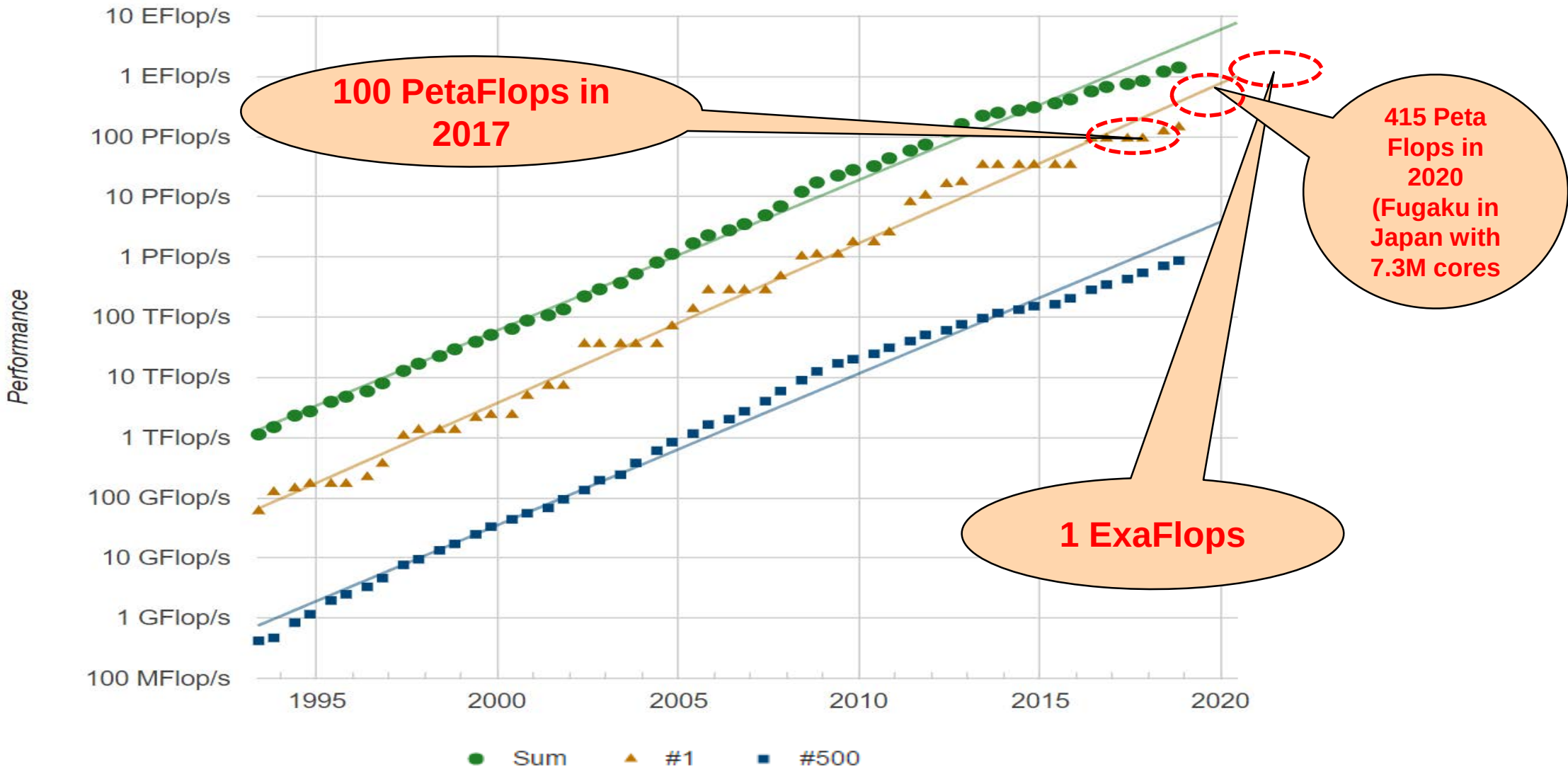
Dhabaleswar K. (DK) Panda

The Ohio State University

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

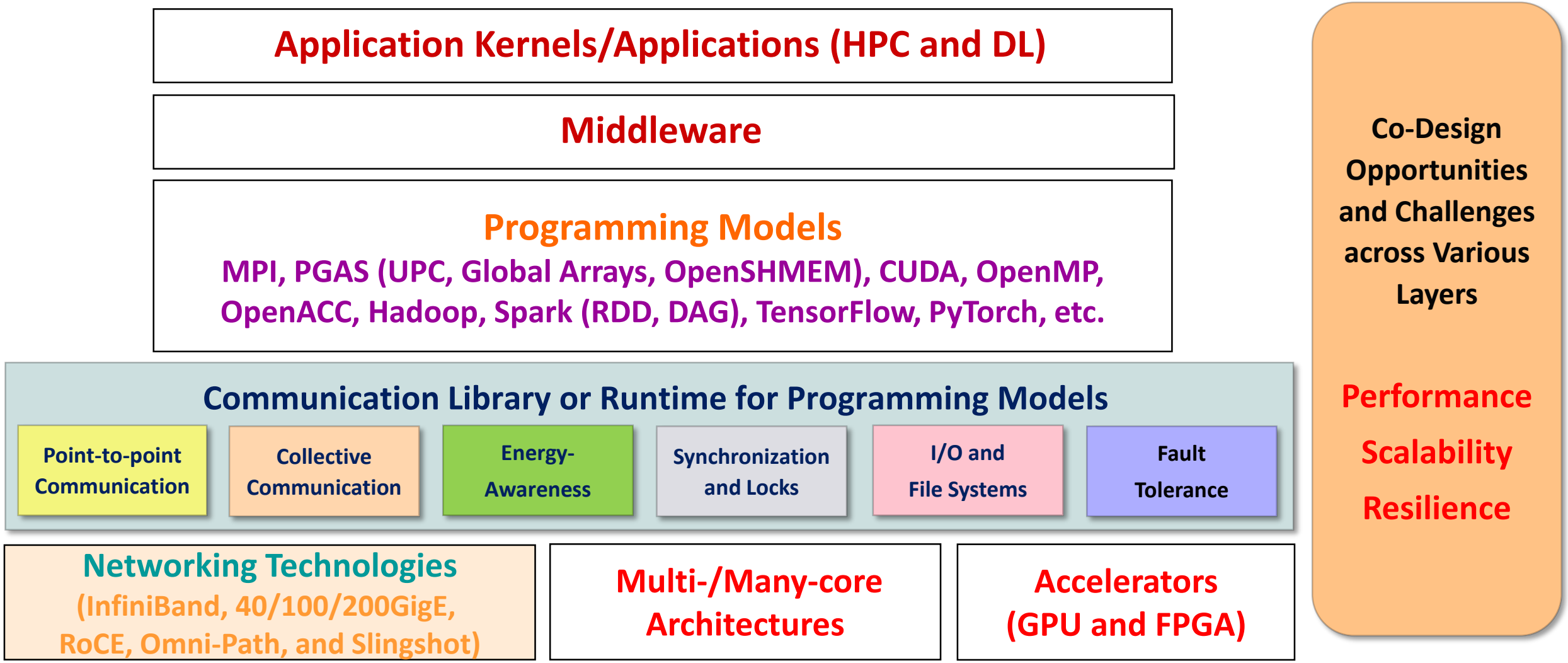
http://www.cse.ohio-state.edu/~panda

High-End Computing (HEC): PetaFlop to ExaFlop



Expected to have an ExaFlop system in 2021!

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



Designing (MPI+X) at Exascale

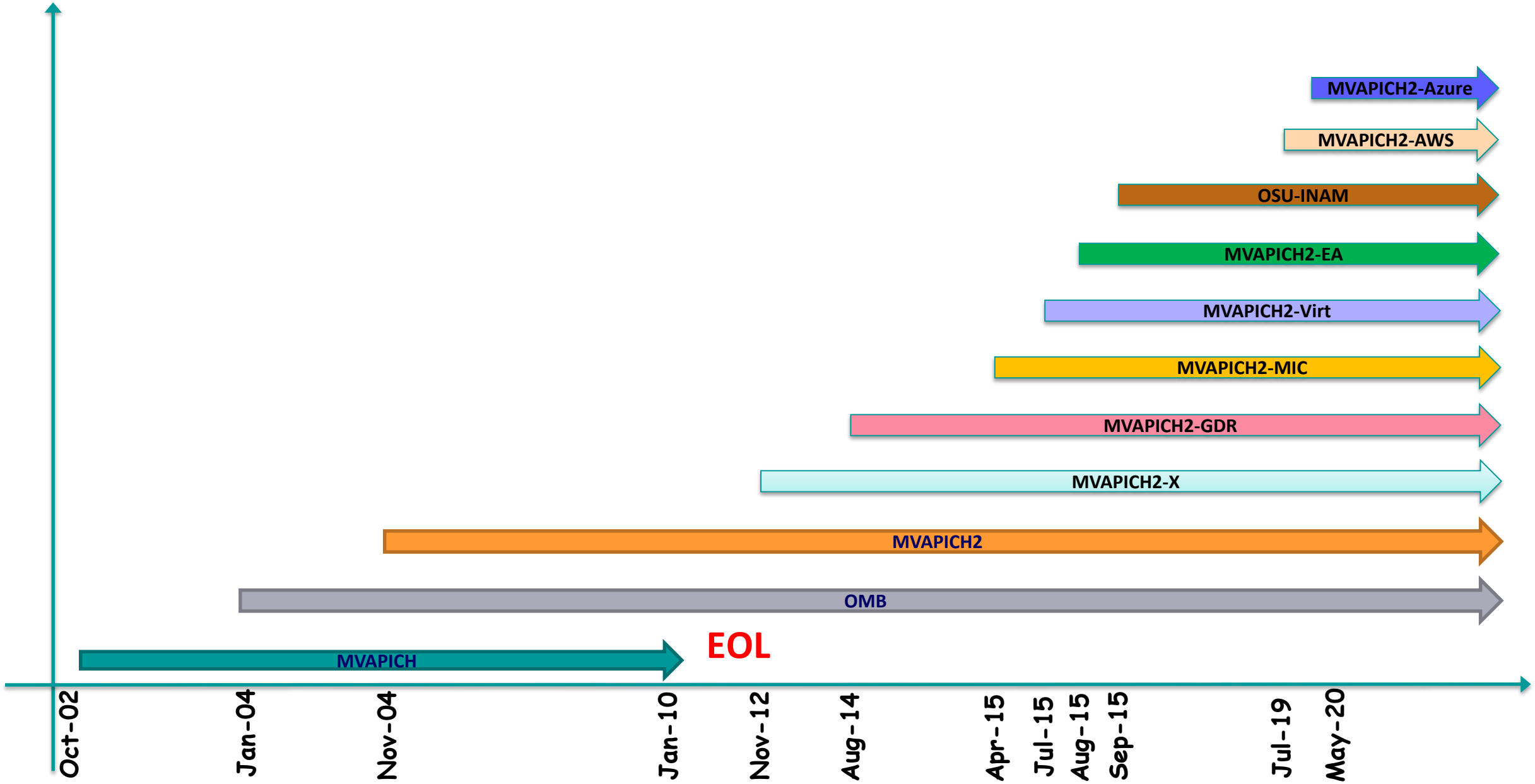
- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Scalable job start-up
  - Low memory footprint
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Balancing intra-node 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

Overview of the MVAPICH2 Project

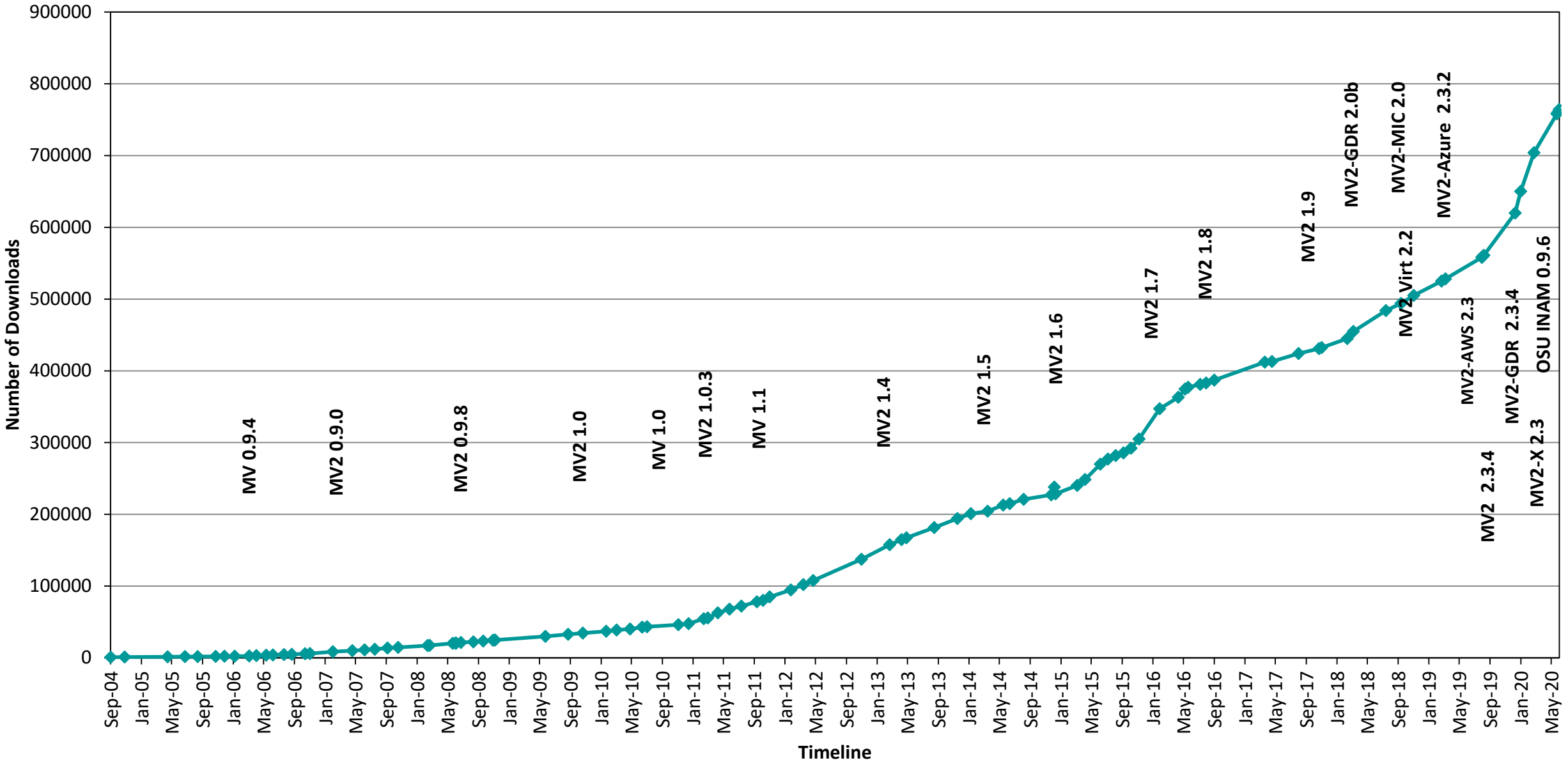
- High Performance open-source MPI Library
- Support for multiple interconnects
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), and AWS EFA
- Support for multiple platforms
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and AMD (upcoming))
- Started in 2001, first open-source 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 3,100 organizations in 89 countries**
- **More than 821,000 (> 0.8 million) downloads from the OSU site directly**
- Empowering many TOP500 clusters (June '20 ranking)
  - 4th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 8th, 448, 448 cores (Frontera) at TACC
  - 12th, 391,680 cores (ABCI) in Japan
  - 18th, 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 8th ranked TACC Frontera system
- **Empowering Top500 systems for more than 15 years**

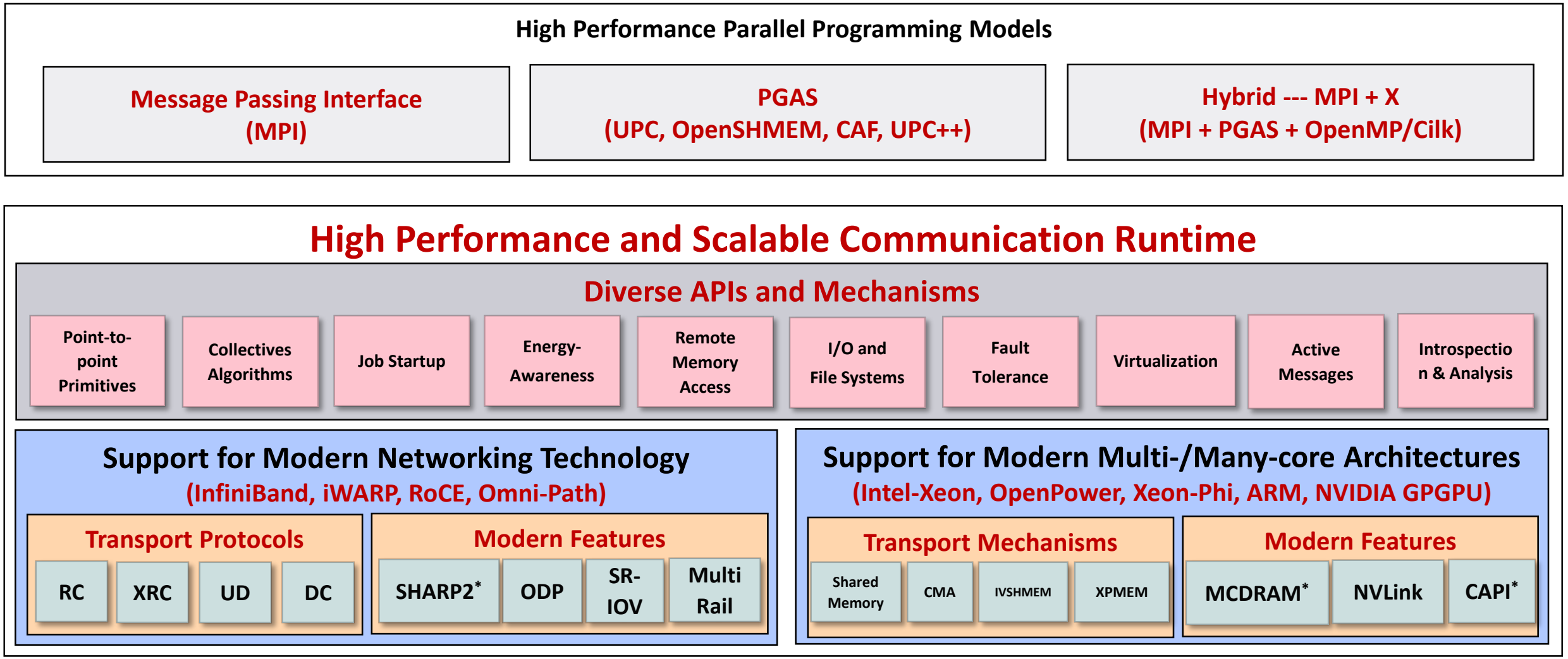
MVAPICH Project Timeline



MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family (for HPC and DL)



* Upcoming

Production Quality Software Design, Development and Release

- 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, Scalapak, and SPEC
  - Spend about 18,000 core hours per commit
  - Performance regression and tuning
  - Applications-based evaluation
  - Evaluation on large-scale systems
- All versions (alpha, beta, RC1 and RC2) go through the above testing

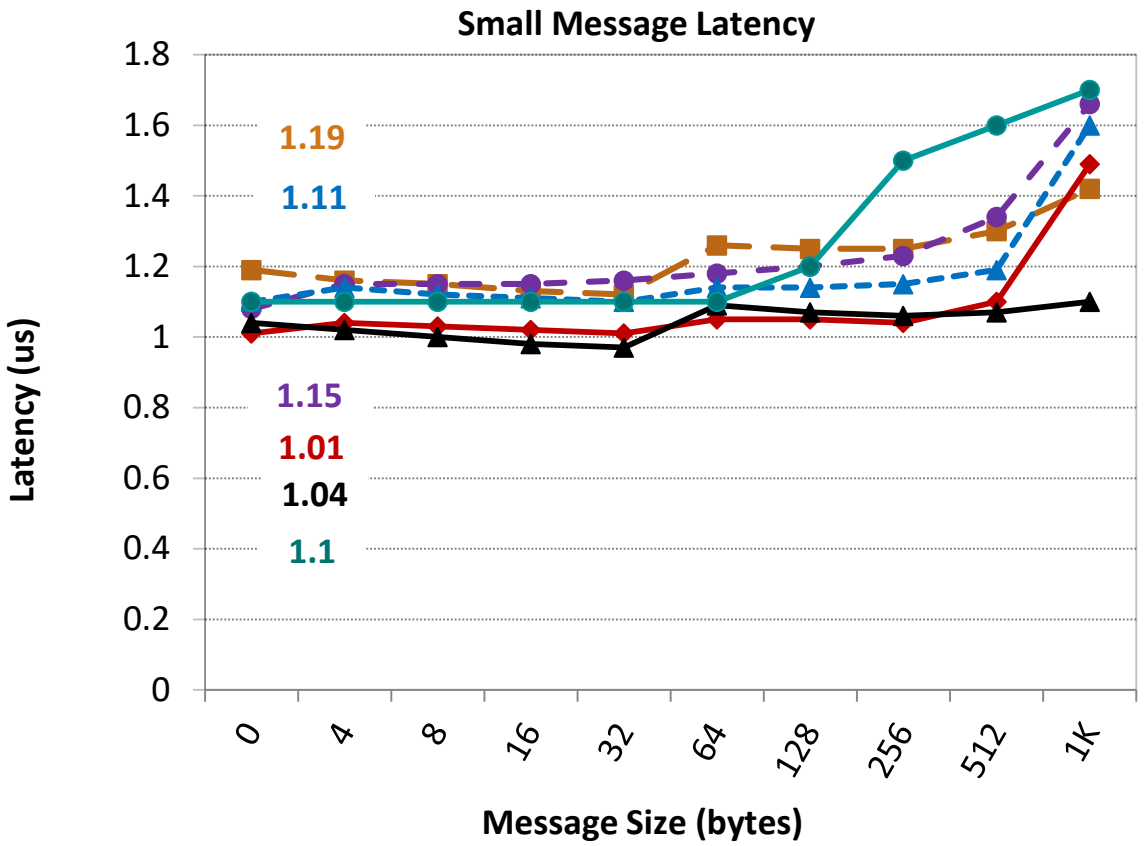
MVAPICH2 Software Family

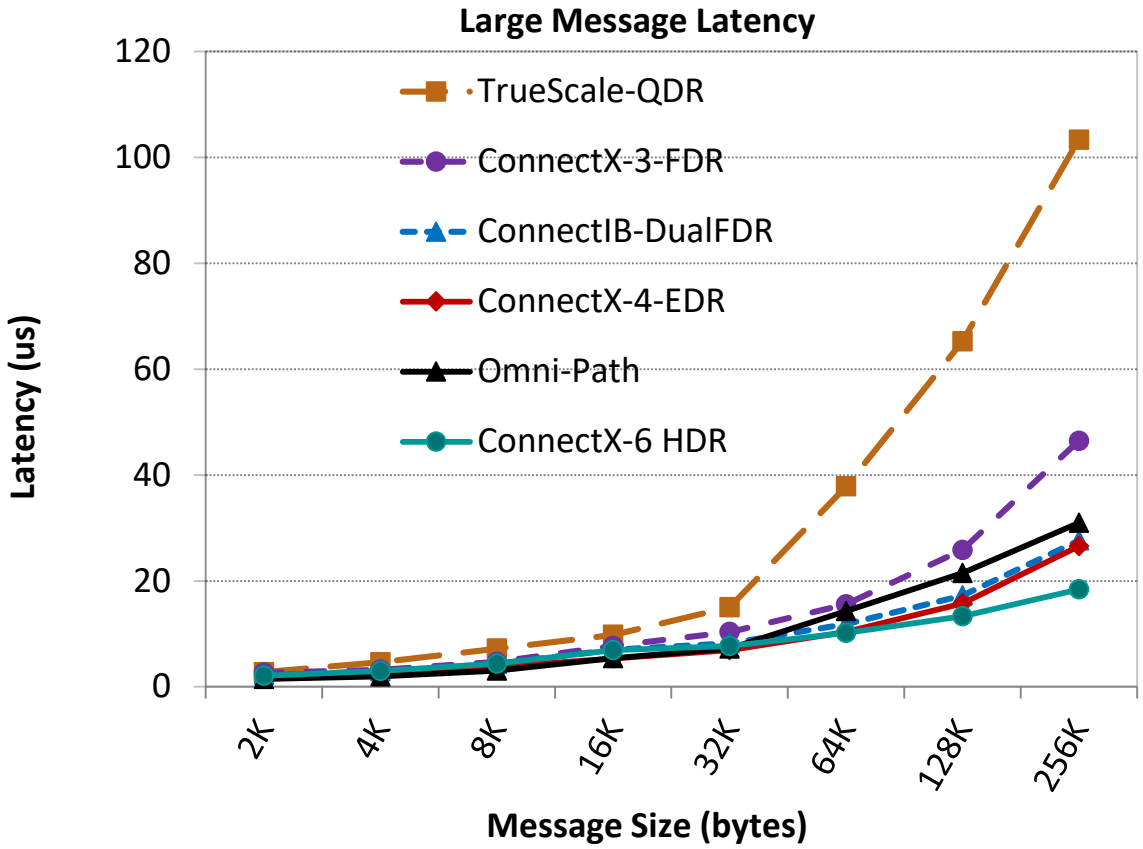
Requirements	Library
MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2
Advanced MPI features/support (UMR, ODP, DC, Core-Direct, SHARP, XPMEM), OSU INAM (InfiniBand Network Monitoring and Analysis), MPI+PGAS	MVAPICH2-X
Optimized Support of MVAPICH2-X for Microsoft Azure Platform with InfiniBand	MVAPICH2-X-Azure
Advanced MPI features (SRD and XPMEM) with support for Amazon Elastic Fabric Adapter (EFA)	MVAPICH2-X-AWS
Optimized MPI for clusters with NVIDIA GPUs and for GPU-enabled Deep Learning Applications	MVAPICH2-GDR
Energy-aware MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

Highlights of MVAPICH2 2.3.4-GA Release

- Support for highly-efficient inter-node and intra-node communication
- Scalable Start-up
- Performance Engineering with MPI-T
- Application Scalability

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

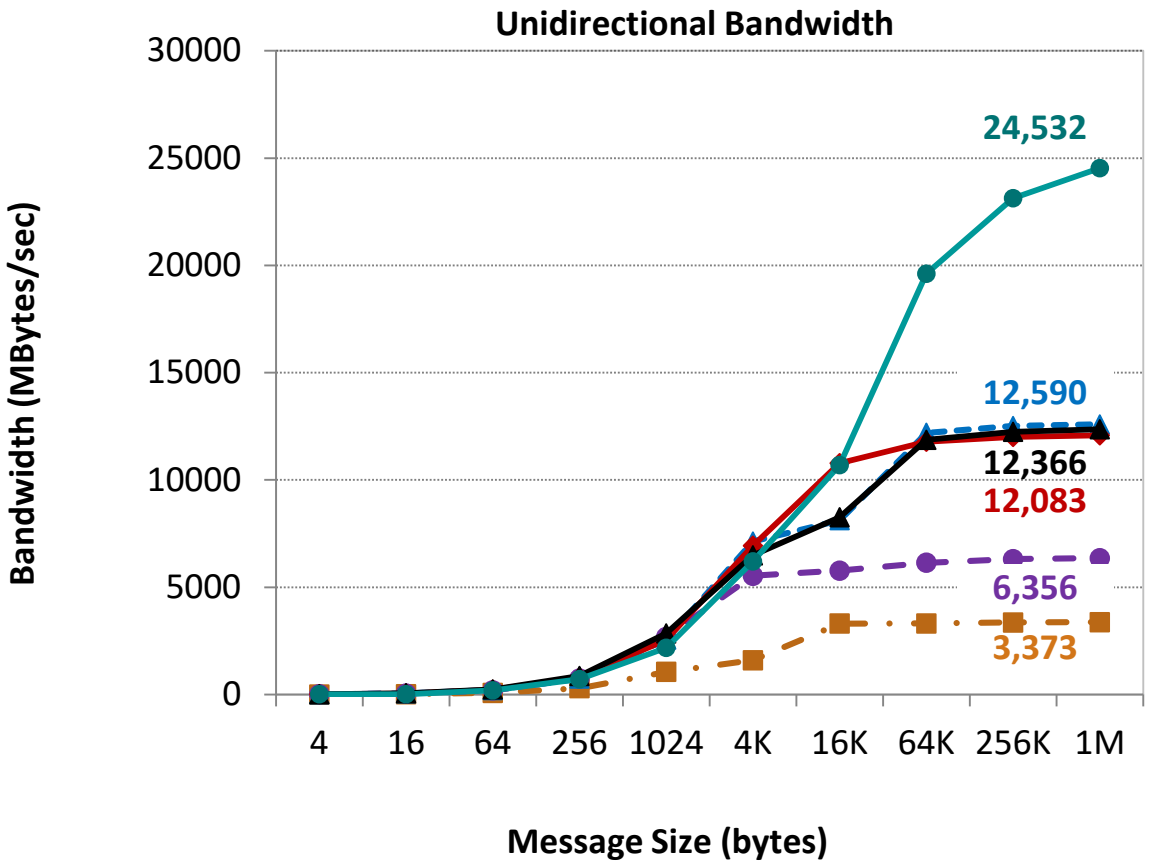
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

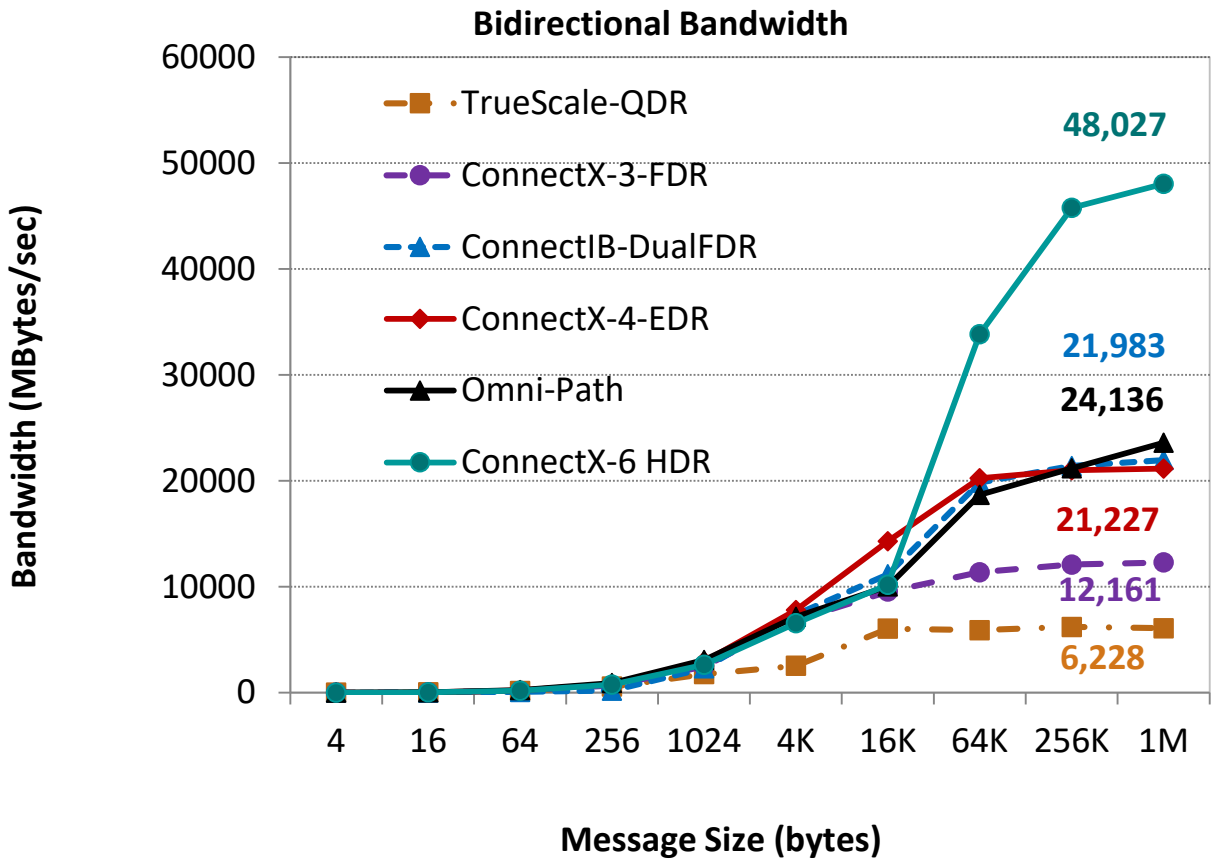
ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

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

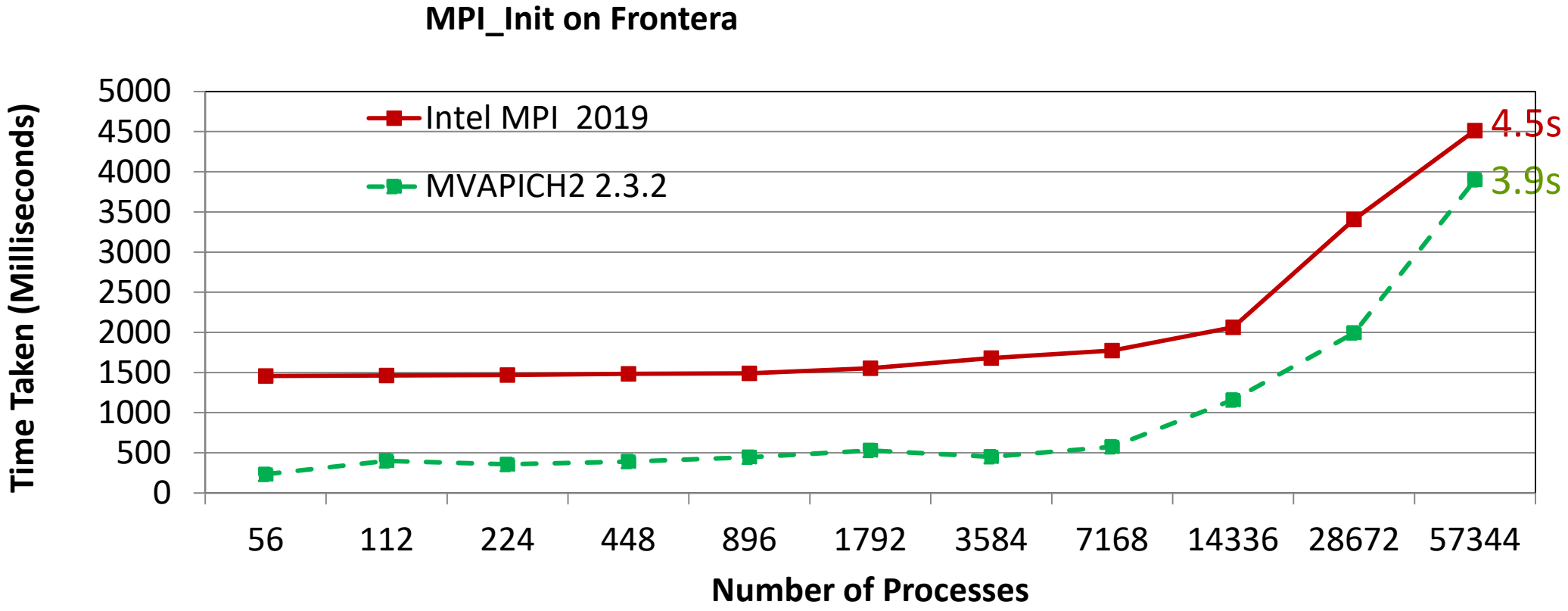
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

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

Startup Performance on TACC Frontera

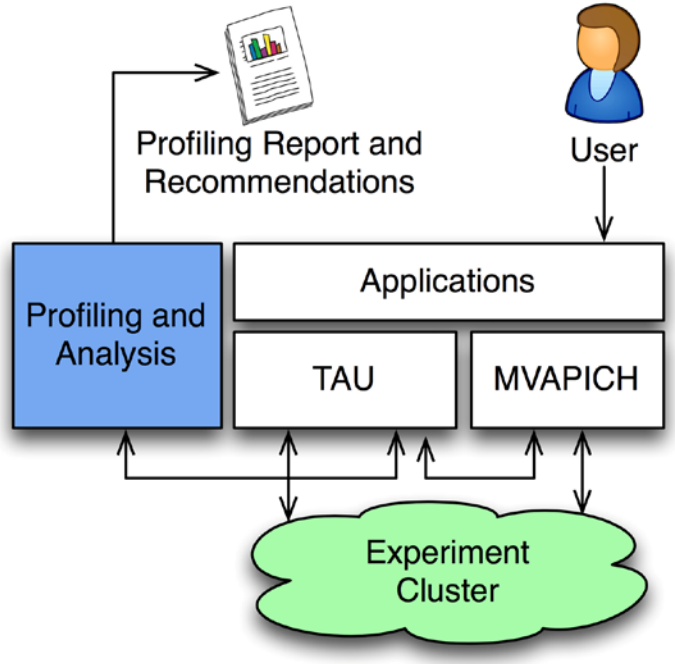


- MPI_Init takes 3.9 seconds on 57,344 processes on 1,024 nodes
- All numbers reported with 56 processes per node

New designs available since MVAPICH2-2.3.2

Performance Engineering Applications using MVAPICH2 and TAU

- Enhance existing support for MPI_T in MVAPICH2 to expose a richer set of performance and control variables
- Get and display MPI Performance Variables (PVARs) made available by the runtime in TAU
- Control the runtime's behavior via MPI Control Variables (CVARs)
- Introduced support for new MPI_T based CVARs to MVAPICH2
  - MPIR_CVAR_MAX_INLINE_MSG_SZ, MPIR_CVAR_VBUF_POOL_SIZE, MPIR_CVAR_VBUF_SECONDARY_POOL_SIZE
- TAU enhanced with support for setting MPI_T CVARs in a non-interactive mode for uninstrumented applications
- S. Ramesh, A. Maheo, S. Shende, A. Malony, H. Subramoni, and D. K. Panda, *MPI Performance Engineering with the MPI Tool Interface: the Integration of MVAPICH and TAU, EuroMPI/USA '17, Best Paper Finalist*
- More details in Sameer Shende's talk (later today)



VBUF usage without CVAR based tuning as displayed by ParaProf

TAU: ParaProf: Context Events for: node 0 - mpit_withoutcvar_bt.C.1k.ppk

Name	MaxValue	MinValue	MeanValue	Std. Dev.	NumSamples	Total
mv2_total_vbuf_memory (Total amount of memory in bytes used for VBUFs)	3,313,056	3,313,056	3,313,056	0	1	3,313,056
mv2_ud_vbuf_allocated (Number of UD VBUFs allocated)	0	0	0	0	0	0
mv2_ud_vbuf_available (Number of UD VBUFs available)	0	0	0	0	0	0
mv2_ud_vbuf_freed (Number of UD VBUFs freed)	0	0	0	0	0	0
mv2_ud_vbuf_inuse (Number of UD VBUFs inuse)	0	0	0	0	0	0
mv2_ud_vbuf_max_use (Maximum number of UD VBUFs used)	0	0	0	0	0	0
mv2_vbuf_allocated (Number of VBUFs allocated)	320	320	320	0	1	320
mv2_vbuf_available (Number of VBUFs available)	255	255	255	0	1	255
mv2_vbuf_freed (Number of VBUFs freed)	25,545	25,545	25,545	0	1	25,545
mv2_vbuf_inuse (Number of VBUFs inuse)	65	65	65	0	1	65
mv2_vbuf_max_use (Maximum number of VBUFs used)	65	65	65	0	1	65
num_calloc_calls (Number of MPIT_calloc calls)	89	89	89	0	1	89

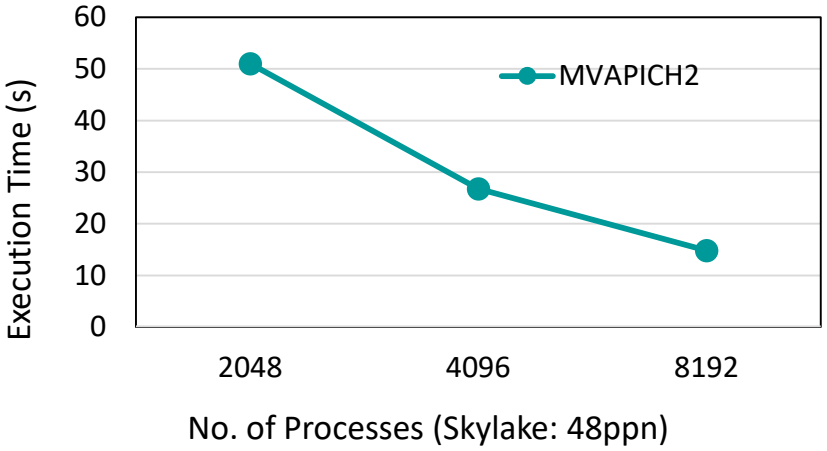
VBUF usage with CVAR based tuning as displayed by ParaProf

TAU: ParaProf: Context Events for: node 0 - bt-mz.E.vbuf_pool_16.1k.ppk

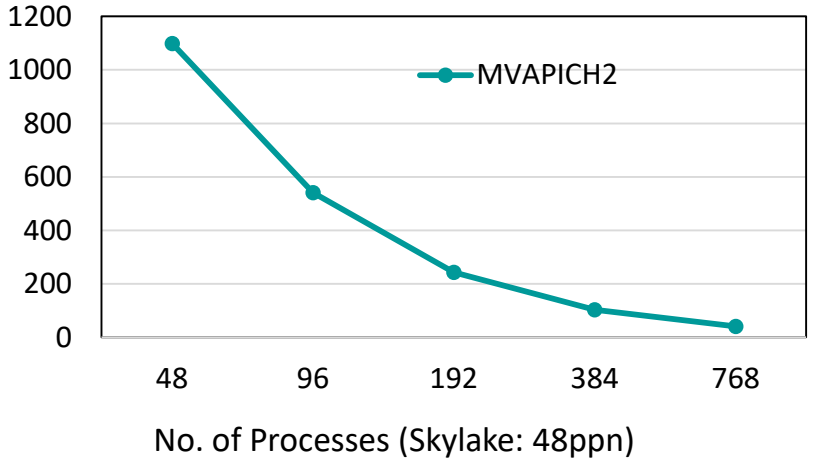
Name	MaxValue	MinValue	MeanValue	Std. Dev.	NumSamp...	Total
mv2_total_vbuf_memory (Total amount of memory in bytes used for VBUFs)	1,815,056	1,815,056	1,815,056	0	1	1,815,056
mv2_ud_vbuf_allocated (Number of UD VBUFs allocated)	0	0	0	0	0	0
mv2_ud_vbuf_available (Number of UD VBUFs available)	0	0	0	0	0	0
mv2_ud_vbuf_freed (Number of UD VBUFs freed)	0	0	0	0	0	0
mv2_ud_vbuf_inuse (Number of UD VBUFs inuse)	0	0	0	0	0	0
mv2_ud_vbuf_max_use (Maximum number of UD VBUFs used)	0	0	0	0	0	0
mv2_vbuf_allocated (Number of VBUFs allocated)	160	160	160	0	1	160
mv2_vbuf_available (Number of VBUFs available)	94	94	94	0	1	94
mv2_vbuf_freed (Number of VBUFs freed)	5,479	5,479	5,479	0	1	5,479
mv2_vbuf_inuse (Number of VBUFs inuse)	66	66	66	0	1	66

Application Scalability on Stampede2

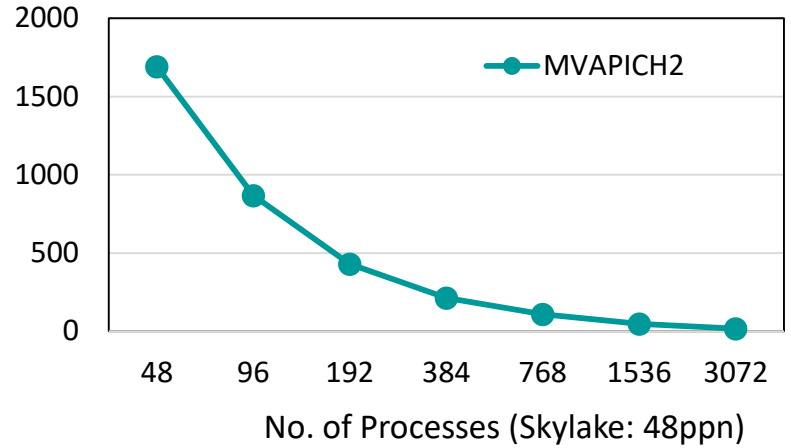
MiniFE (1300x1300x1300 ~ 910 GB)

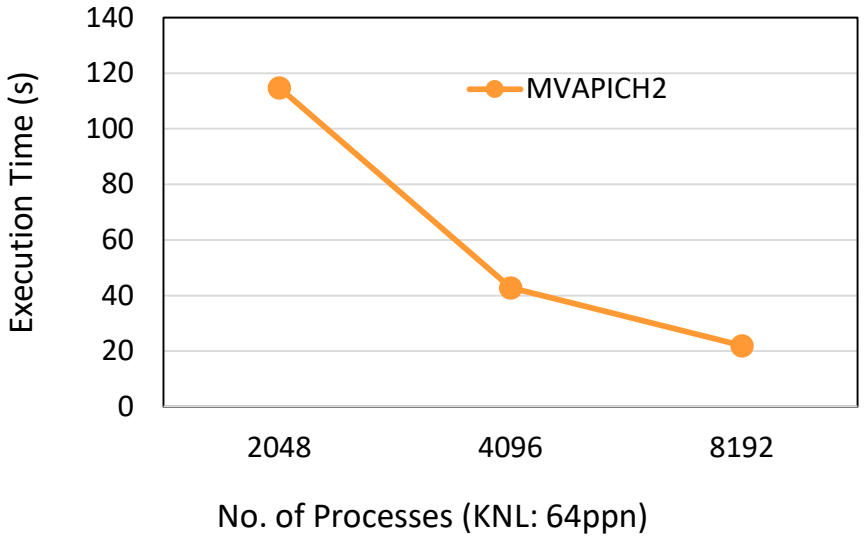


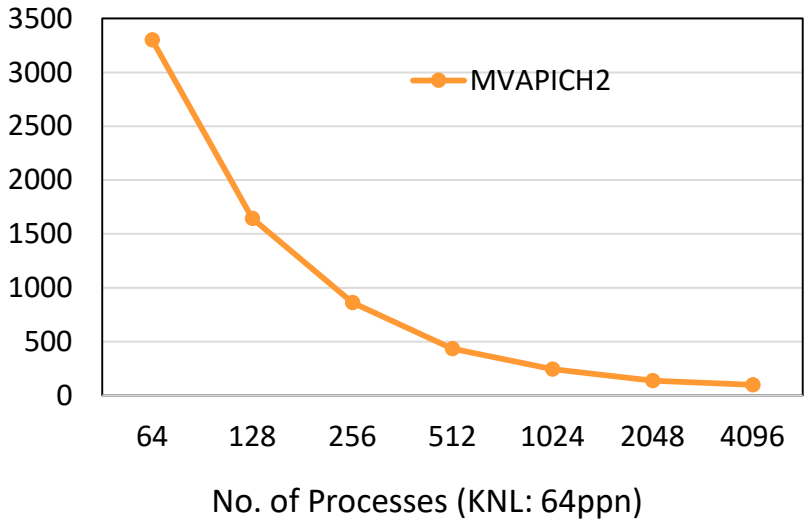
NEURON (YuEtAl2012)

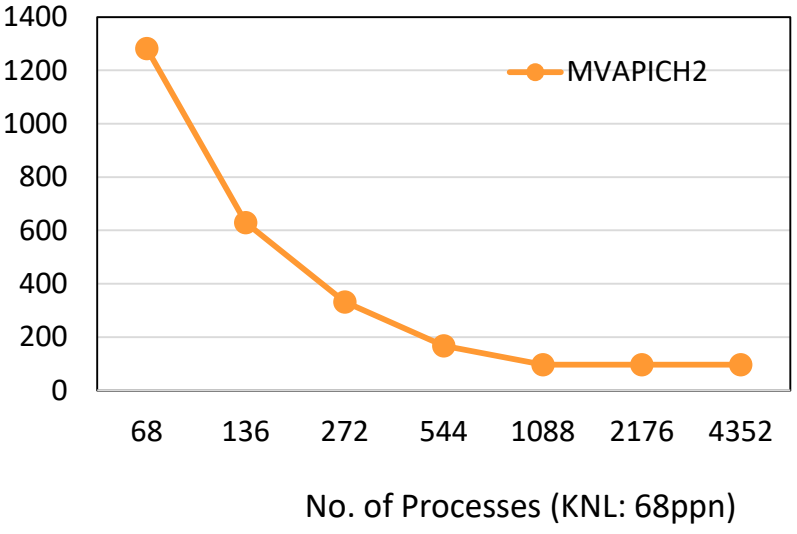


Cloverleaf (bm64) MPI+OpenMP, NUM_OMP_THREADS = 2









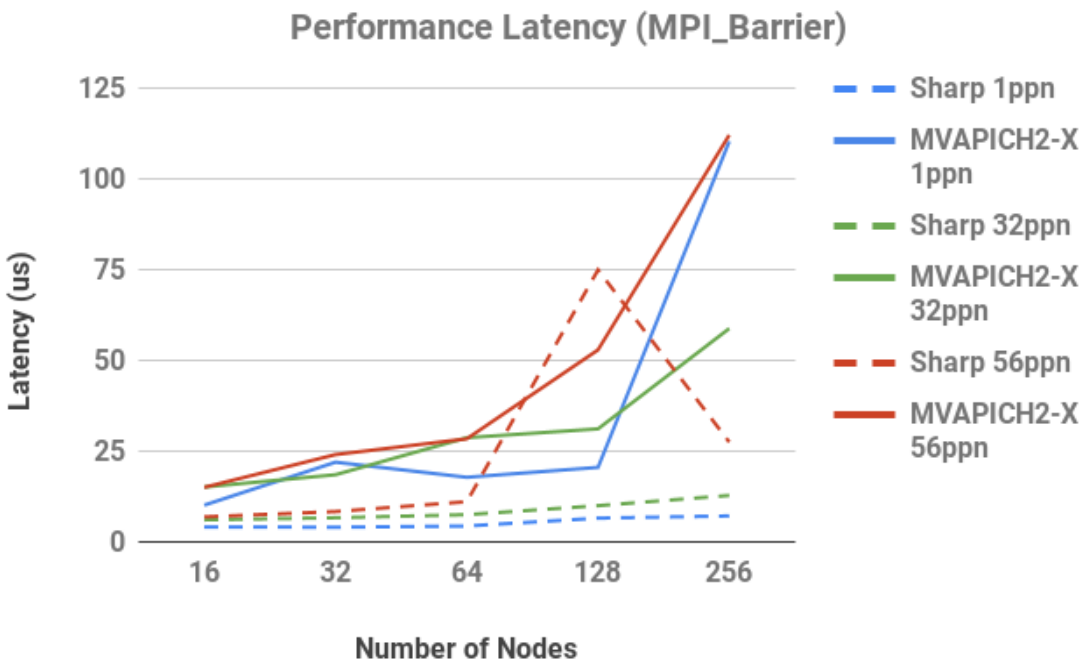
Courtesy: Mahidhar Tatineni @SDSC, Dong Ju (DJ) Choi@SDSC, and Samuel Khuvis@OSC ---- Testbed: TACC Stampede2 using MVAPICH2-2.3b

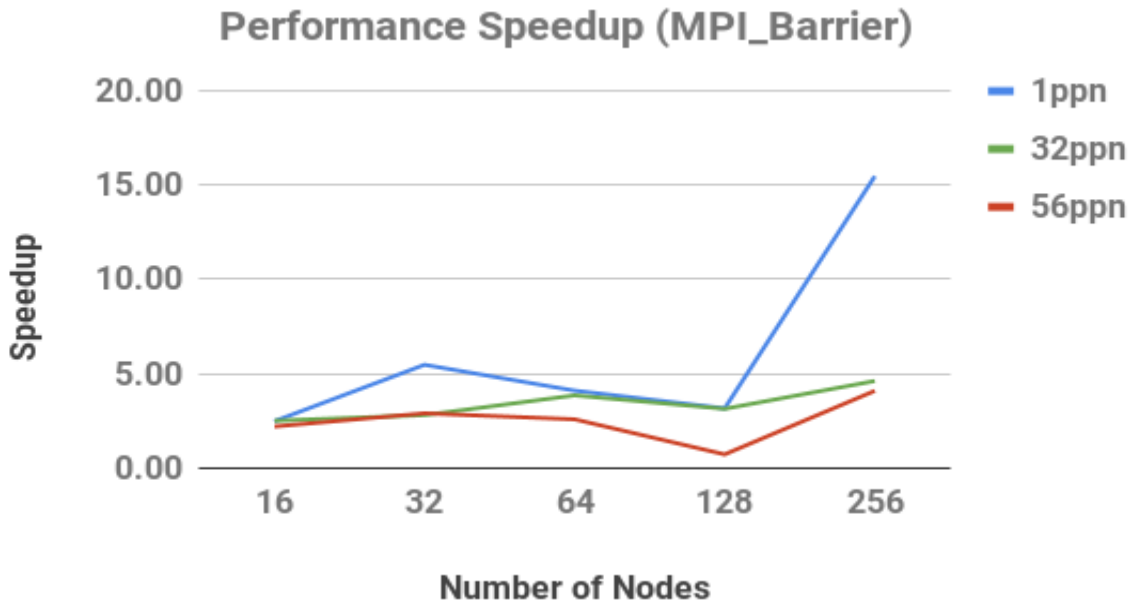
Runtime parameters: *MV2_SMPI_LENGTH_QUEUE=524288 PSM2_MQ_RNDV_SHM_THRESH=128K PSM2_MQ_RNDV_HFI_THRESH=128K*

MVAPICH2 Upcoming Features

- Integration of SHARPv2 and associated Collective Optimizations
- Support for Broadcom RoCEv2 adapter

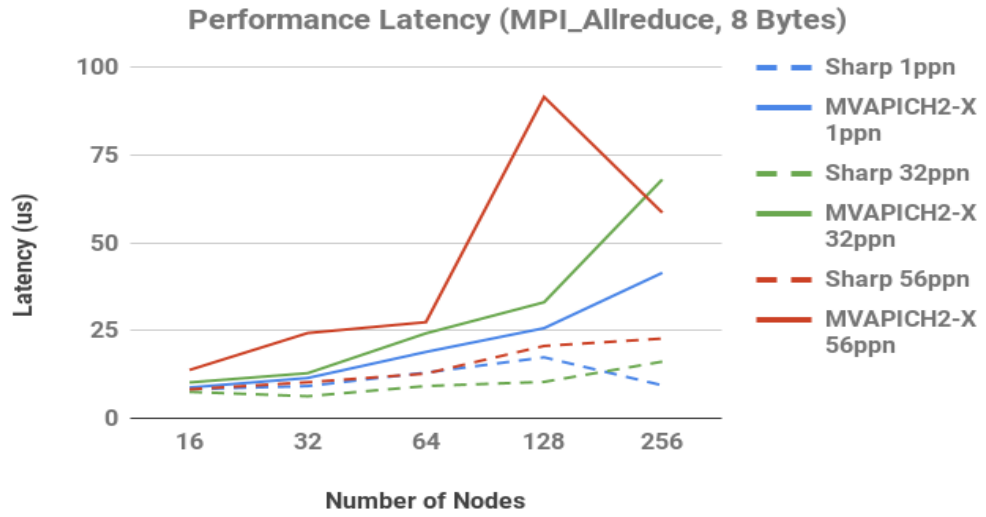
Impact of SHARPv2 on Performance of MPI_Barrier

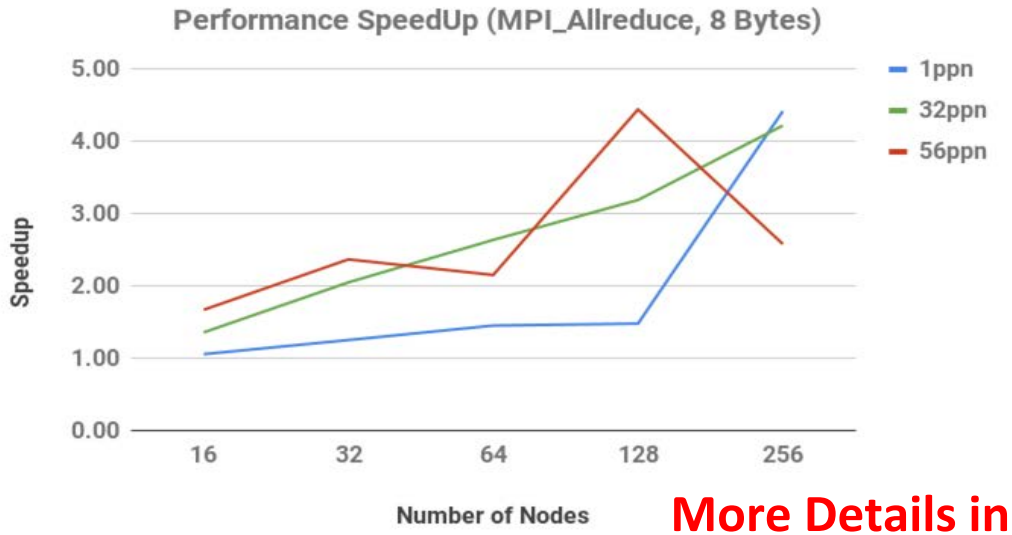




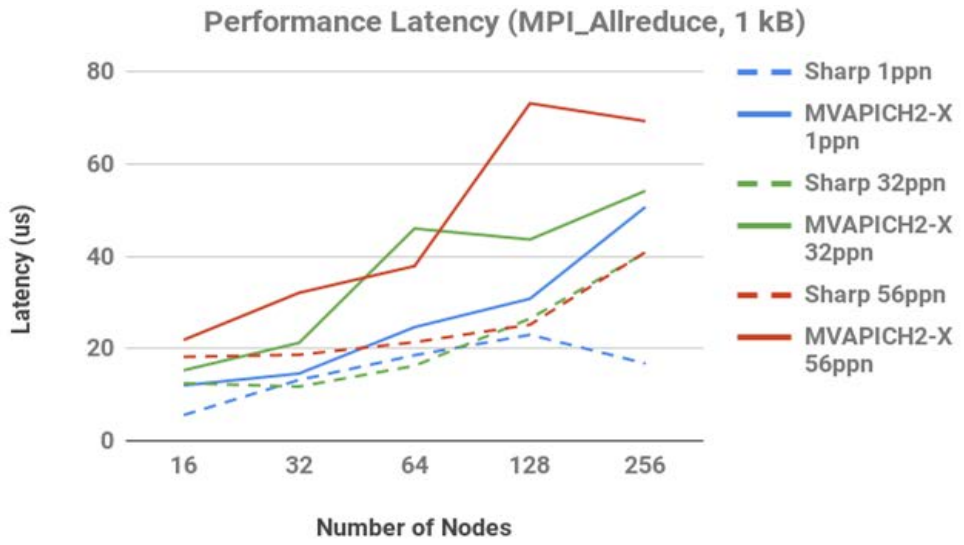
Impact of SHARP on performance of MVAPICH2-X using osu_barrier on different scales. SHARP-enabled MVAPICH2-X has sustained performance benefit as the number of the nodes increases.

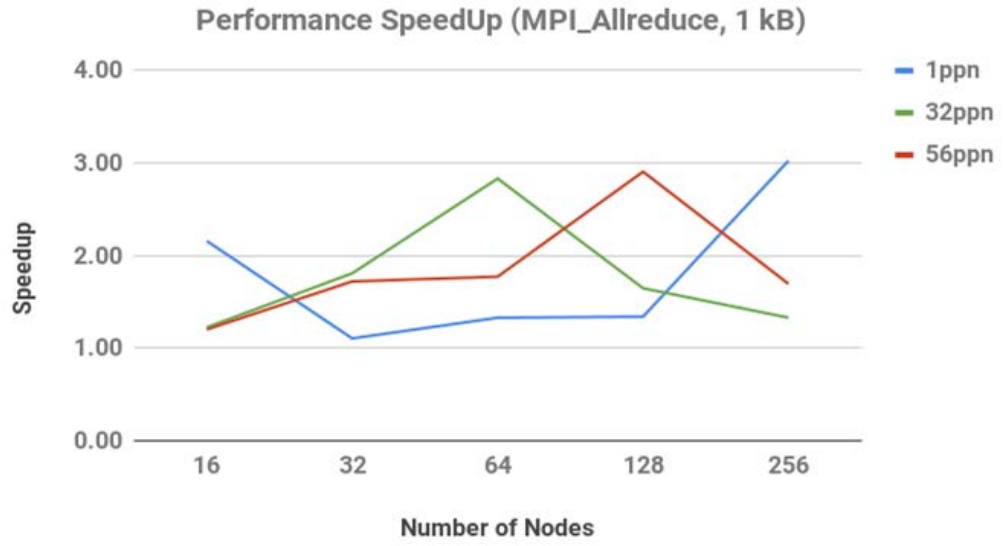
Impact of SHARPv2 on Performance of MPI_Allreduce





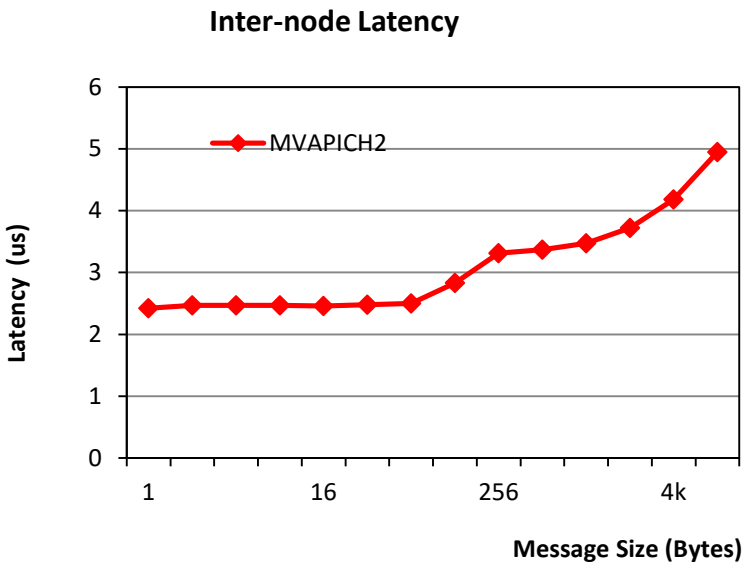
More Details in TACC Talk (Tomorrow)

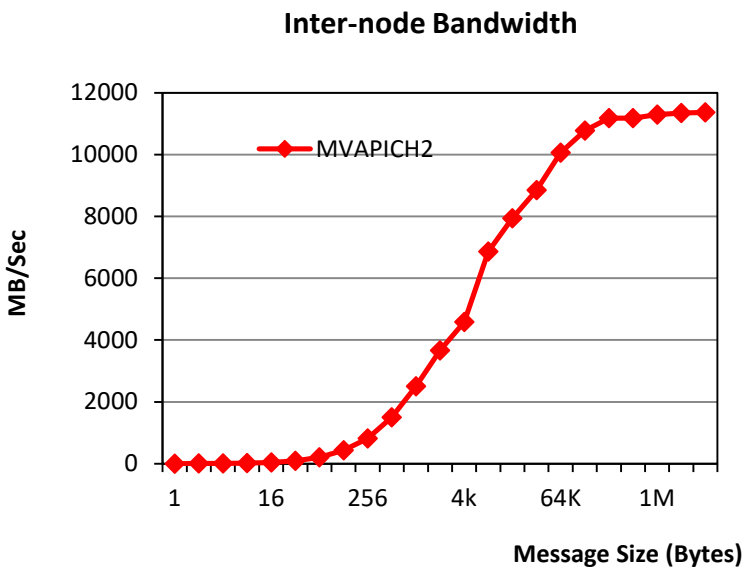


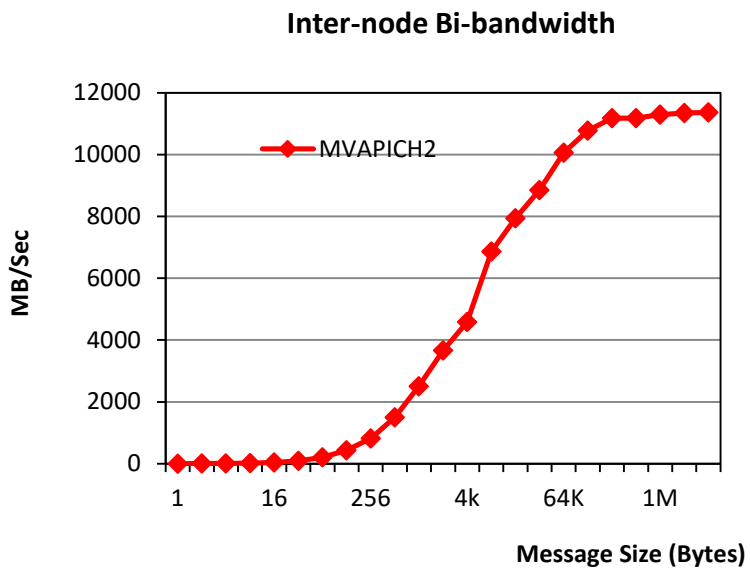


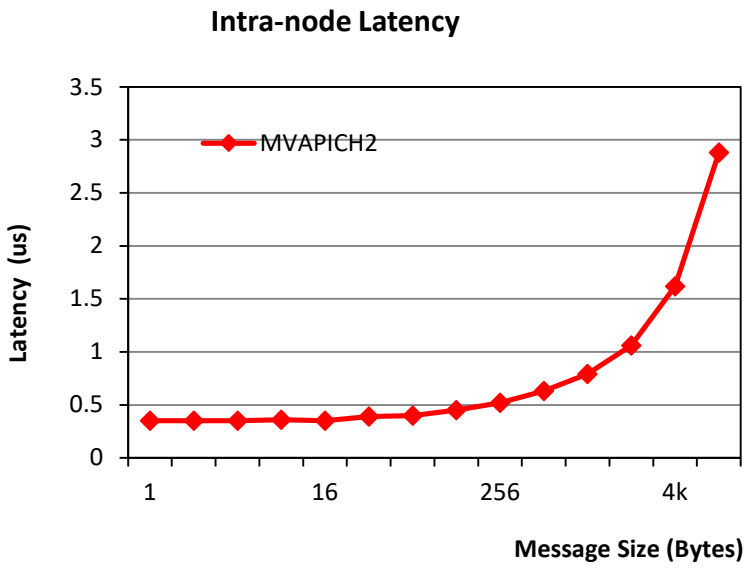
SHARP-enabled MVAPICH2-x has sustained performance benefit as the number of the nodes increases.

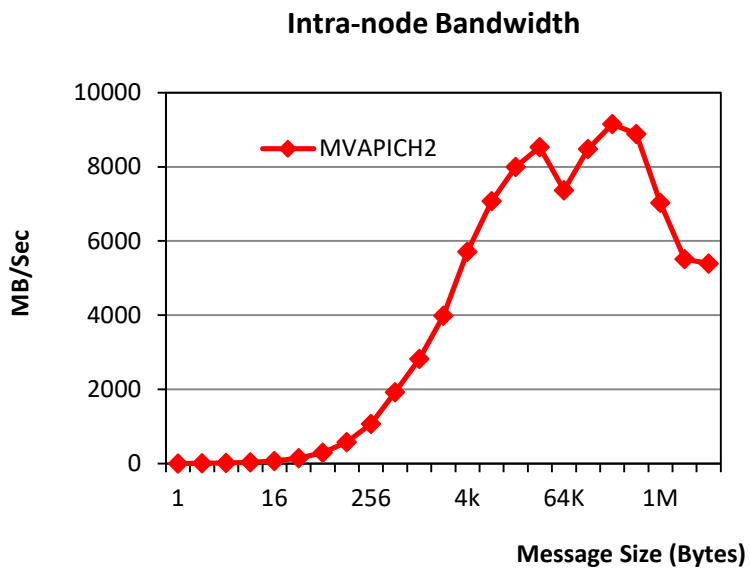
Point-to-Point Performance on Broadcom RoCE V2 Adapter

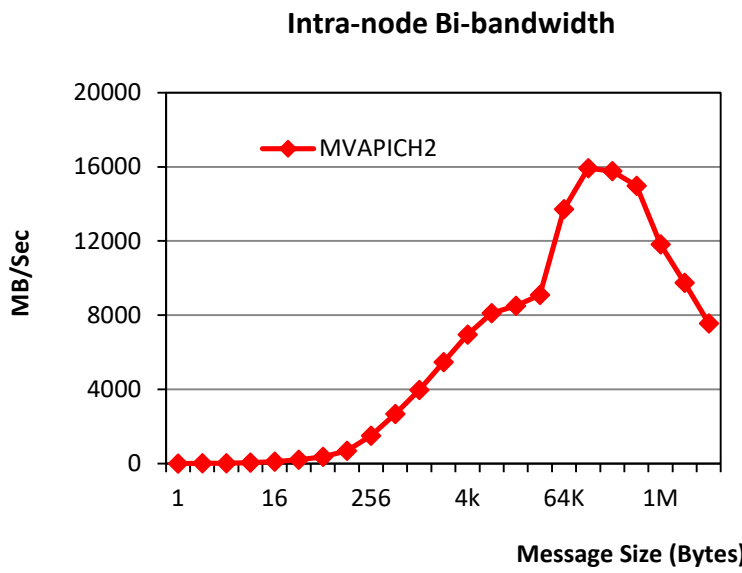












More Results in Broadcom Talk (Later Today)

MVAPICH2 Software Family

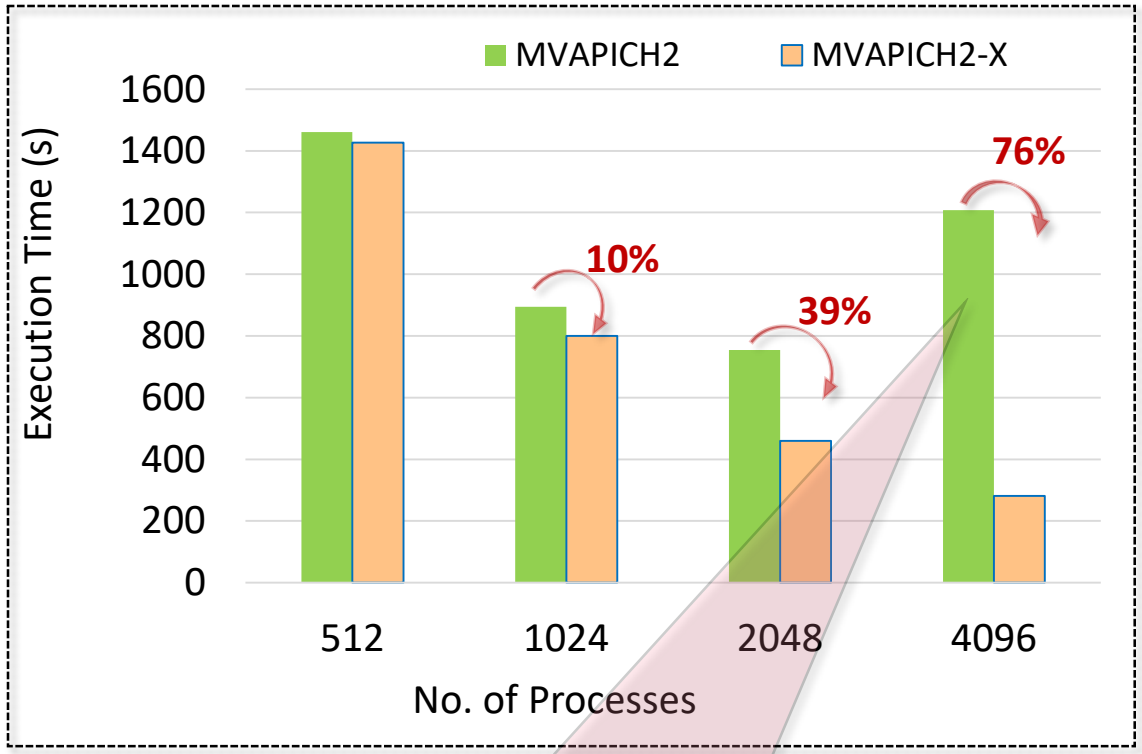
Requirements	Library
MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2
Advanced MPI features/support (UMR, ODP, DC, Core-Direct, SHARP, XPMEM), OSU INAM (InfiniBand Network Monitoring and Analysis), MPI+PGAS	MVAPICH2-X
Optimized Support of MVAPICH2-X for Microsoft Azure Platform with InfiniBand	MVAPICH2-X-Azure
Advanced MPI features (SRD and XPMEM) with support for Amazon Elastic Fabric Adapter (EFA)	MVAPICH2-X-AWS
Optimized MPI for clusters with NVIDIA GPUs and for GPU-enabled Deep Learning Applications	MVAPICH2-GDR
Energy-aware MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

Overview of Some of the MVAPICH2-X Features

- Direct Connect (DC) Transport
- Co-operative Rendezvous Protocol
- CMA-based Collectives
- Asynchronous Progress
- XPMEM-based Reduction Collectives

Impact of DC Transport Protocol on Neuron

Neuron with YuEtAl2012



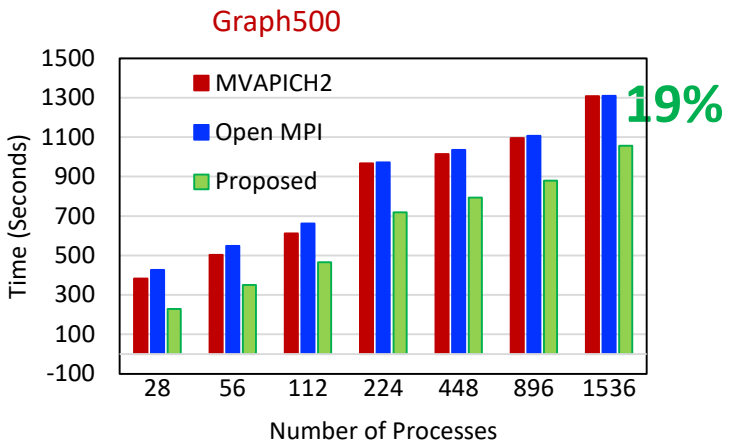
Overhead of RC protocol for connection establishment and communication

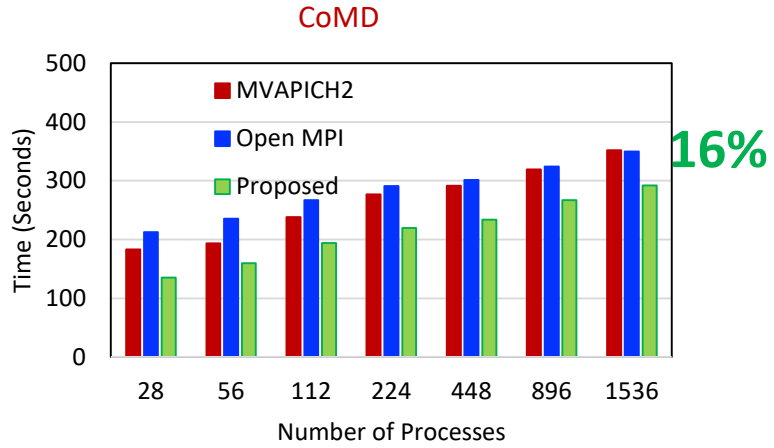
- Up to 76% benefits over MVAPICH2 for Neuron using Direct Connected transport protocol at scale
  - VERSION 7.6.2 master (f5a1284) 2018-08-15
- Numbers taken on bbpv2.epfl.ch
  - Knights Landing nodes with 64 ppn
  - ./x86_64/special -mpi -c stop_time=2000 -c is_split=1 parinit.hoc
  - Used "runtime" reported by execution to measure performance
- Environment variables used
  - MV2_USE_DC=1
  - MV2_NUM_DC_TGT=64
  - MV2_SMALL_MSG_DC_POOL=96
  - MV2_LARGE_MSG_DC_POOL=96
  - MV2_USE_RDMA_CM=0

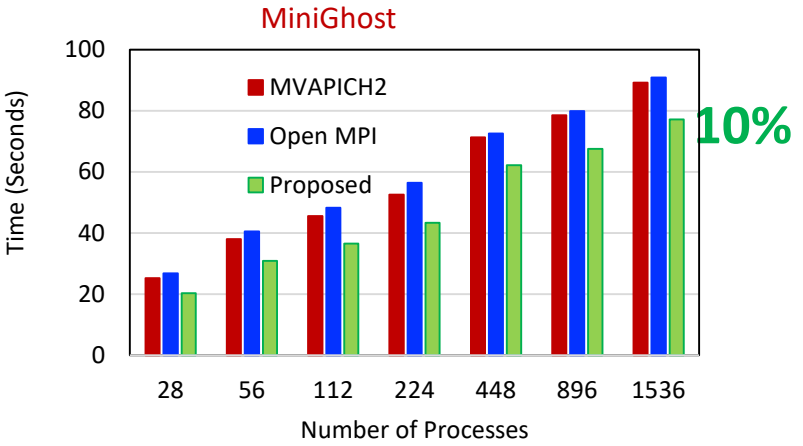
More details in EPFL Talk (Tomorrow)

Available from MVAPICH2-X 2.3rc2 onwards

Cooperative Rendezvous Protocols







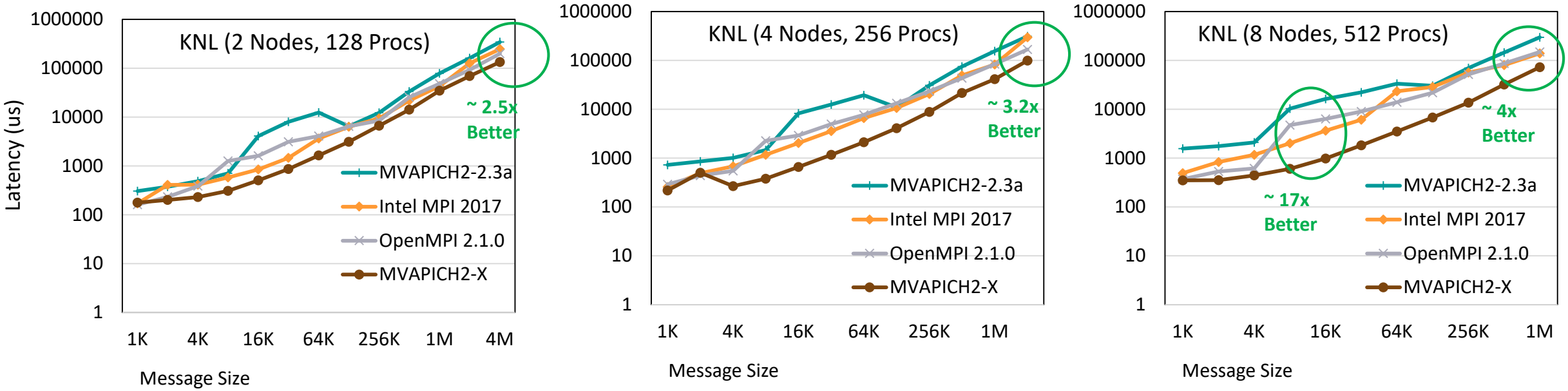
- Use both sender and receiver CPUs to progress communication concurrently
- Dynamically select rendezvous protocol based on communication primitives and sender/receiver availability (load balancing)
- Up to 2x improvement in large message latency and bandwidth
- Up to 19% improvement for Graph500 at 1536 processes

Cooperative Rendezvous Protocols for Improved Performance and Overlap
S. Chakraborty, M. Bayatpour, J Hashmi, H. Subramoni, and DK Panda,
SC '18 (Best Student Paper Award Finalist)

Platform: 2x14 core Broadwell 2680 (2.4 GHz)
Mellanox EDR ConnectX-5 (100 GBps)
Baseline: MVAPICH2X-2.3rc1, Open MPI v3.1.0

Available in MVAPICH2-X 2.3rc2

Optimized CMA-based Collectives for Large Messages



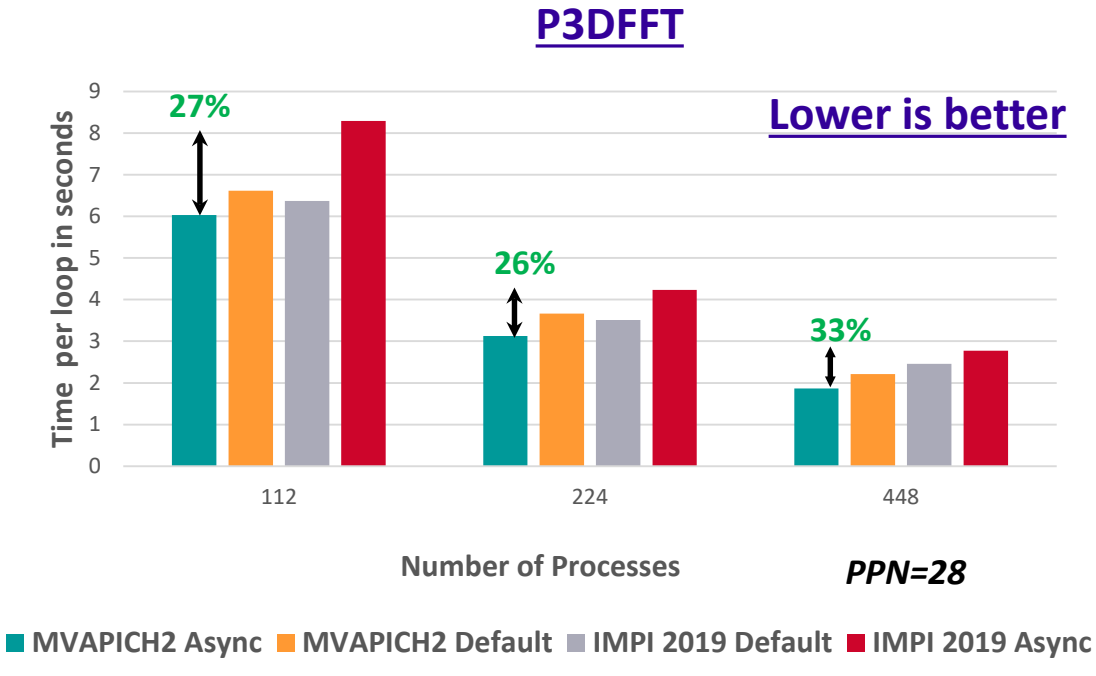
Performance of MPI_Gather on KNL nodes (64PPN)

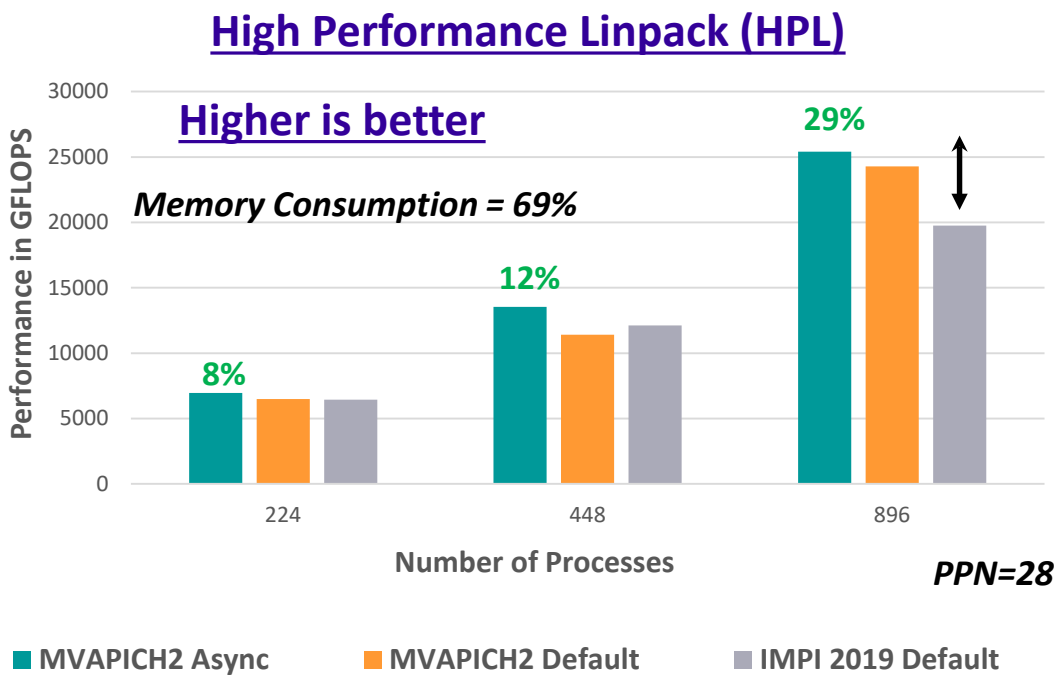
- Significant improvement over existing implementation for Scatter/Gather with 1MB messages (up to 4x on KNL, 2x on Broadwell, 14x on OpenPOWER)
- New two-level algorithms for better scalability
- Improved performance for other collectives (Bcast, Allgather, and Alltoall)

S. Chakraborty, H. Subramoni, and D. K. Panda, **Contention Aware Kernel-Assisted MPI Collectives for Multi/Many-core Systems,** ***IEEE Cluster '17, BEST Paper Finalist***

Available since MVAPICH2-X 2.3b

Benefits of the New Asynchronous Progress Design: Broadwell + InfiniBand





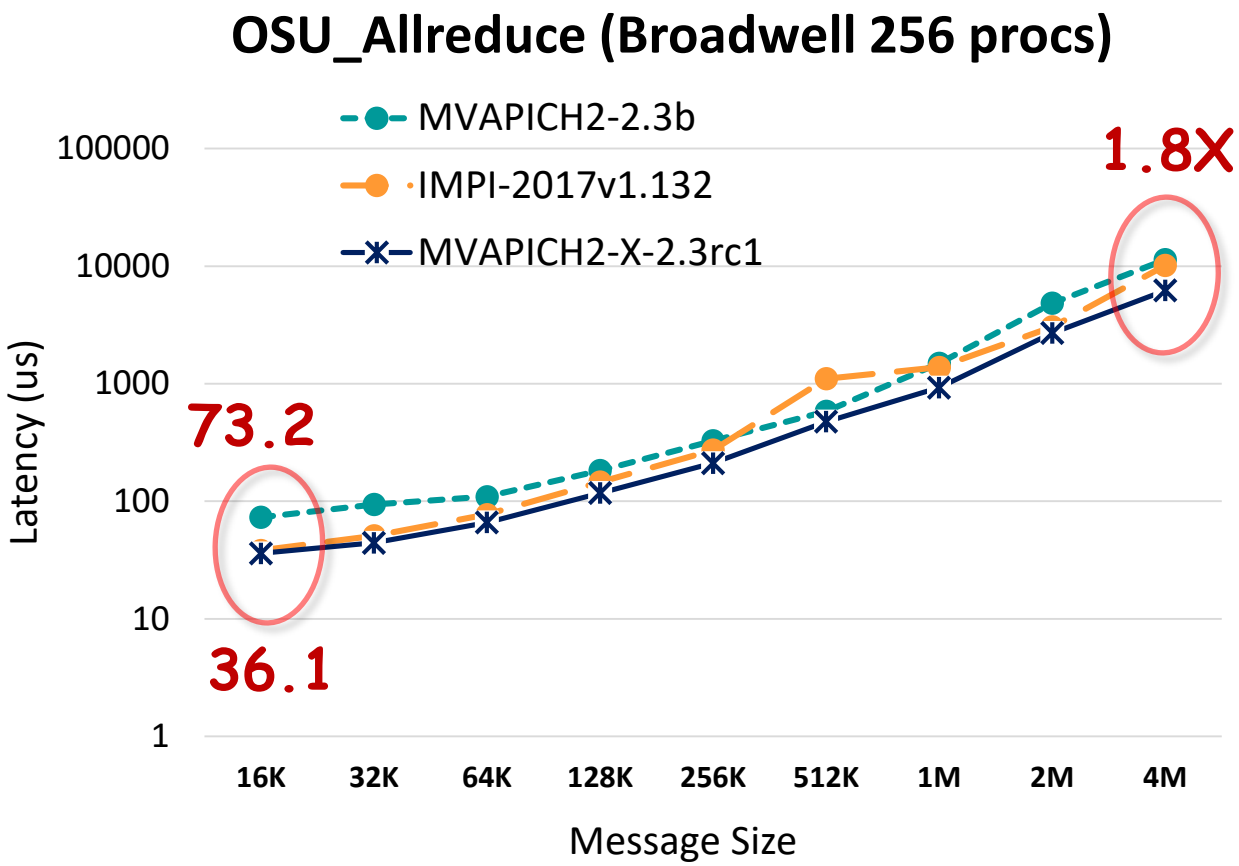
Up to 33% performance improvement in P3DFFT application with 448 processes

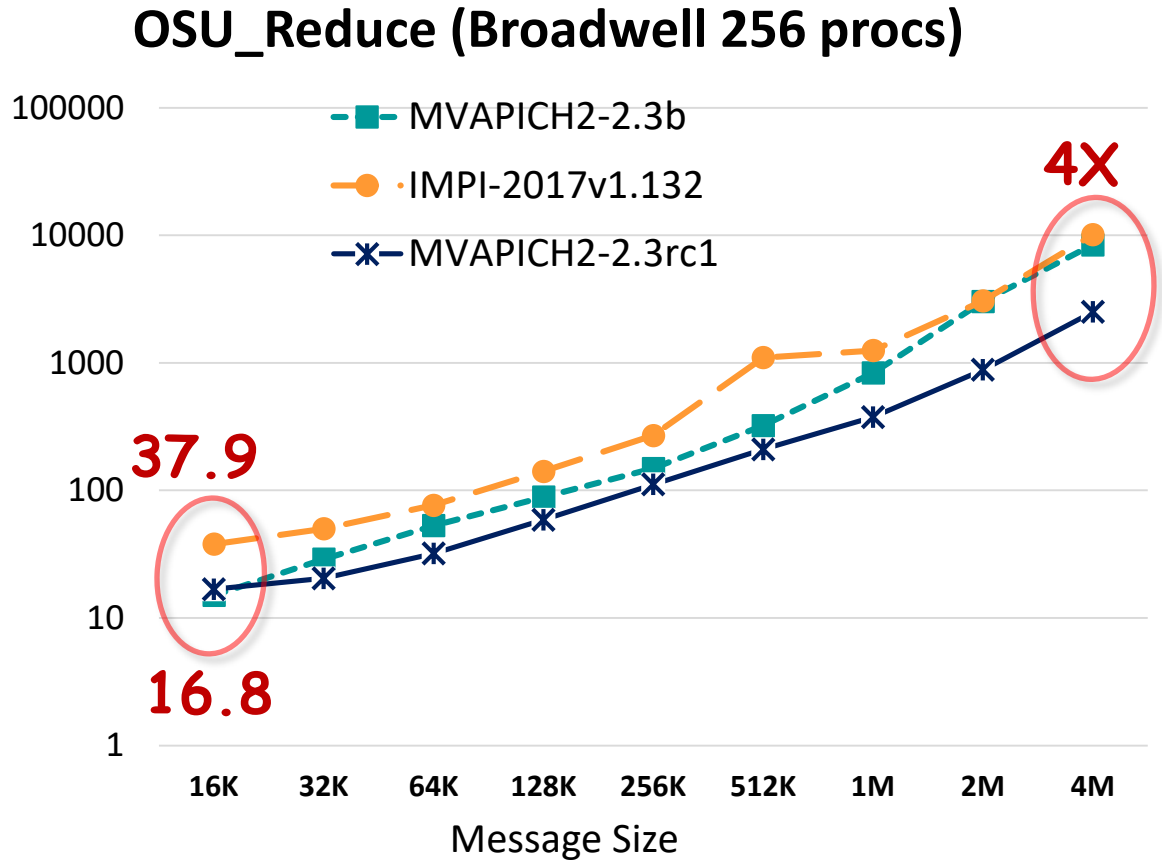
Up to 29% performance improvement in HPL application with 896 processes

A. Ruhela, H. Subramoni, S. Chakraborty, M. Bayatpour, P. Kousha, and D.K. Panda, "Efficient design for MPI Asynchronous Progress without Dedicated Resources", Parallel Computing 2019

Available since MVAPICH2-X 2.3rc1

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

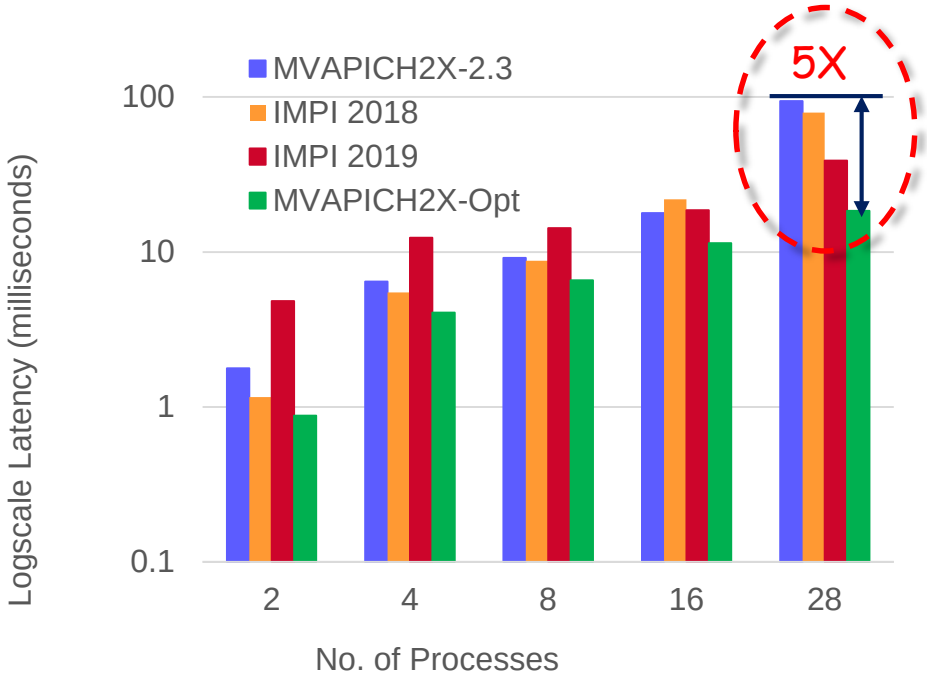
J. Hashmi, S. Chakraborty, M. Bayatpour, H. Subramoni, and D. Panda, Designing Efficient Shared Address Space Reduction Collectives for Multi-/Many-cores, International Parallel & Distributed Processing Symposium (IPDPS '18), May 2018.

Available since MVAPICH2-X 2.3rc1

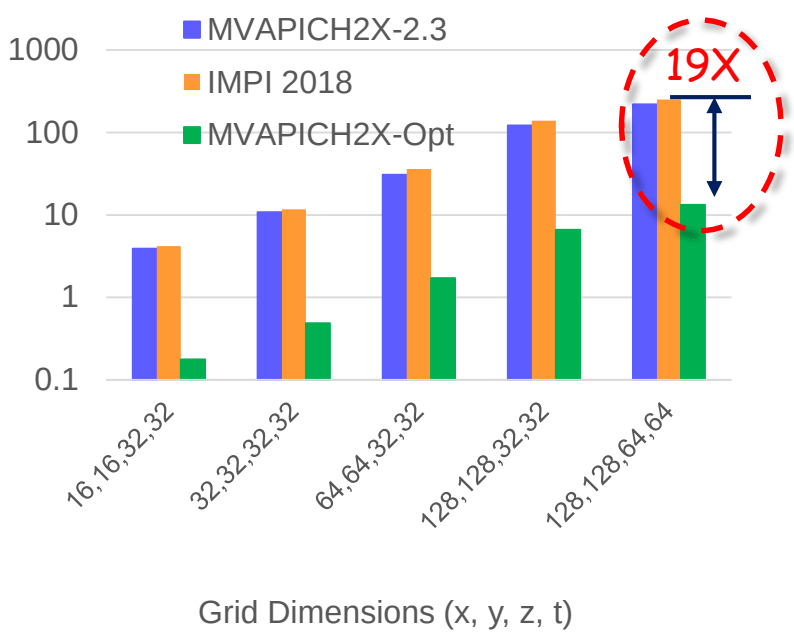
MVAPICH2-X Upcoming Features

- XPMEM-based MPI Derived Datatype Designs
- Exploiting Hardware Tag Matching
- Neighborhood Collectives

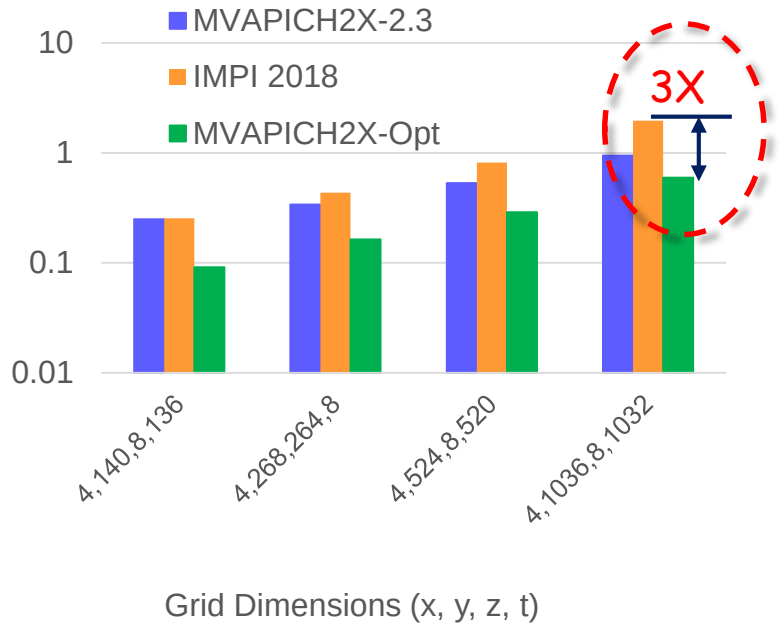
Efficient Zero-copy MPI Datatypes for Emerging Architectures



3D-Stencil Datatype Kernel on **Broadwell** (2x14 core)



MILC Datatype Kernel on **KNL 7250** in Flat-Quadrant Mode (64-core)



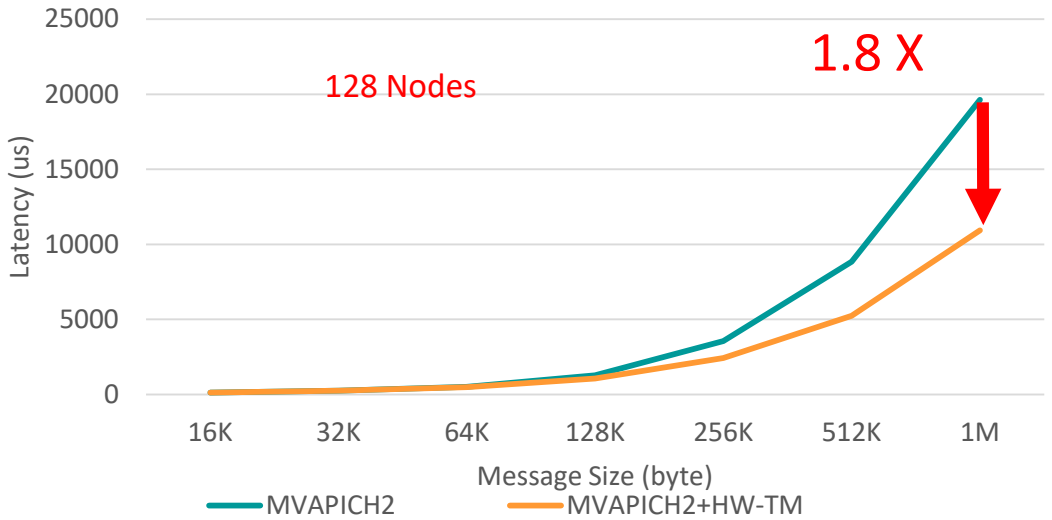
NAS-MG Datatype Kernel on **OpenPOWER** (20-core)

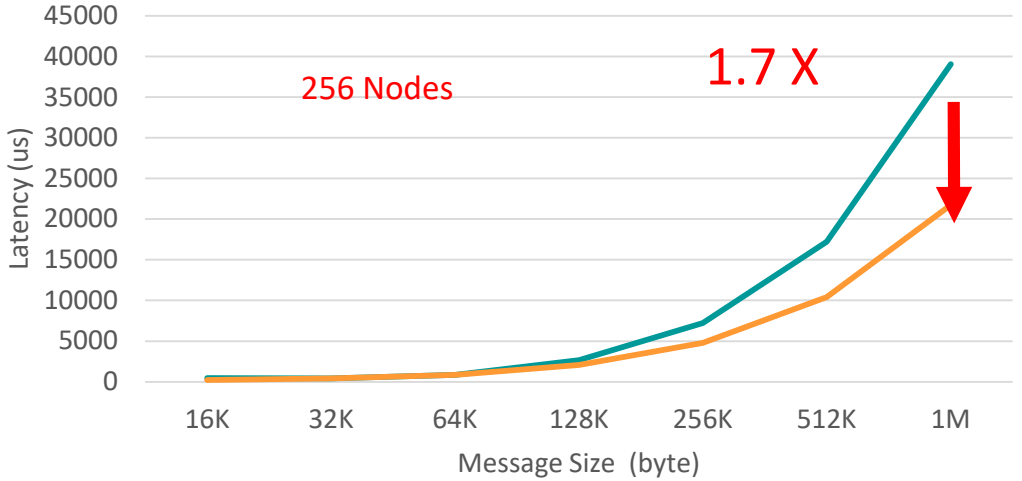
- New designs for efficient zero-copy based MPI derived datatype processing
- Efficient schemes mitigate datatype translation, packing, and exchange overheads
- Demonstrated benefits over prevalent MPI libraries for various application kernels
- To be available in the upcoming MVAPICH2-X release

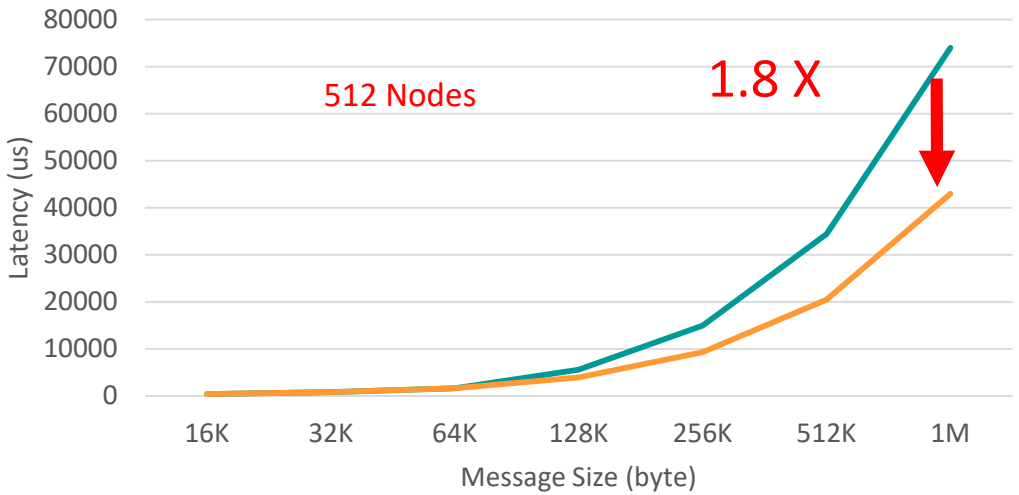
Hardware Tag Matching Support

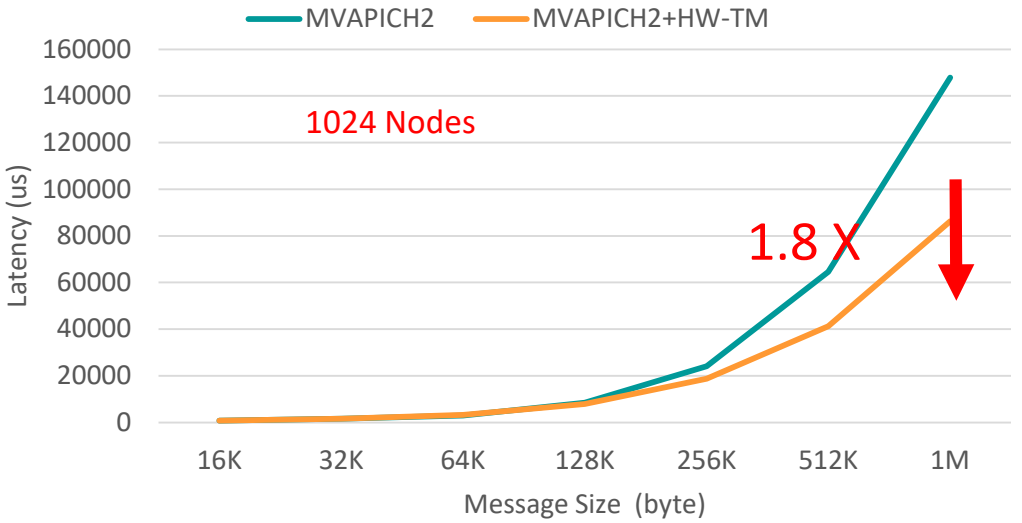
- Offloads the processing of point-to-point MPI messages from the host processor to HCA
- Enables zero copy of MPI message transfers
  - Messages are written directly to the user's buffer without extra buffering and copies
- Provides rendezvous progress offload to HCA
  - Increases the overlap of communication and computation

Performance of MPI_Iscatterv using HW Tag Matching



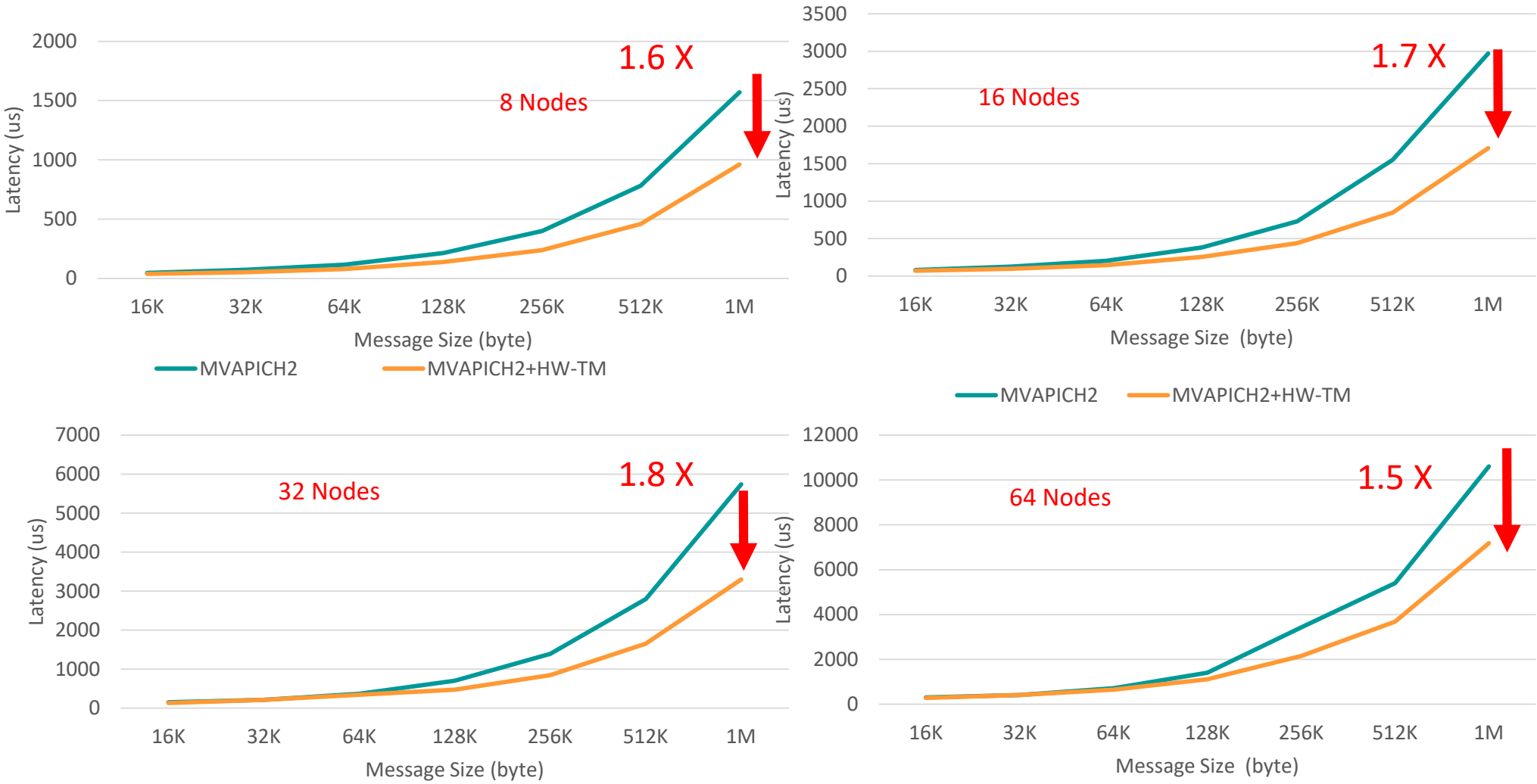






- Up to 1.8x Performance Improvement, Sustained benefits as system size increases

Performance of MPI_Ialltoall using HW Tag Matching



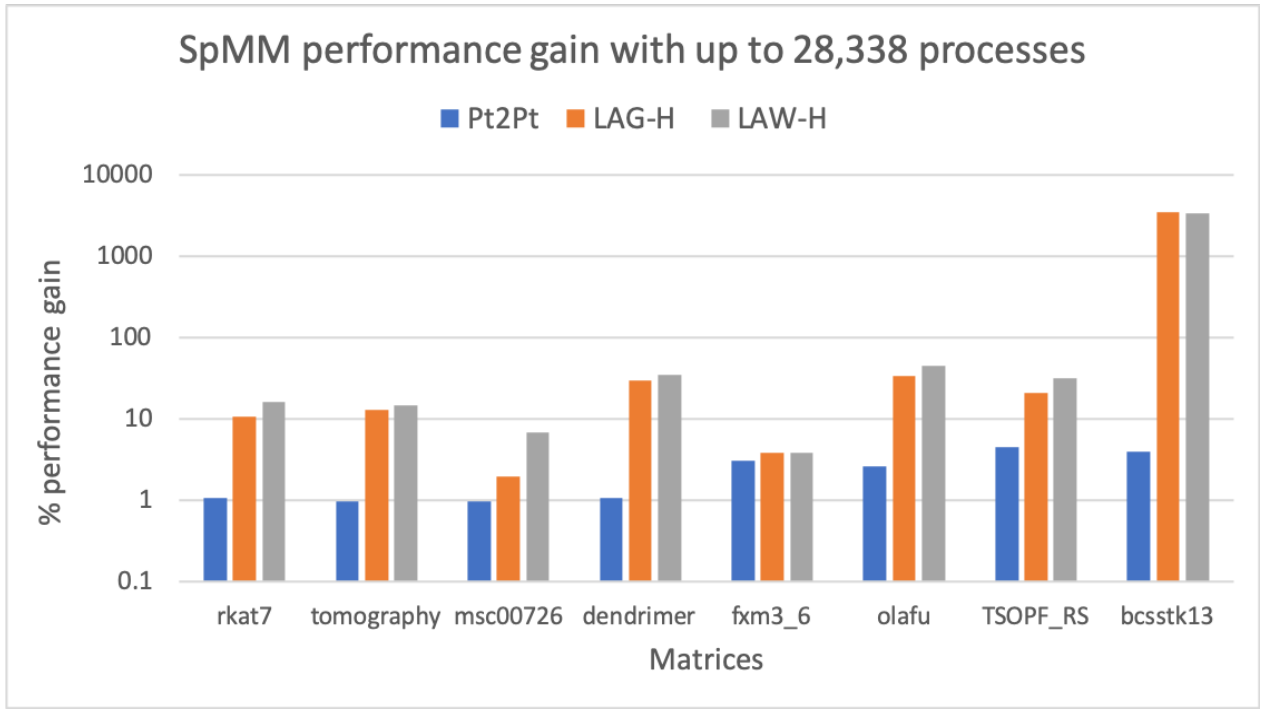
More details in Bayatpour's Talk (later today)

- Up to 1.8x Performance Improvement, Sustained benefits as system size increases

Neighborhood Collectives – Performance Benefits

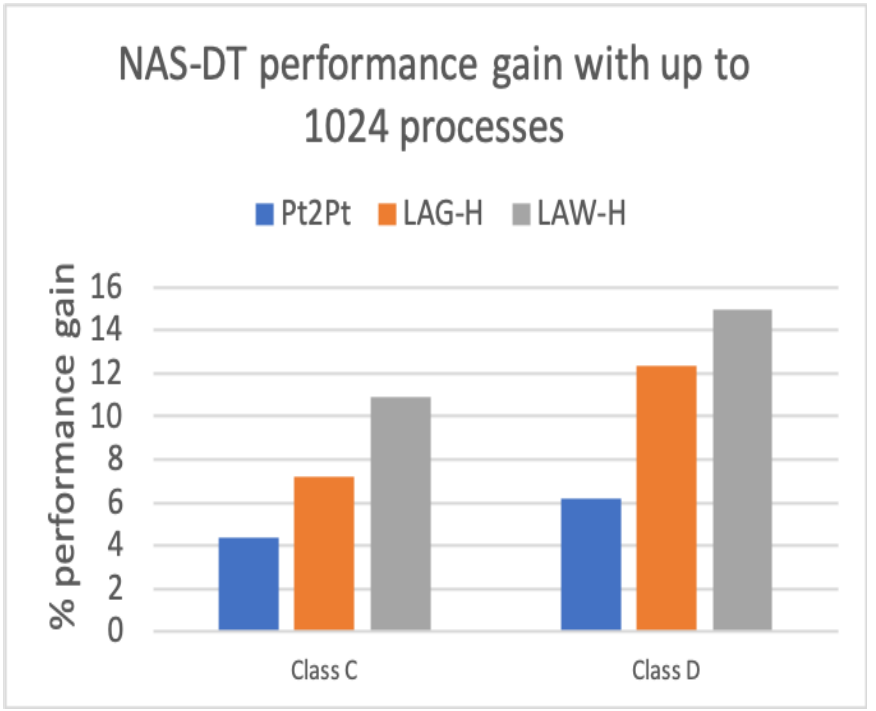
More details in Ghazimirsaeed's Talk (Later today)

- SpMM



up to 34x speedup

- NAS DT



up to 15% improvement

M. Ghazimirsaeed, Q. Zhou, A. Ruhela, M. Bayatpour, H. Subramoni, and DK Panda, "A Hierarchical and Load-Aware Design for Large Message Neighborhood Collectives", SC '20

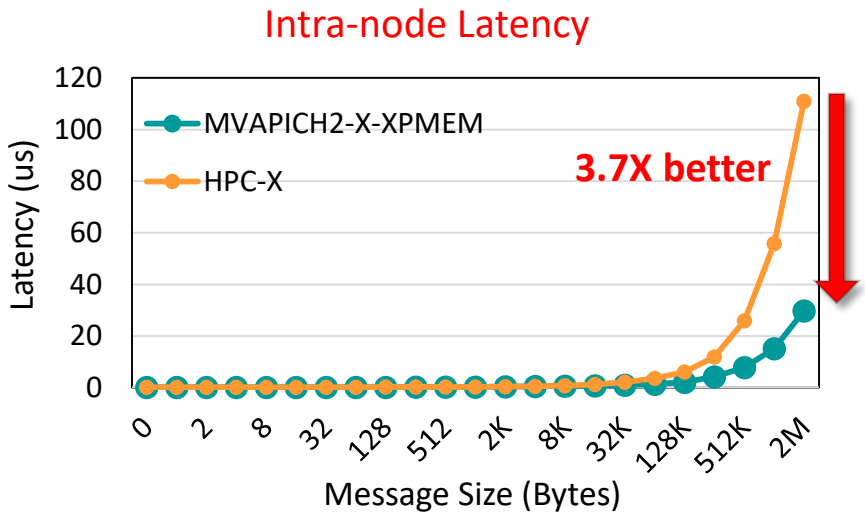
MVAPICH2 Software Family

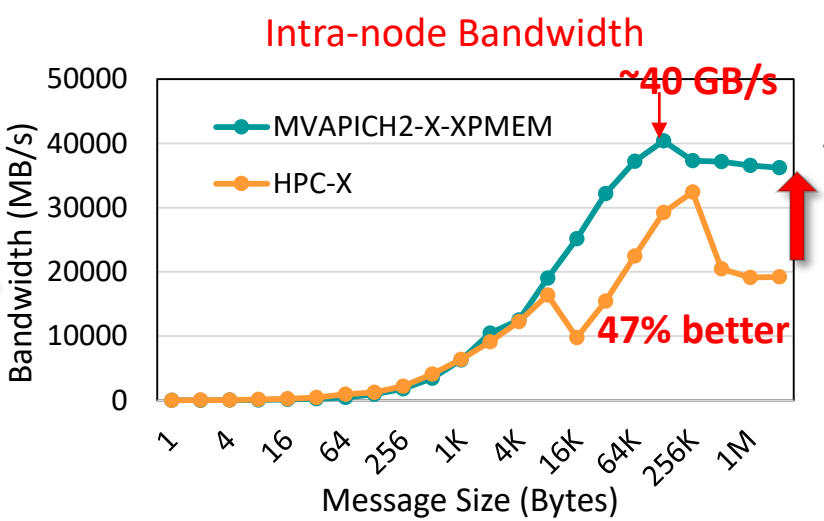
Requirements	Library
MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2
Advanced MPI features/support (UMR, ODP, DC, Core-Direct, SHARP, XPMEM), OSU INAM (InfiniBand Network Monitoring and Analysis), MPI+PGAS	MVAPICH2-X
Optimized Support of MVAPICH2-X for Microsoft Azure Platform with InfiniBand	MVAPICH2-X-Azure
Advanced MPI features (SRD and XPMEM) with support for Amazon Elastic Fabric Adapter (EFA)	MVAPICH2-X-AWS
Optimized MPI for clusters with NVIDIA GPUs and for GPU-enabled Deep Learning Applications	MVAPICH2-GDR
Energy-aware MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

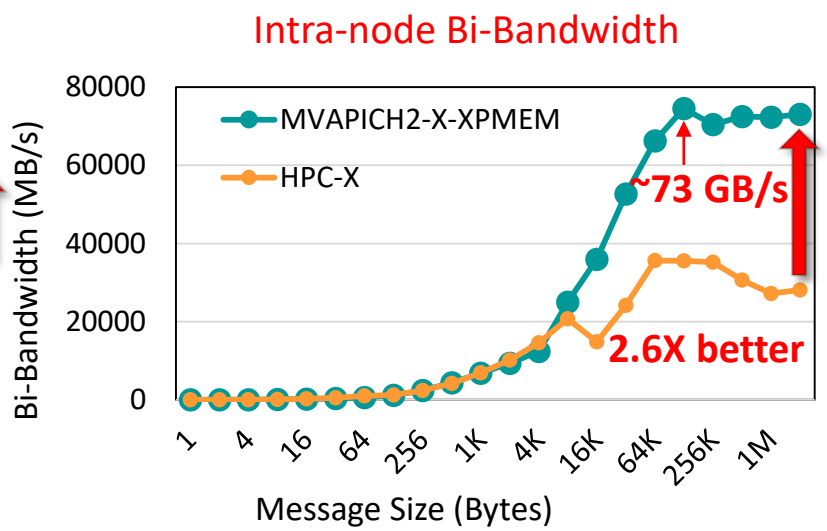
MVAPICH2-Azure Deployment

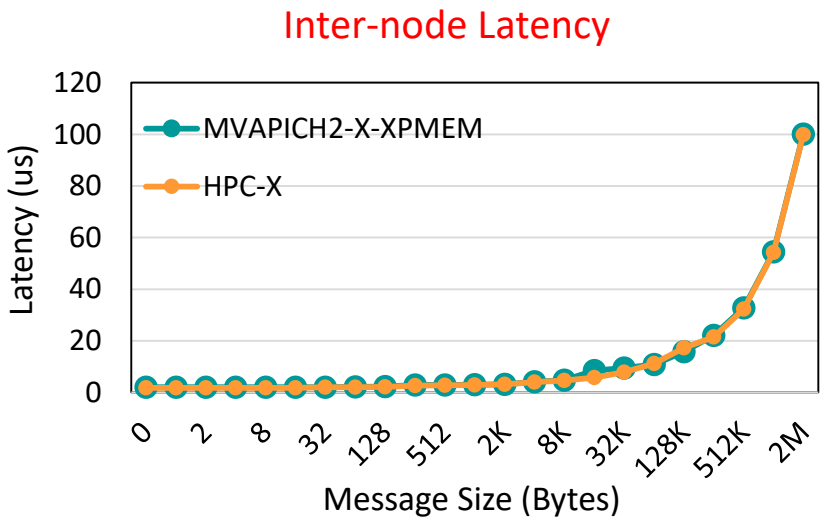
- **Released on 05/20/2020**
- Integrated Azure CentOS HPC Images
  - **https://github.com/Azure/azhpc-images/releases/tag/centos-7.6-hpc-20200417**
- MVAPICH2 2.3.3
  - **CentOS Images (7.6, 7.7 and 8.1)**
  - **Tested with multiple VM instances**
- MVAPICH2-X 2.3.RC3
  - **CentOS Images (7.6, 7.7 and 8.1)**
  - **Tested with multiple VM instances**
- **More details from Azure Blog Post**
  - **https://techcommunity.microsoft.com/t5/azure-compute/mvapich2-on-azure-hpc-clusters/ba-p/1404305**

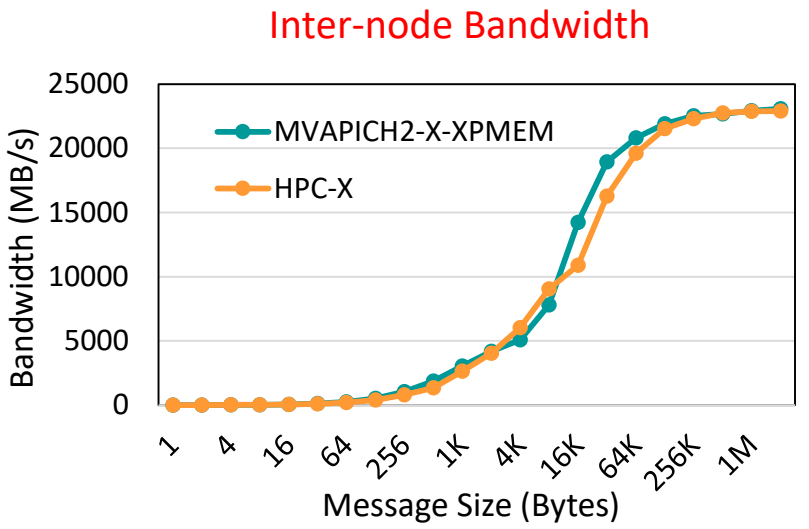
Point-to-Point Evaluation on Azure HBv2 (AMD Rome)

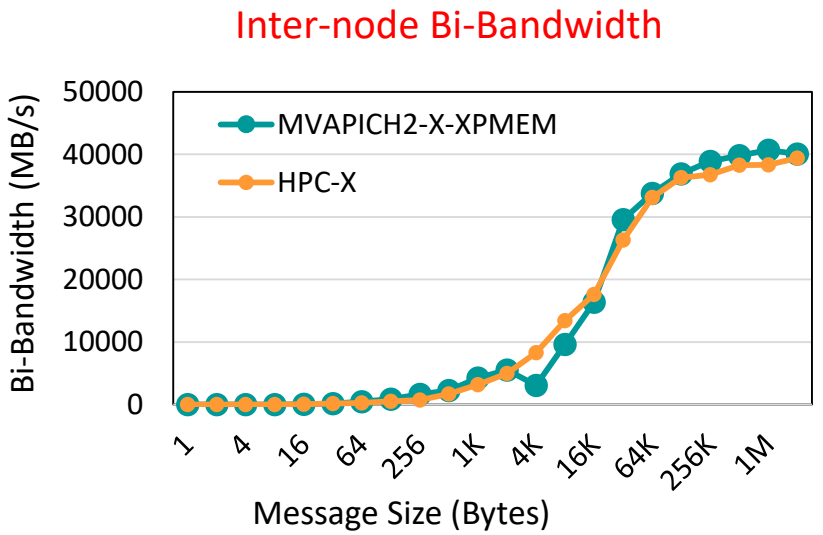






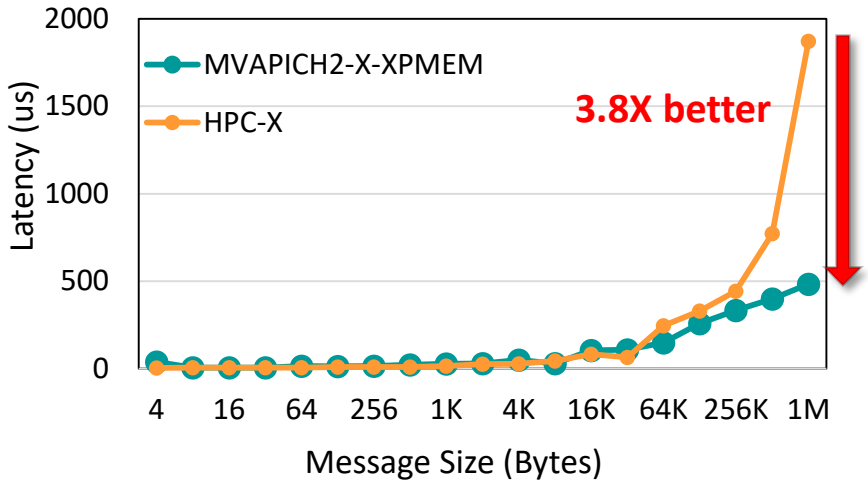




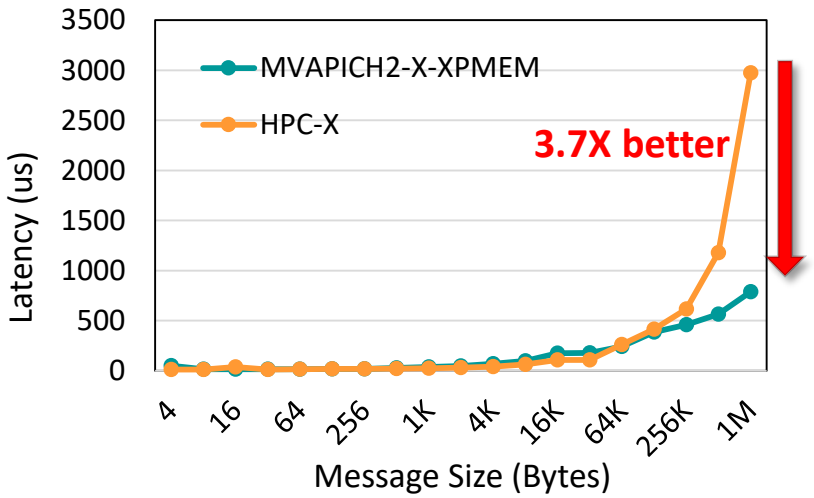


Collectives Performance on Azure HBv2 (AMD Rome)

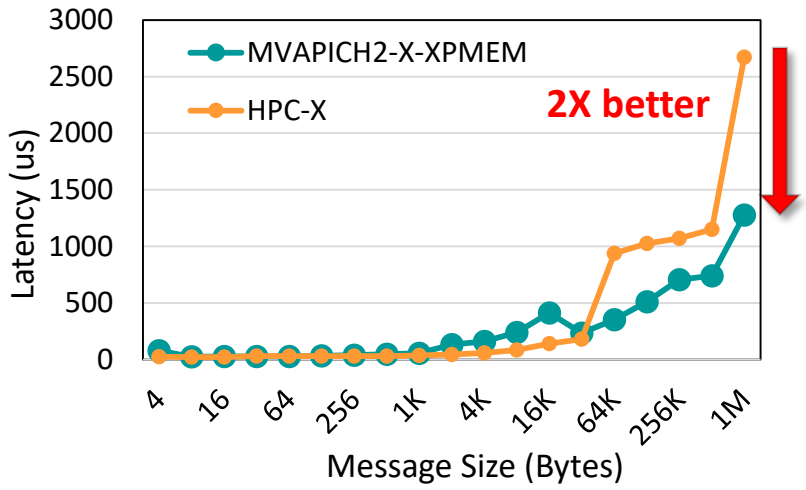
osu_allreduce (NP=120, PPN=120)



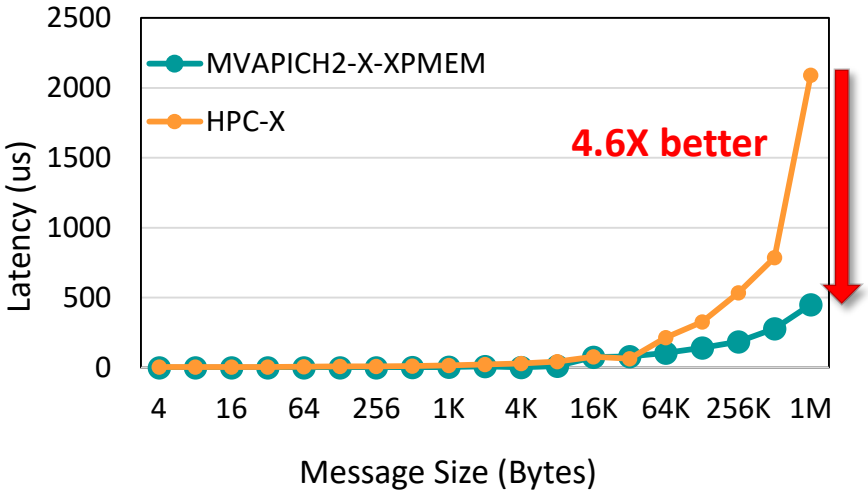
osu_allreduce (NP=480, PPN=120)



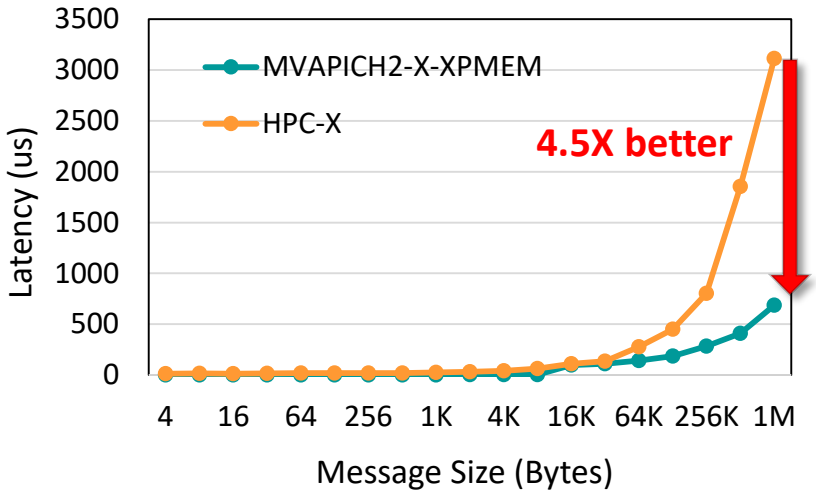
osu_allreduce (NP=960, PPN=120)



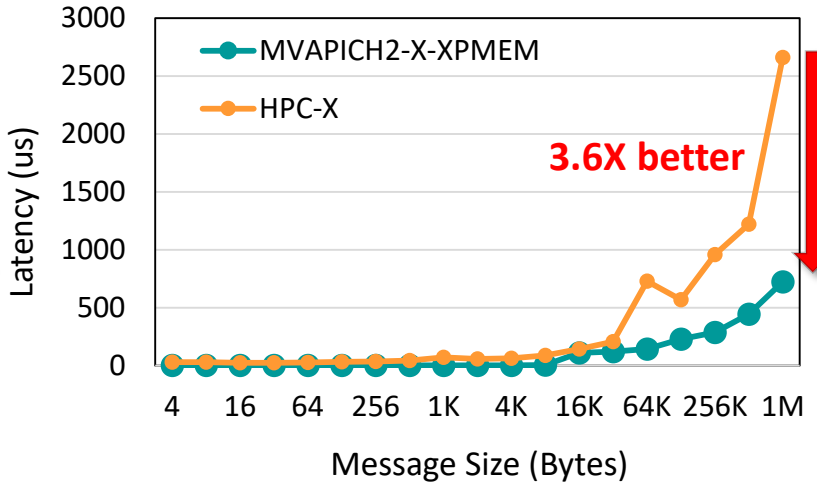
osu_reduce (NP=120, PPN=120)



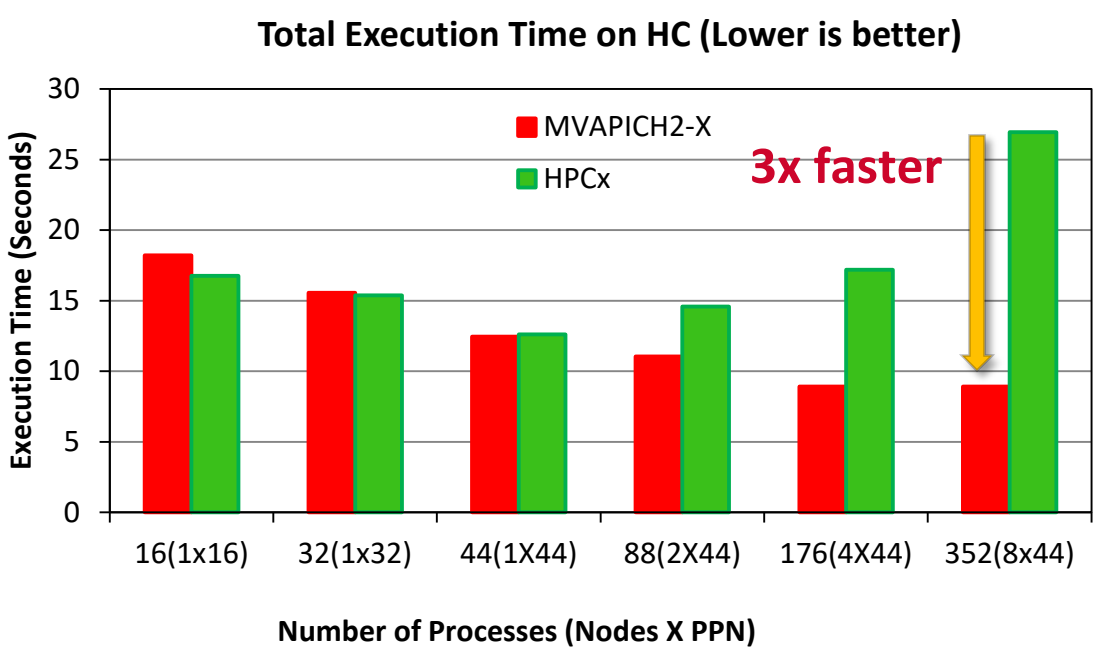
osu_reduce (NP=480, PPN=120)

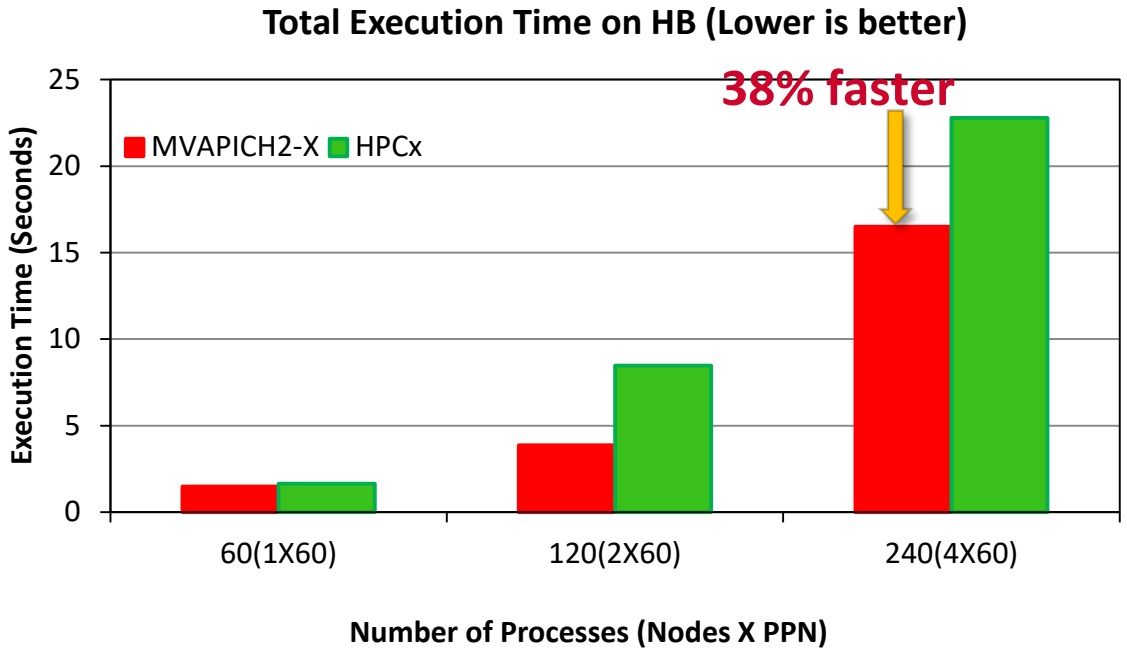


osu_reduce (NP=960, PPN=120)



Performance of Radix

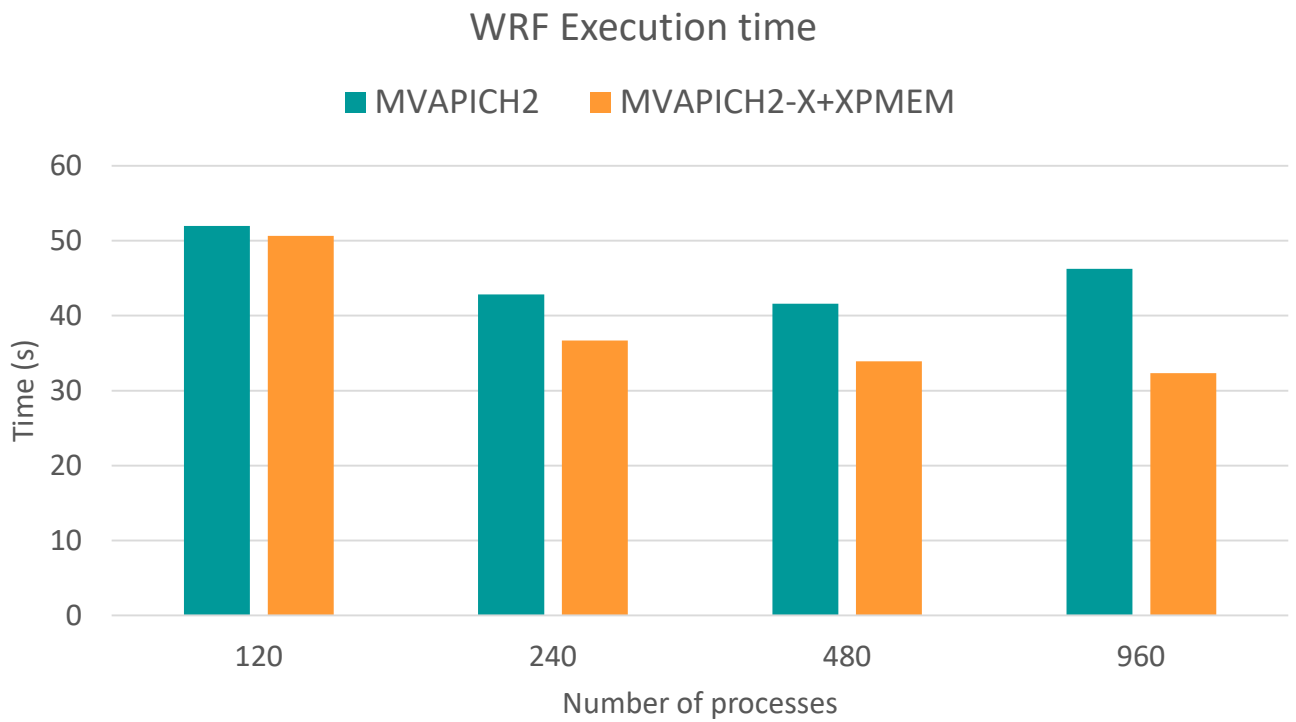




More Results in Microsoft Azure Talk (Later Today)

WRF Application Results on HBv2 (AMD Rome)

- Performance of WRF with MVAPICH2 and MVAPICH2-X-XPMEM



- WRF 3.6
  - https://github.com/hanschen/WRFV3
- Benchmark: 12km resolution case over the Continental U.S. (CONUS) domain
  - https://www2.mmm.ucar.edu/wrf/WG2/benchv3/#_Toc212961288
- Update io_form_history in namelist.input to 102
  - https://www2.mmm.ucar.edu/wrf/users/namelist_best_prac_wrf.html#io_form_history

MVAPICH2-X-XPMEM is able to deliver better performance and scalability

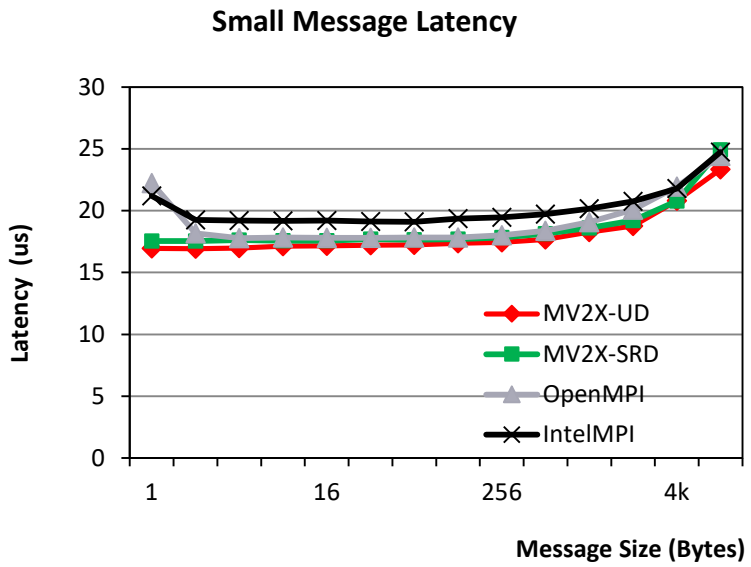
MVAPICH2 Software Family

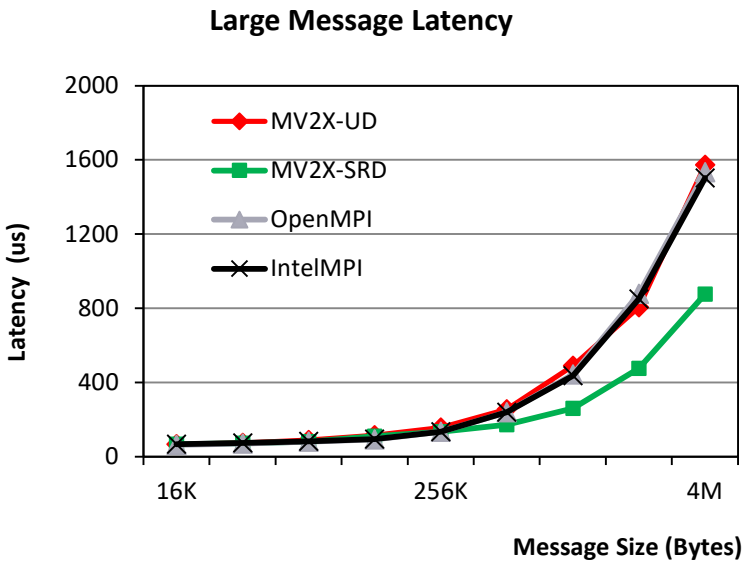
Requirements	Library
MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2
Advanced MPI features/support (UMR, ODP, DC, Core-Direct, SHARP, XPMEM), OSU INAM (InfiniBand Network Monitoring and Analysis), MPI+PGAS	MVAPICH2-X
Optimized Support of MVAPICH2-X for Microsoft Azure Platform with InfiniBand	MVAPICH2-X-Azure
Advanced MPI features (SRD and XPMEM) with support for Amazon Elastic Fabric Adapter (EFA)	MVAPICH2-X-AWS
Optimized MPI for clusters with NVIDIA GPUs and for GPU-enabled Deep Learning Applications	MVAPICH2-GDR
Energy-aware MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

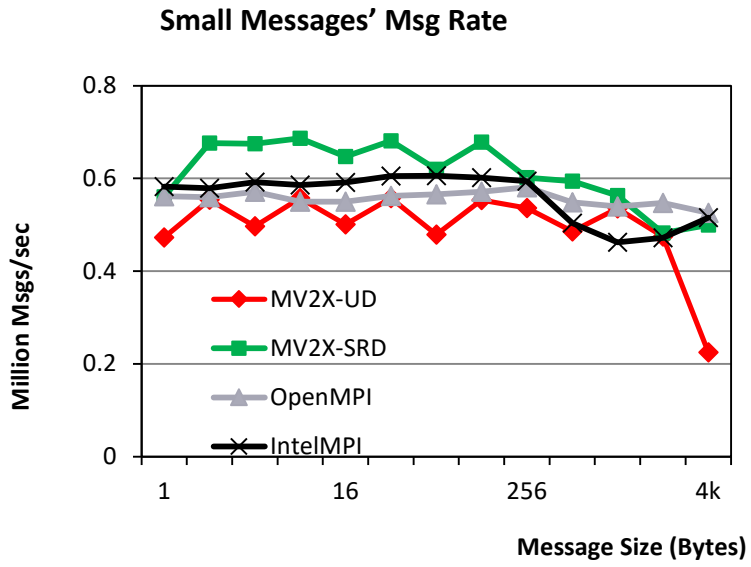
MVAPICH2-X-AWS 2.3

- **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**

Point-to-Point Inter-node Performance



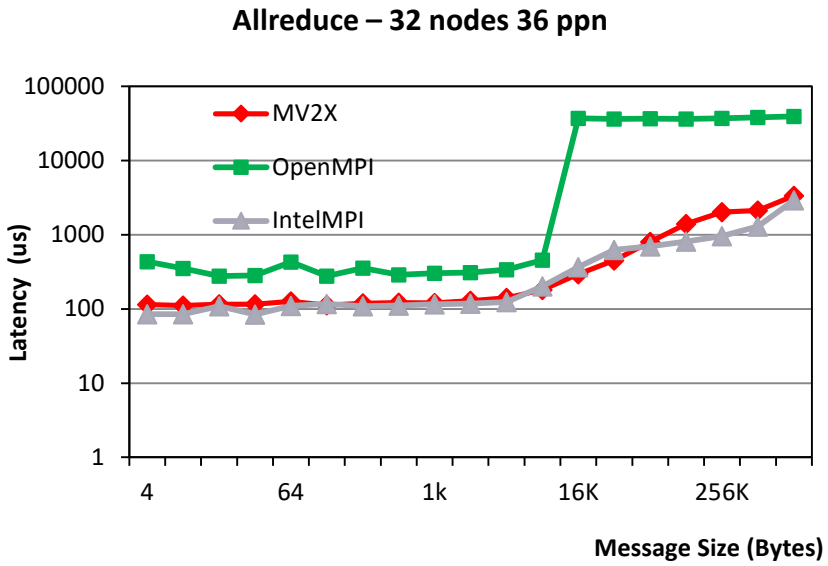


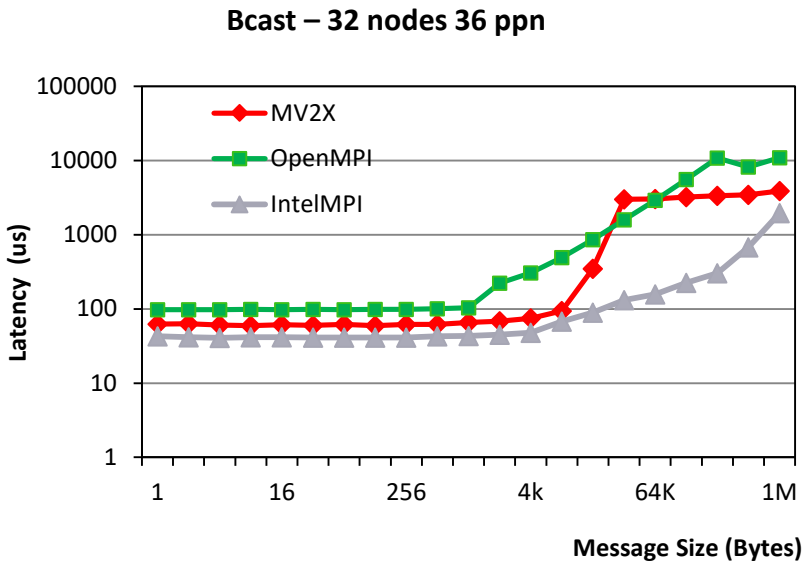


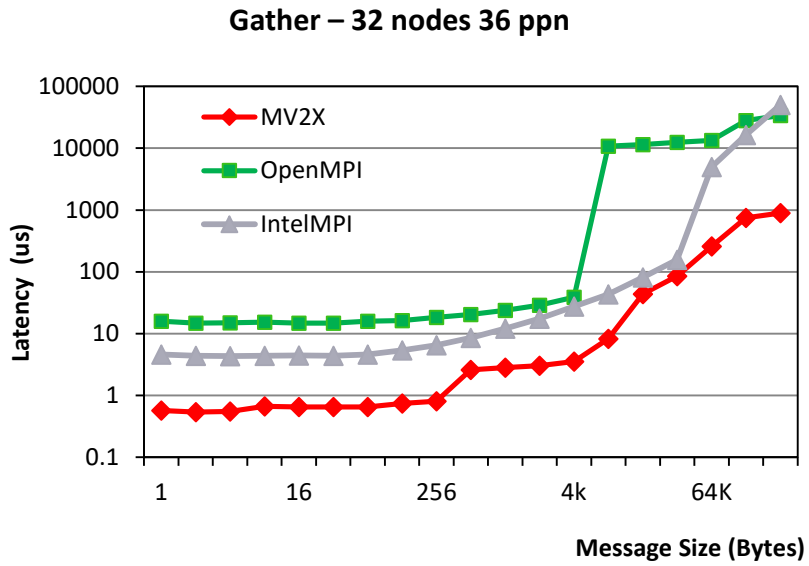
- Both UD and SRD shows similar latency for small messages
- SRD shows higher message rate due to lack of software reliability overhead
- SRD is faster for large messages due to larger MTU size
- UD show lower message rate for messages that larger than MTU size (4K)

S. Chakraborty, S. Xu, H. Subramoni, and D. K. Panda, *Designing Scalable and High-performance MPI Libraries on Amazon Elastic Fabric Adapter,* 26th Symposium on High Performance Interconnects, (HOTI '19)

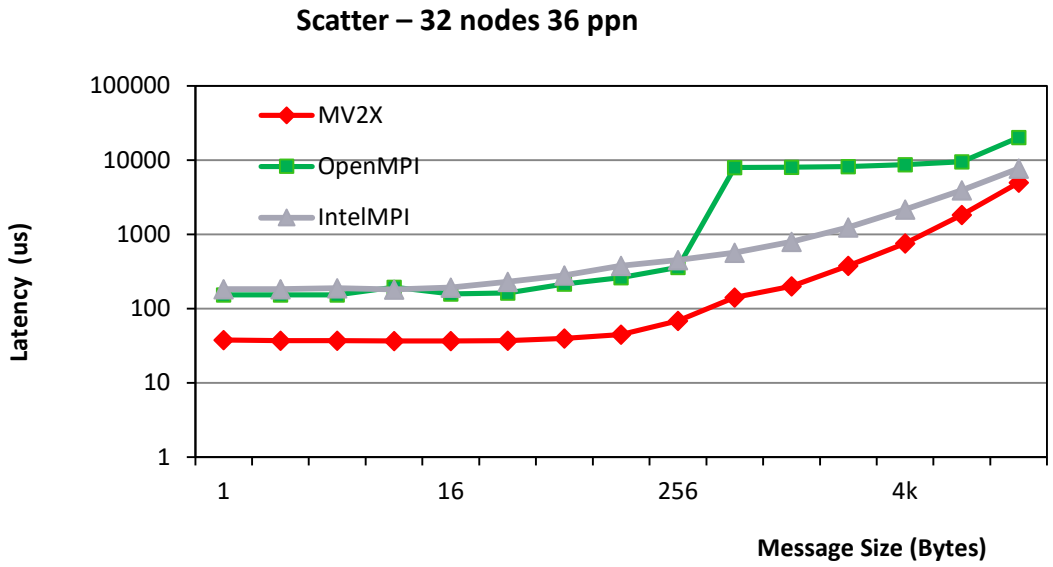
Collective Performance (1,152 processes)





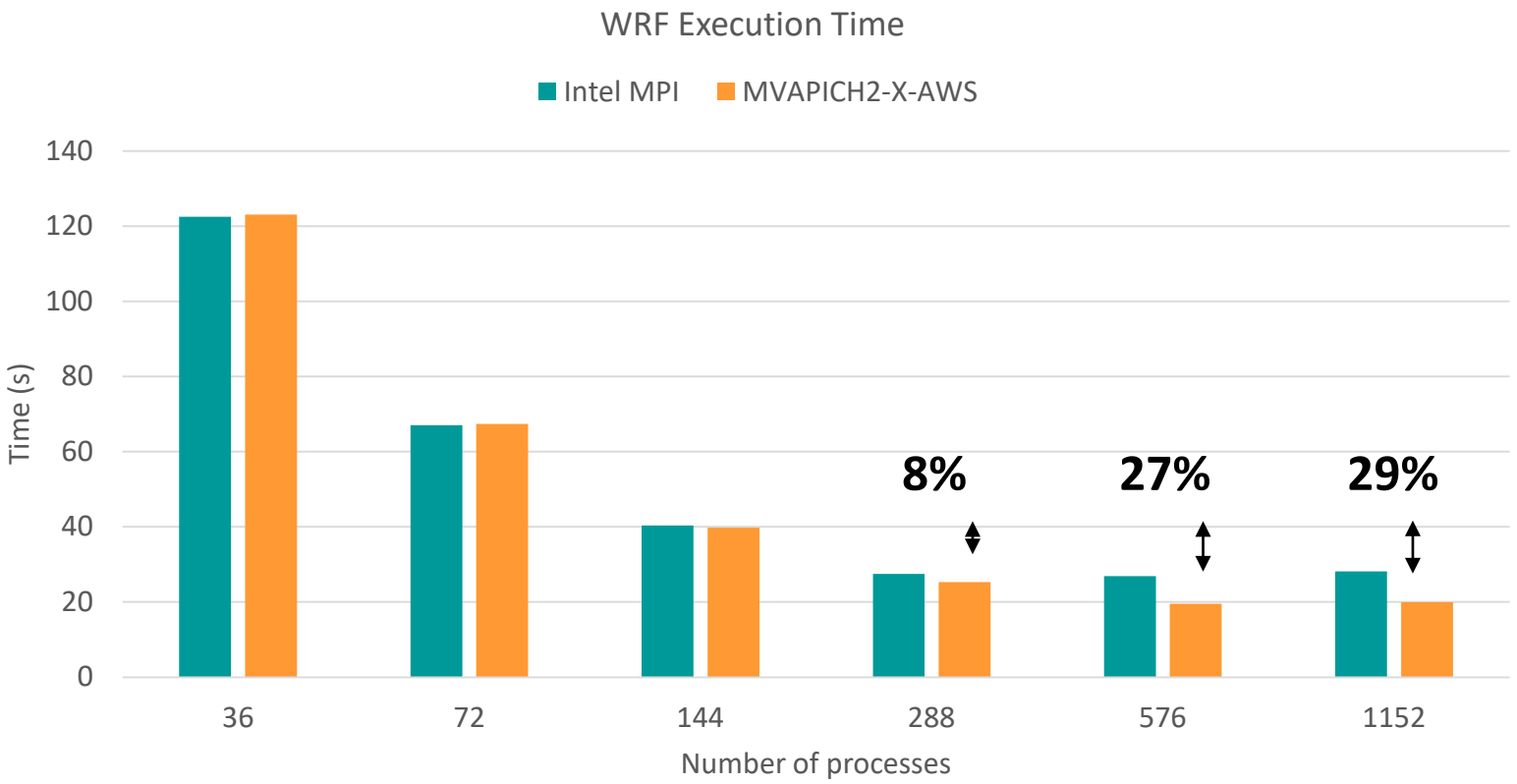


- Instance type: c5n.18xlarge
- CPU: Intel Xeon Platinum 8124M @ 3.00GHz
- MVAPICH2 version: Latest MVAPICH2-X + SRD support
- OpenMPI version: Open MPI v4.0.3 with libfabric 1.8
- IntelMPI version: Intel MPI 2019.7.217 with libfabric 1.8



WRF Application Results

- Performance of WRF with Intel MPI 2019.7.217 vs MVAPICH2-X-AWS v2.3



- WRF 3.6
  - https://github.com/hanschen/WRFV3
- Benchmark: 12km resolution case over the Continental U.S. (CONUS) domain
  - https://www2.mmm.ucar.edu/wrf/WG2/benchv3/#_Toc212961288
- Update io_form_history in namelist.input to 102
  - https://www2.mmm.ucar.edu/wrf/users/namelist_best_prac_wrf.html#io_form_history

More Results in AWS Talk (Later Today)

MVAPICH2 Software Family

Requirements	Library
MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2
Advanced MPI features/support (UMR, ODP, DC, Core-Direct, SHARP, XPMEM), OSU INAM (InfiniBand Network Monitoring and Analysis), MPI+PGAS	MVAPICH2-X
Optimized Support of MVAPICH2-X for Microsoft Azure Platform with InfiniBand	MVAPICH2-X-Azure
Advanced MPI features (SRD and XPMEM) with support for Amazon Elastic Fabric Adapter (EFA)	MVAPICH2-X-AWS
Optimized MPI for clusters with NVIDIA GPUs and for GPU-enabled Deep Learning Applications	MVAPICH2-GDR
Energy-aware MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

Overview of Some of the MVAPICH2-GDR Features for HPC and DL

- CUDA-Aware MPI
- Scalable Collectives
- Kernel-based Support for Datatypes
- High-Performance Deep Learning

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

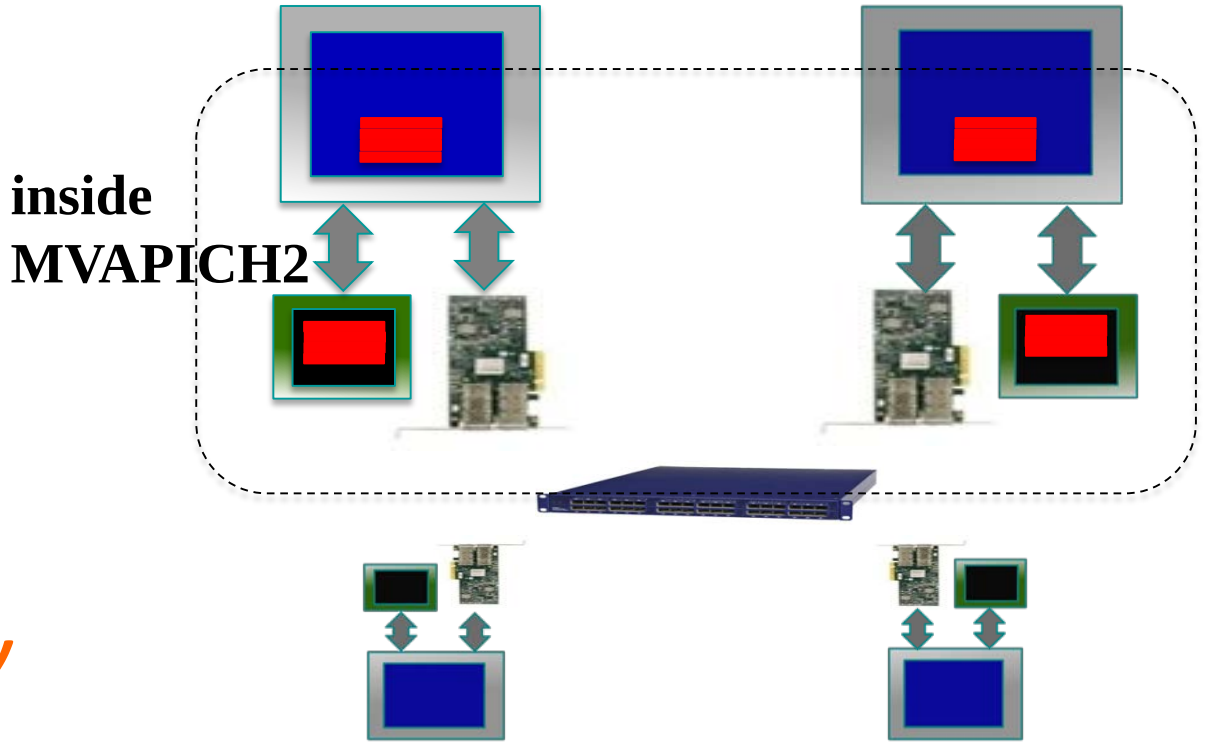
- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

```
MPI_Recv(r_devbuf, size, …);
```

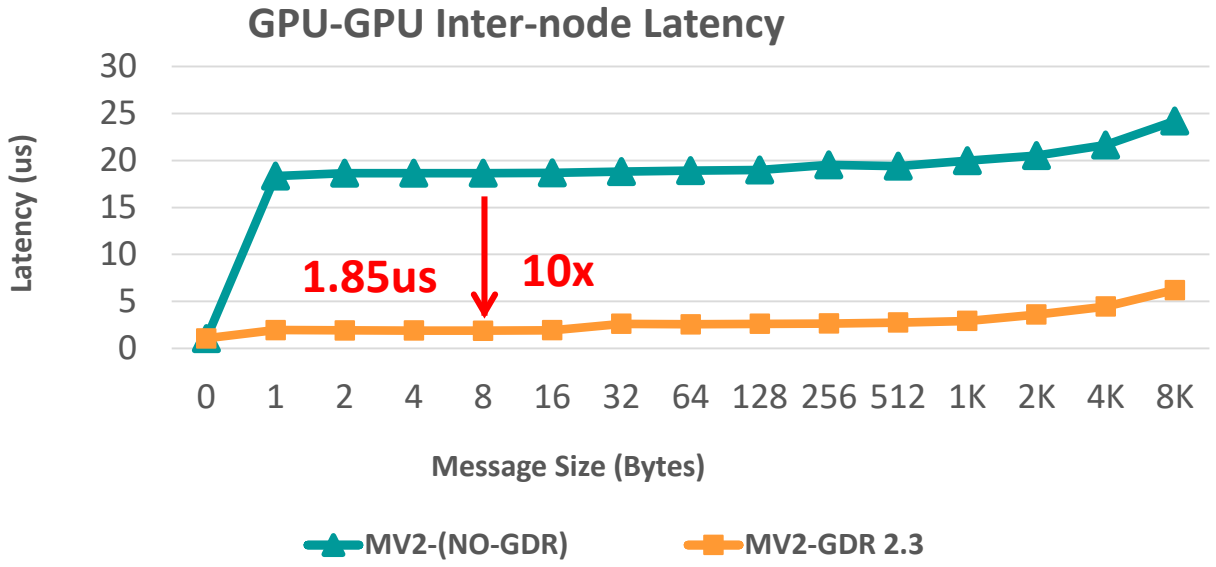


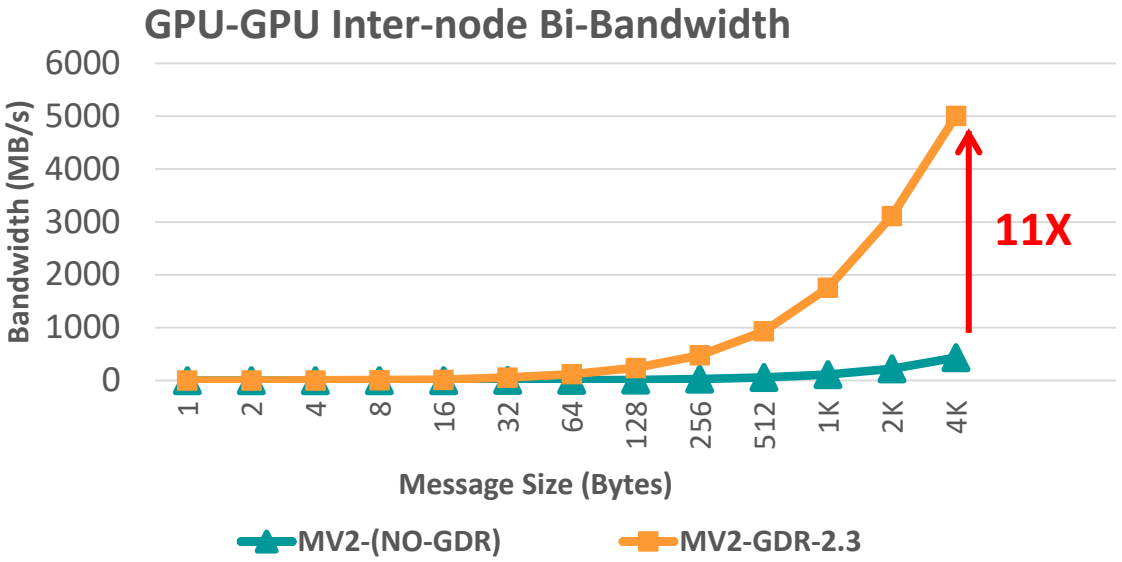
High Performance and High Productivity

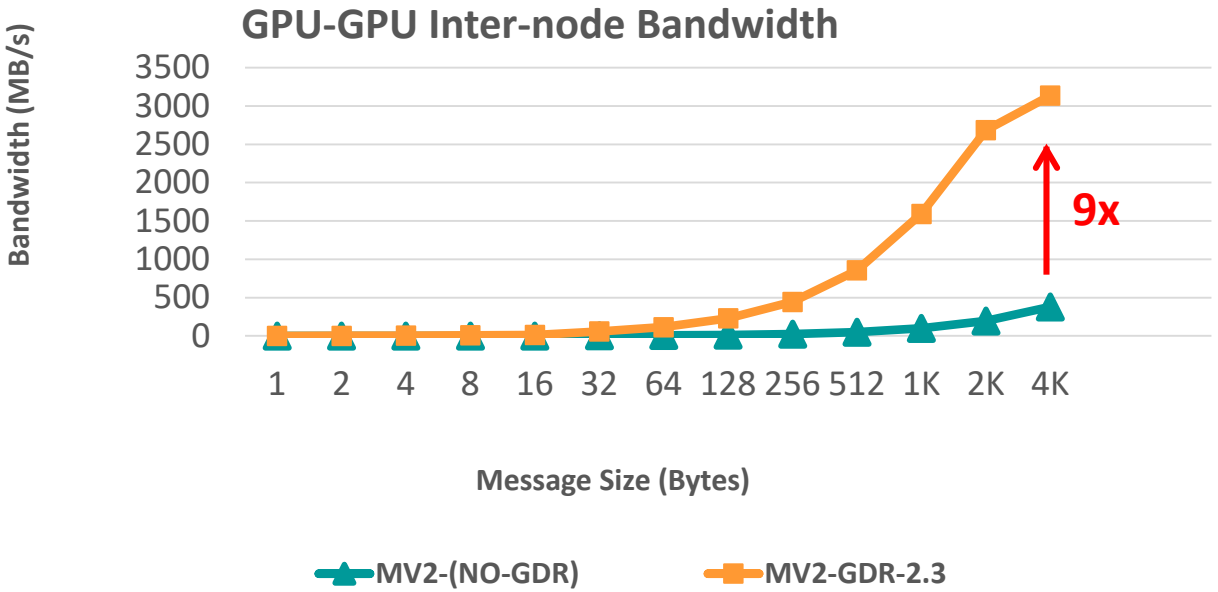
CUDA-Aware MPI: MVAPICH2-GDR 1.8-2.3.4 Releases

- Support for MPI communication from NVIDIA GPU device memory
- High performance RDMA-based inter-node point-to-point communication (GPU-GPU, GPU-Host and Host-GPU)
- High performance intra-node point-to-point communication for multi-GPU adapters/node (GPU-GPU, GPU-Host and Host-GPU)
- Taking advantage of CUDA IPC (available since CUDA 4.1) in intra-node communication for multiple GPU adapters/node
- Optimized and tuned collectives for GPU device buffers
- MPI datatype support for point-to-point and collective communication from GPU device buffers
- Unified memory

Optimized MVAPICH2-GDR Design



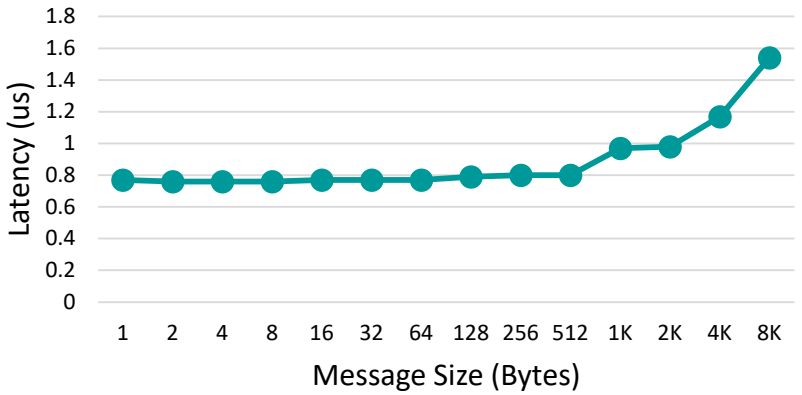




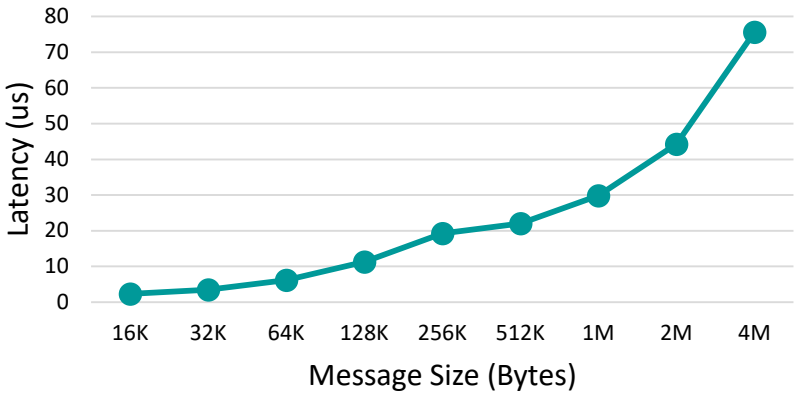
MVAPICH2-GDR-2.3
Intel Haswell (E5-2687W @ 3.10 GHz) node - 20 cores
NVIDIA Volta V100 GPU
Mellanox Connect-X4 EDR HCA
CUDA 9.0
Mellanox OFED 4.0 with GPU-Direct-RDMA

D-to-D Performance on OpenPOWER w/ GDRCopy (NVLink2 + Volta)

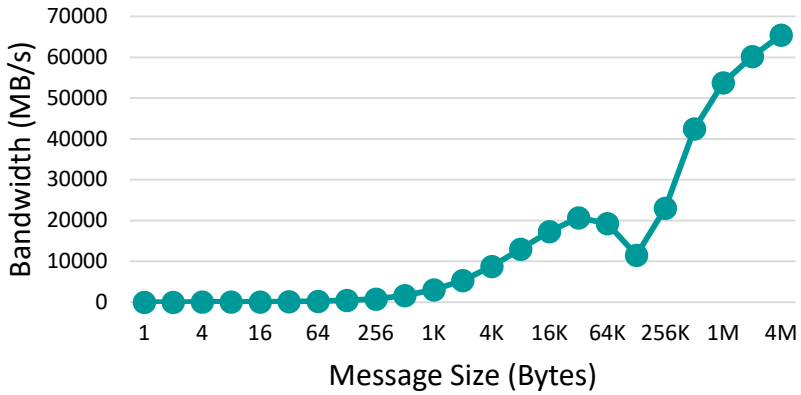
Intra-Node Latency (Small Messages)



Intra-Node Latency (Large Messages)



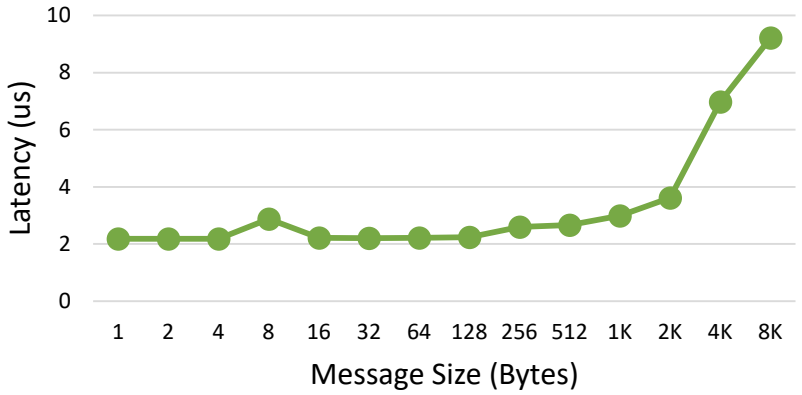
Intra-Node Bandwidth



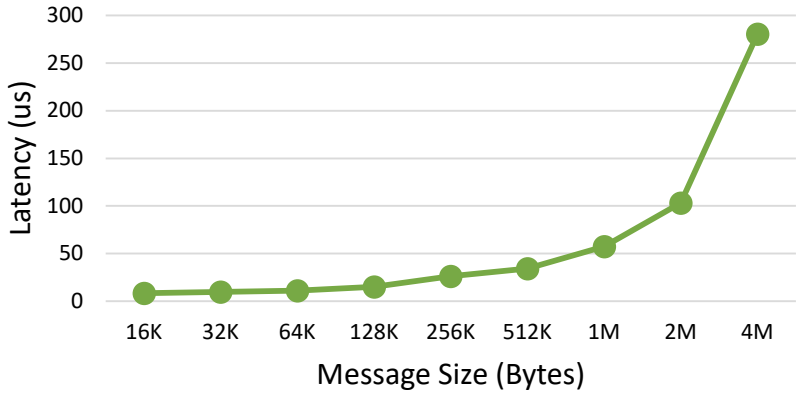
Intra-node Latency: 0.76 us (with GDRCopy)

Intra-node Bandwidth: 65.48 GB/sec for 4MB (via NVLINK2)

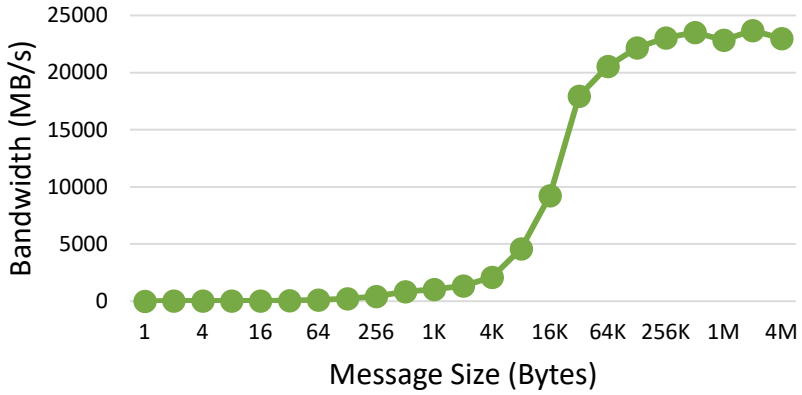
Inter-Node Latency (Small Messages)



Inter-Node Latency (Large Messages)



Inter-Node Bandwidth



Inter-node Latency: 2.18 us (with GDRCopy 2.0)

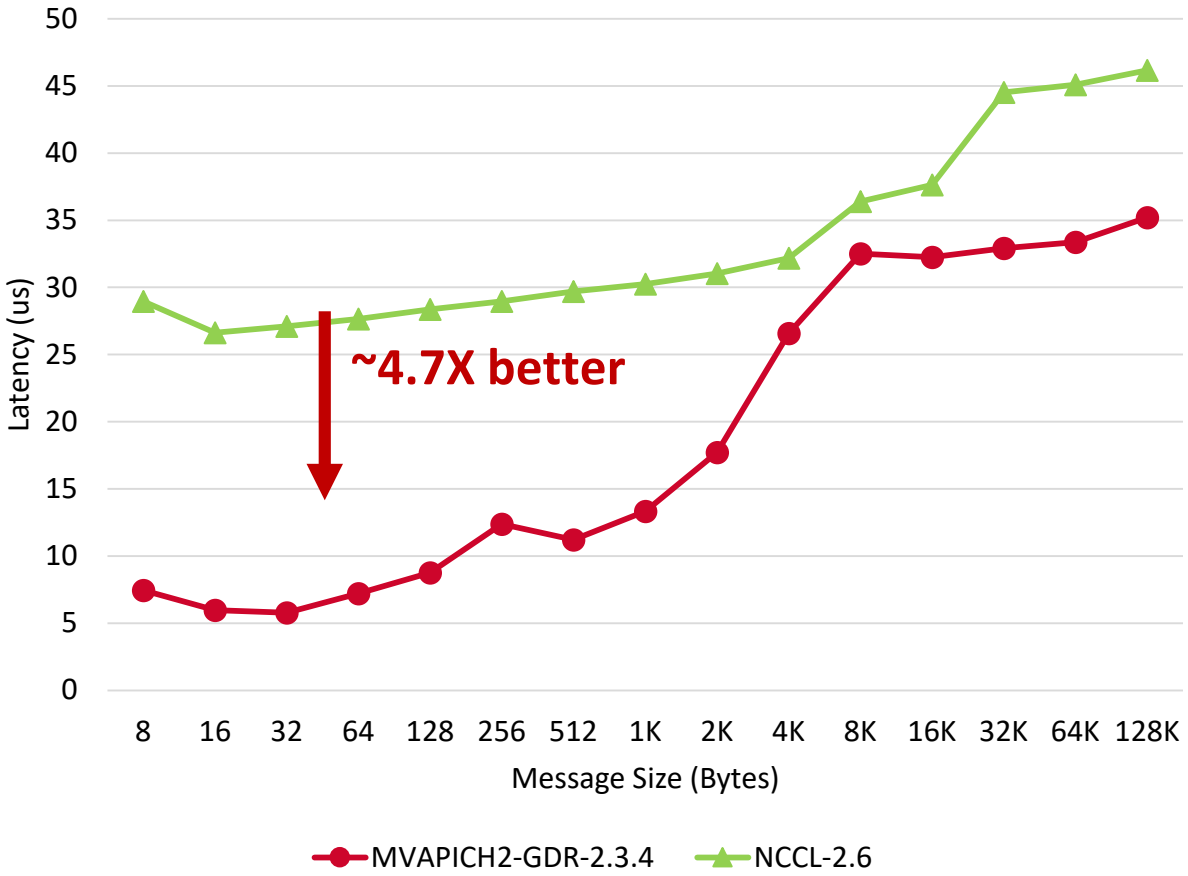
Inter-node Bandwidth: 23 GB/sec for 4MB (via 2 Port EDR)

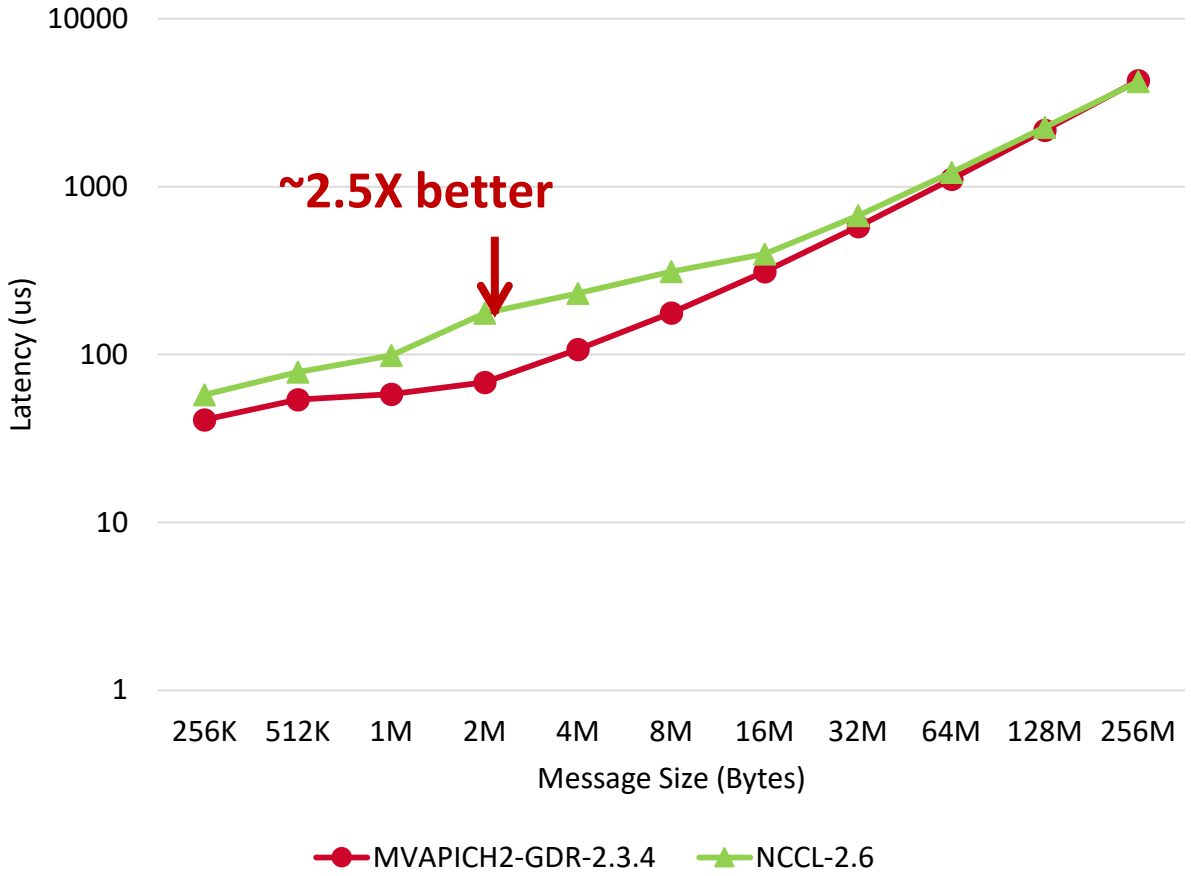
Platform: OpenPOWER (POWER9-ppc64le) nodes equipped with a dual-socket CPU, 4 Volta V100 GPUs, and 2port EDR InfiniBand Interconnect

MVAPICH2-GDR vs. NCCL2 – Allreduce Operation (DGX-2)

- Optimized designs in MVAPICH2-GDR offer better/comparable performance for most cases
- MPI_Allreduce (MVAPICH2-GDR) vs. ncclAllreduce (NCCL2) on 1 DGX-2 node (16 Volta GPUs)

Platform: Nvidia DGX-2 system (16 Nvidia Volta GPUs connected with NVSwitch), CUDA 10.1



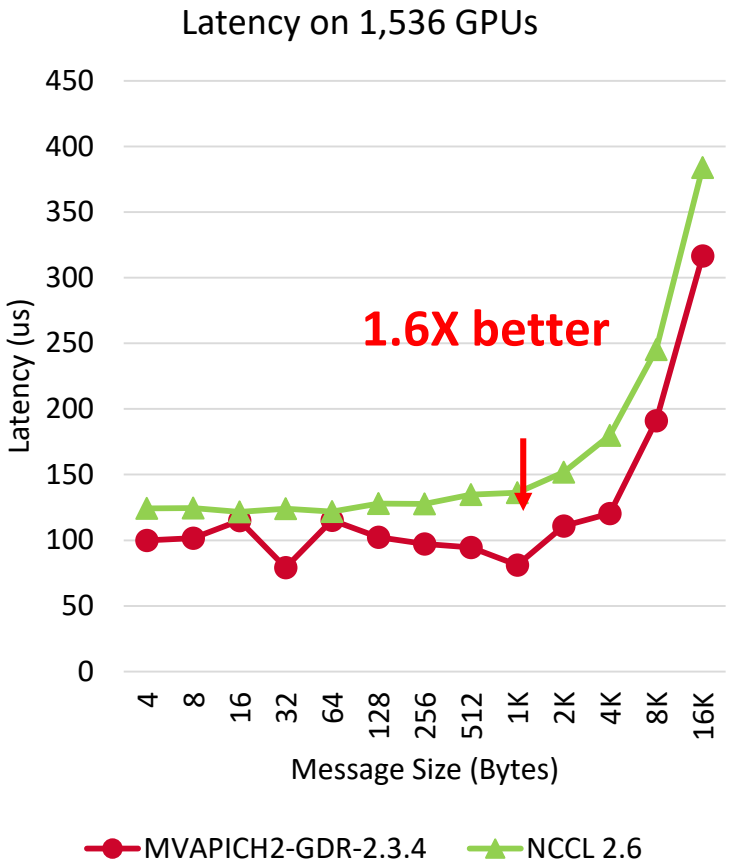


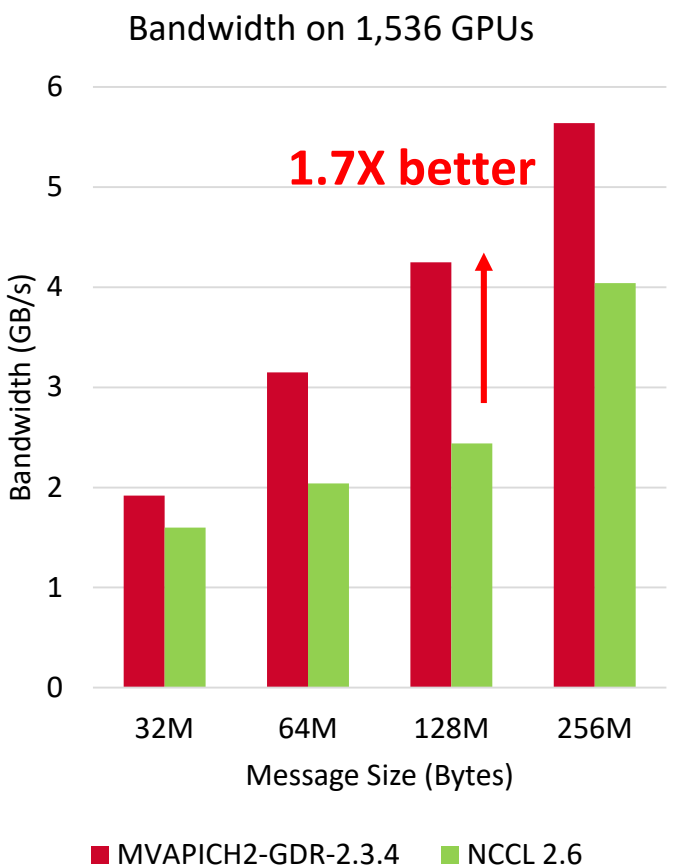
C.-H. Chu, P. Kousha, A. Awan, K. S. Khorassani, H. Subramoni and D. K. Panda, "NV-Group: Link-Efficient Reductions for Distributed Deep Learning on Modern Dense GPU Systems, " ICS-2020, June-July 2020.

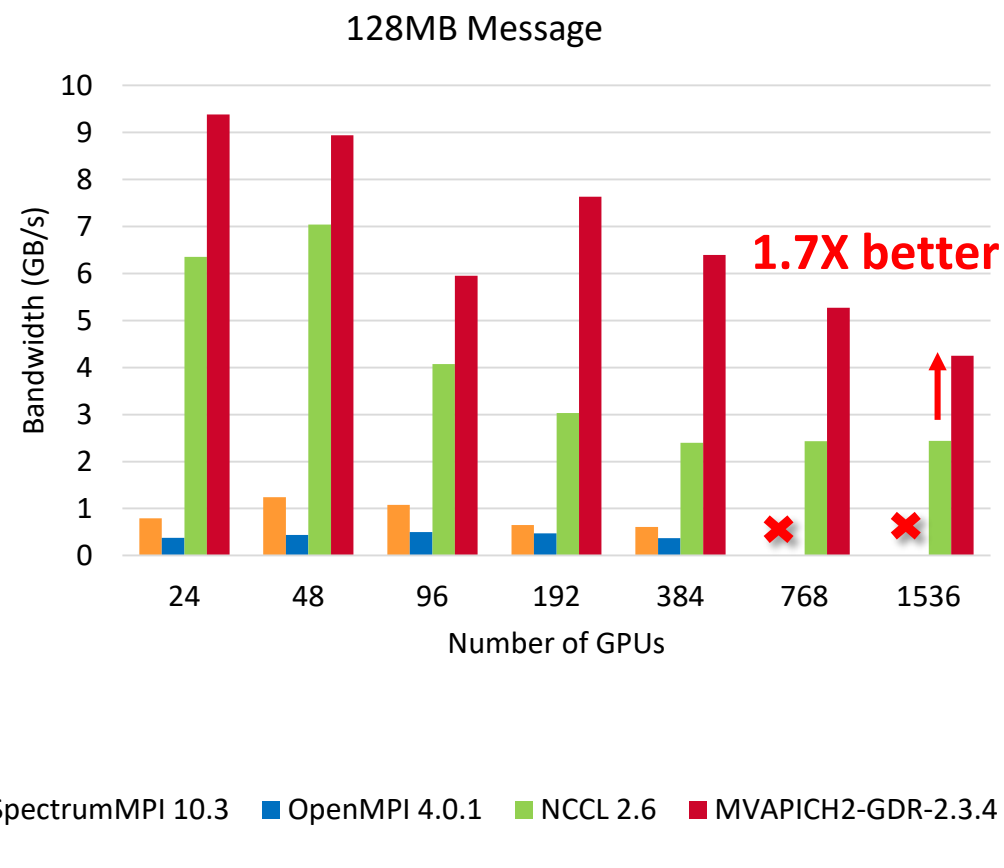
MVAPICH2-GDR: MPI_Allreduce at Scale (ORNL Summit)

- Optimized designs in 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



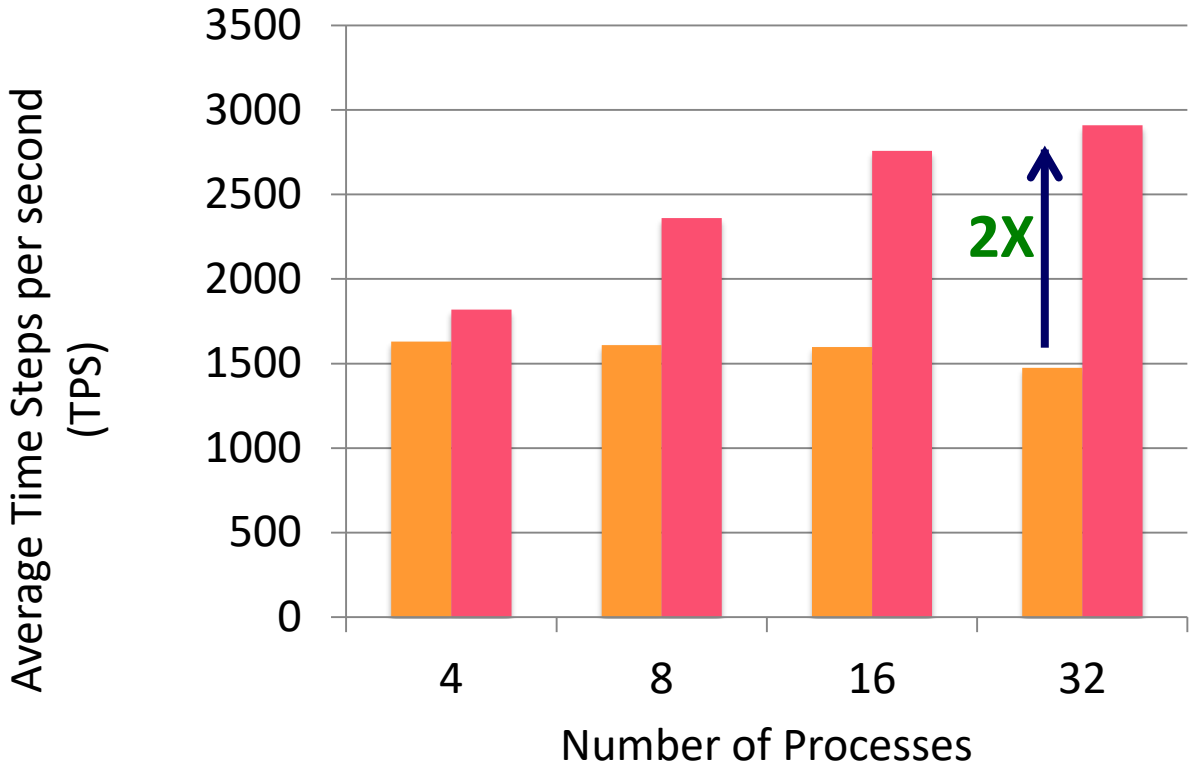




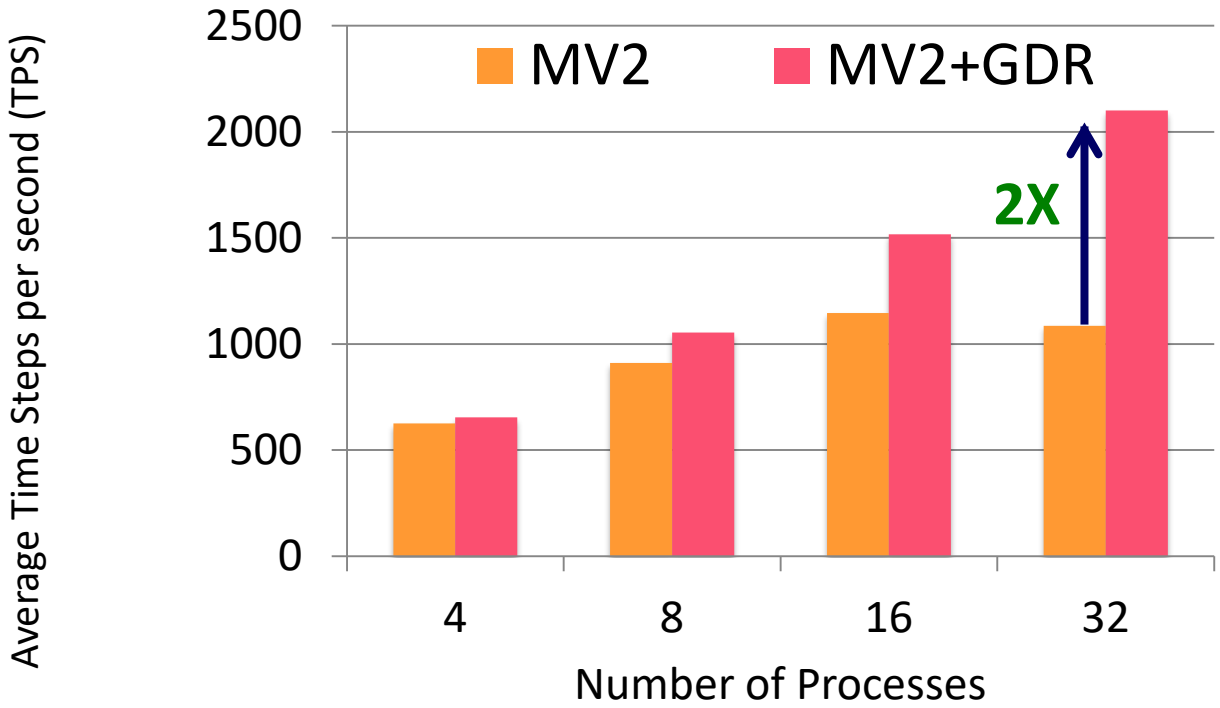
C.-H. Chu, P. Kousha, A. Awan, K. S. Khorassani, H. Subramoni and D. K. Panda, "NV-Group: Link-Efficient Reductions for Distributed Deep Learning on Modern Dense GPU Systems, " ICS-2020, June-July 2020.

Application-Level Evaluation (HOOMD-blue)

64K Particles



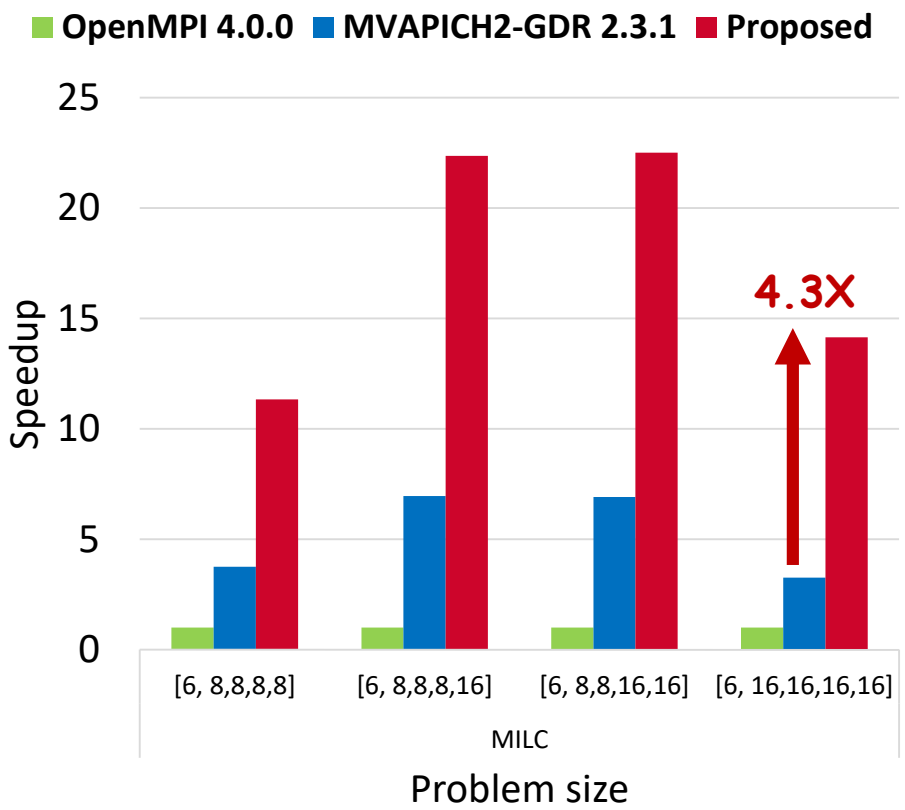
256K Particles



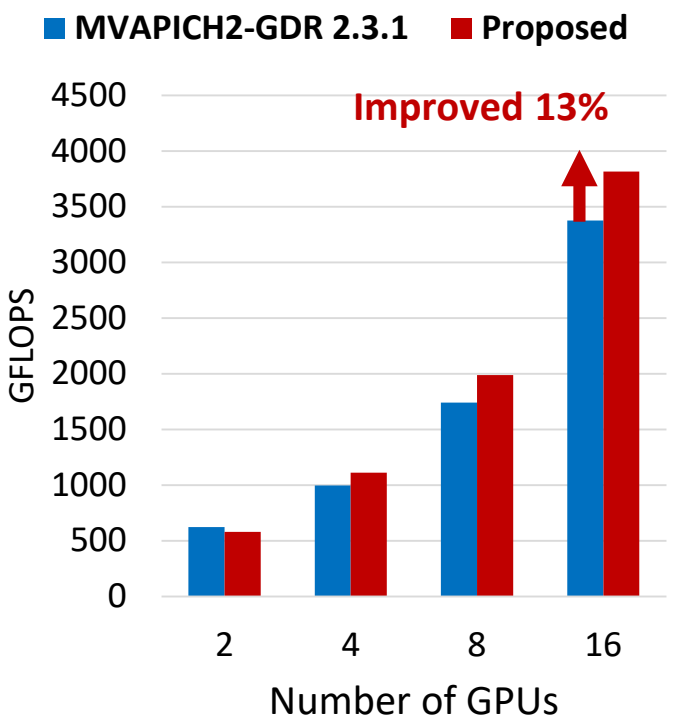
- Platform: Wilkes (Intel Ivy Bridge + NVIDIA Tesla K20c + Mellanox Connect-IB)
- HoomDBlue Version 1.0.5
  - GDRCOPY enabled: MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_0 MV2_IBA_EAGER_THRESHOLD=32768 MV2_VBUF_TOTAL_SIZE=32768 MV2_USE_GPUDIRECT_LOOPBACK_LIMIT=32768 MV2_USE_GPUDIRECT_GDRCOPY=1 MV2_USE_GPUDIRECT_GDRCOPY_LIMIT=16384

Performance Impacts of Kernel-Based Datatypes on Application Kernels

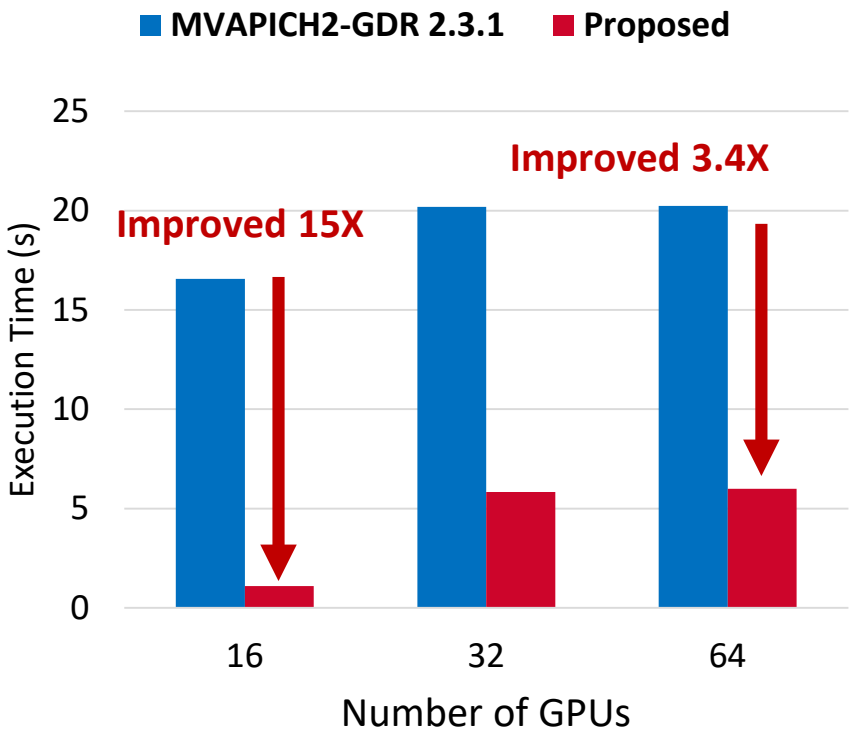
MILC communication kernel



Jacobi – 2DStencil Computation



COSMO Model



NVLink Platform: Nvidia DGX-2 system

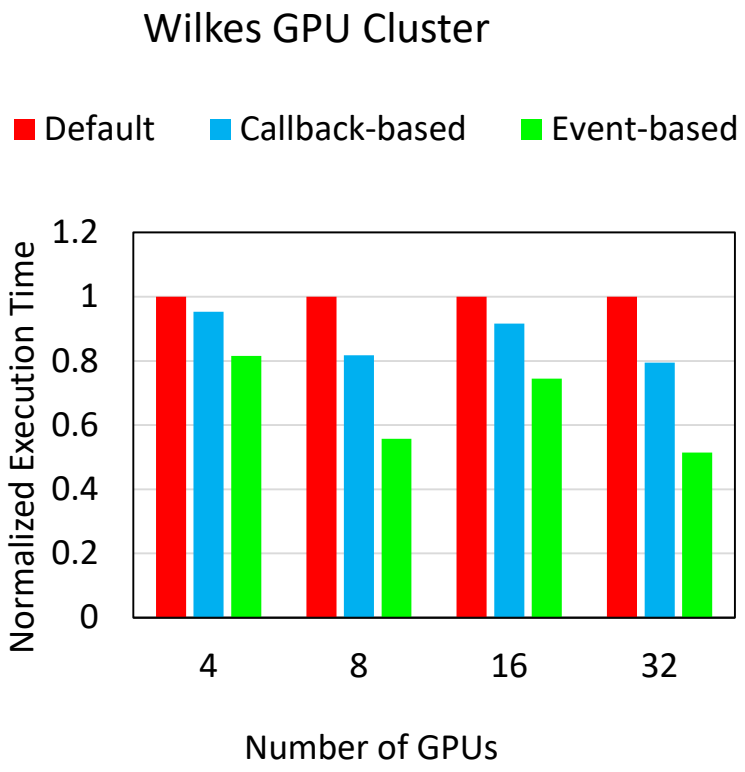
PCIe Platform: Cray CS-Storm

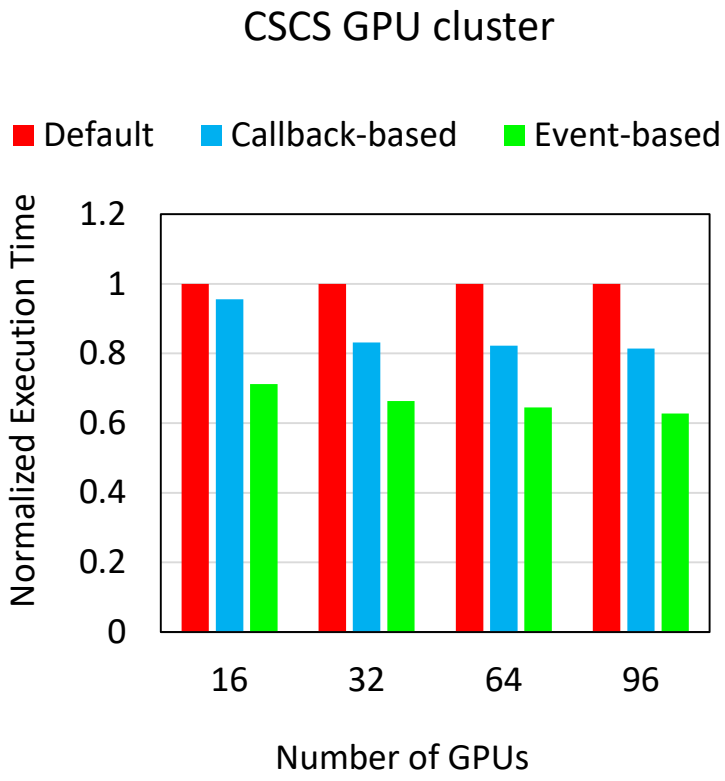
- **Performance improvements for various HPC applications on Dense GPU systems**

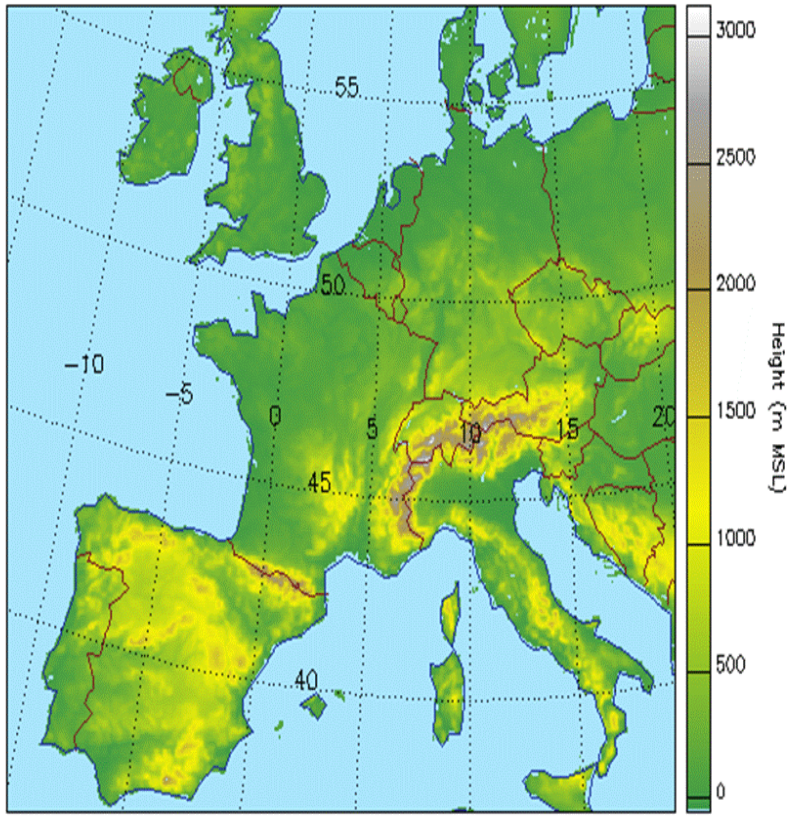
C.-H. Chu, J. M. Hashmi, K. S. Khorassani, H. Subramoni, and D. K. Panda,
"High-Performance Adaptive MPI Derived Datatype Communication for Modern Multi-GPU Systems" HiPC 2019.

Available in MVAPICH2-GDR 2.3.4

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







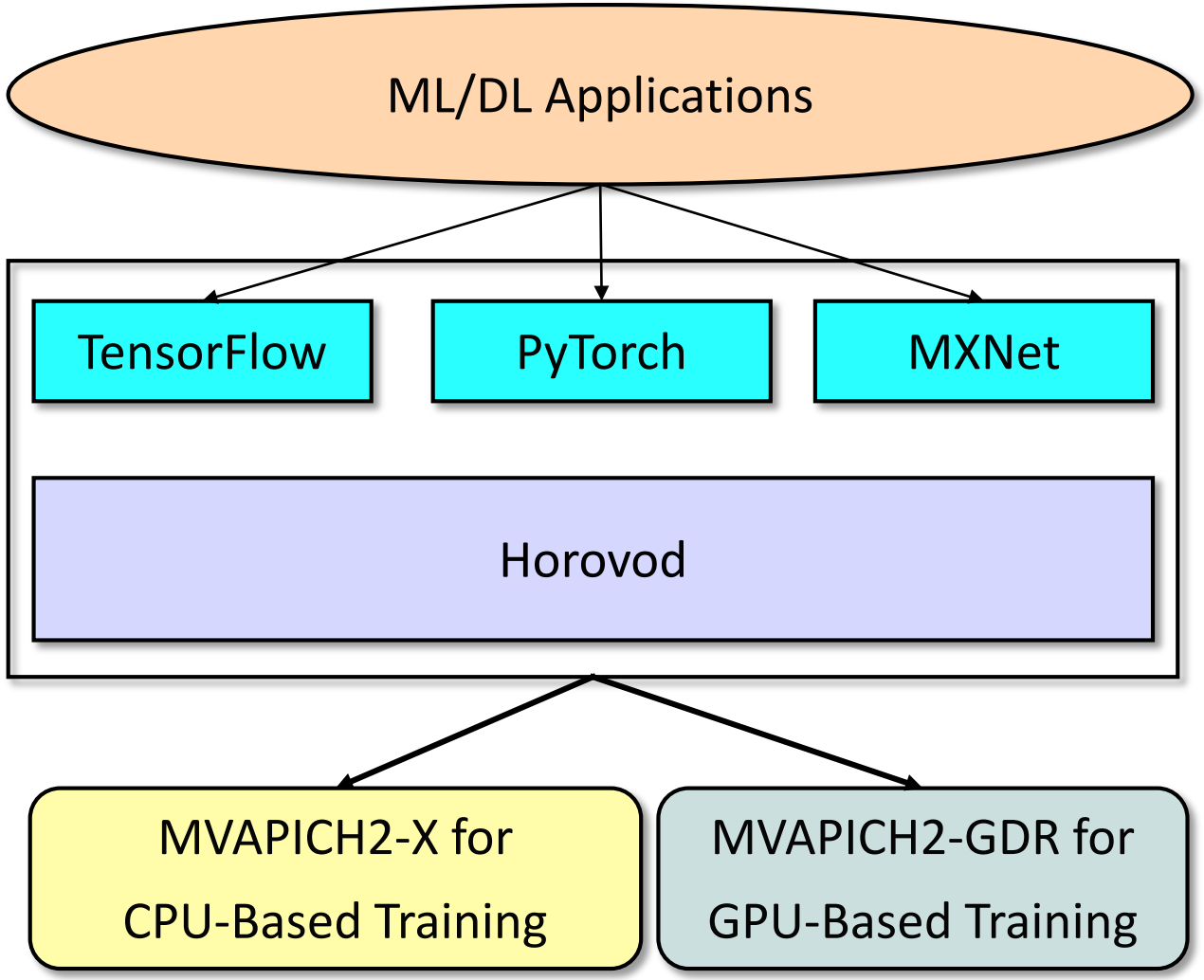
- 2X improvement on 32 GPUs nodes
- 30% improvement on 96 GPU nodes (8 GPUs/node)

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

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

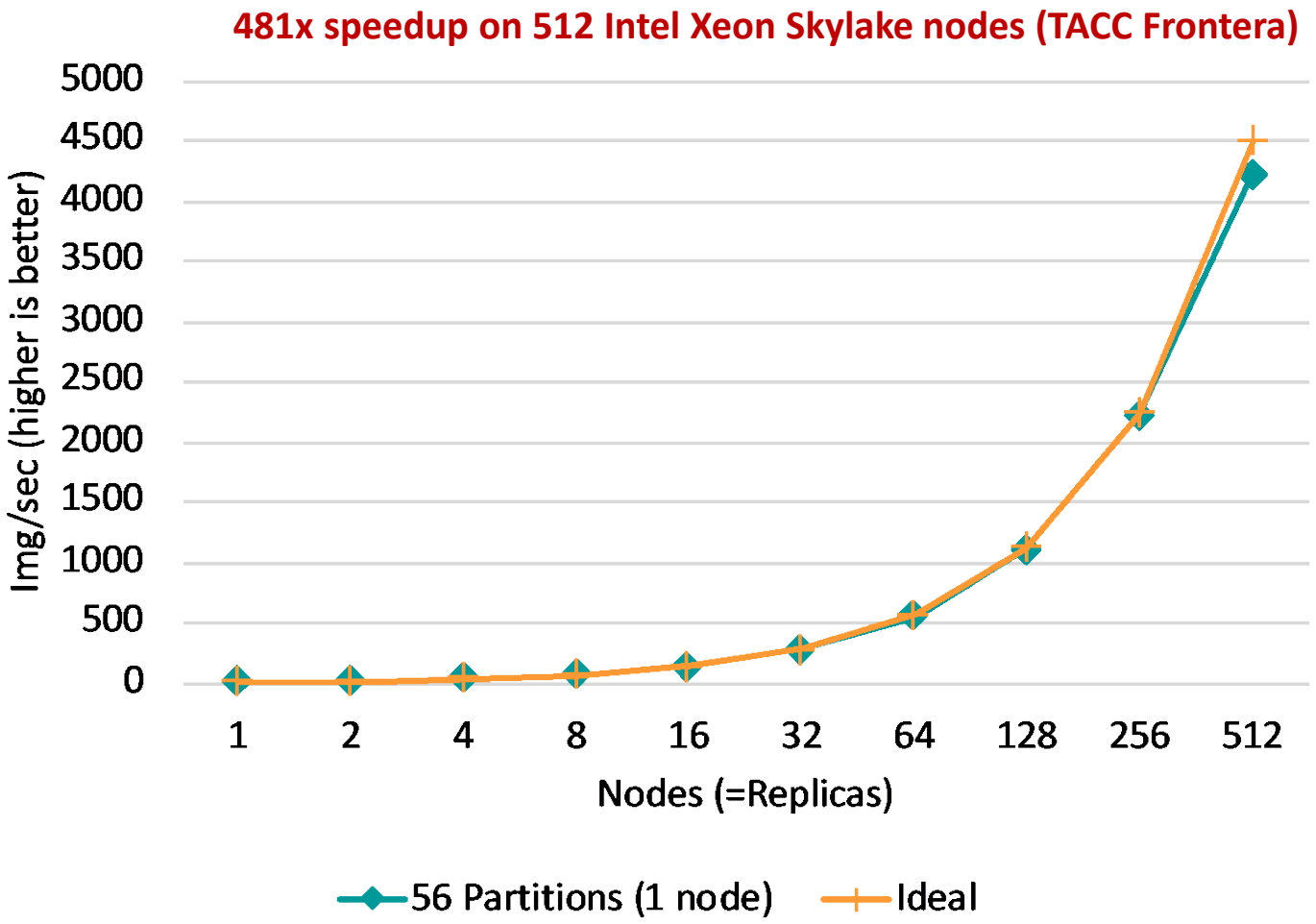
MVAPICH2 (MPI)-driven Infrastructure for ML/DL Training



More details available from:
http://hidl.cse.ohio-state.edu

Out-of-core Training with HyPar-Flow (512 nodes on TACC Frontera)

- ResNet-1001 with variable batch size
- Approach:
  - 48 model-partitions for 56 cores
  - 512 model-replicas for 512 nodes
  - Total cores: 48 x 512 = 24,576
- Speedup
  - **253X** on 256 nodes
  - **481X** on 512 nodes
- Scaling Efficiency
  - **98%** up to 256 nodes
  - **93.9%** for 512 nodes

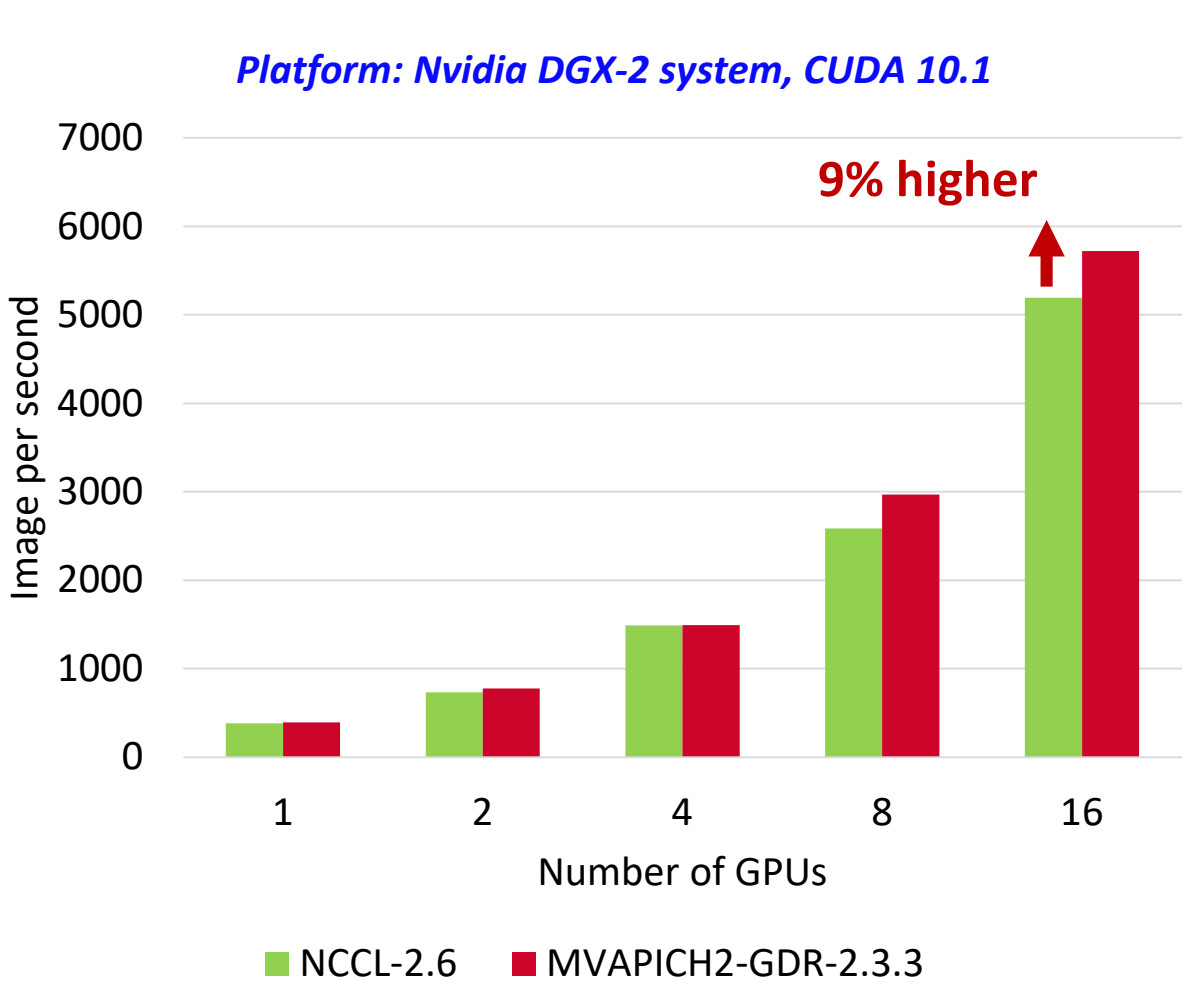


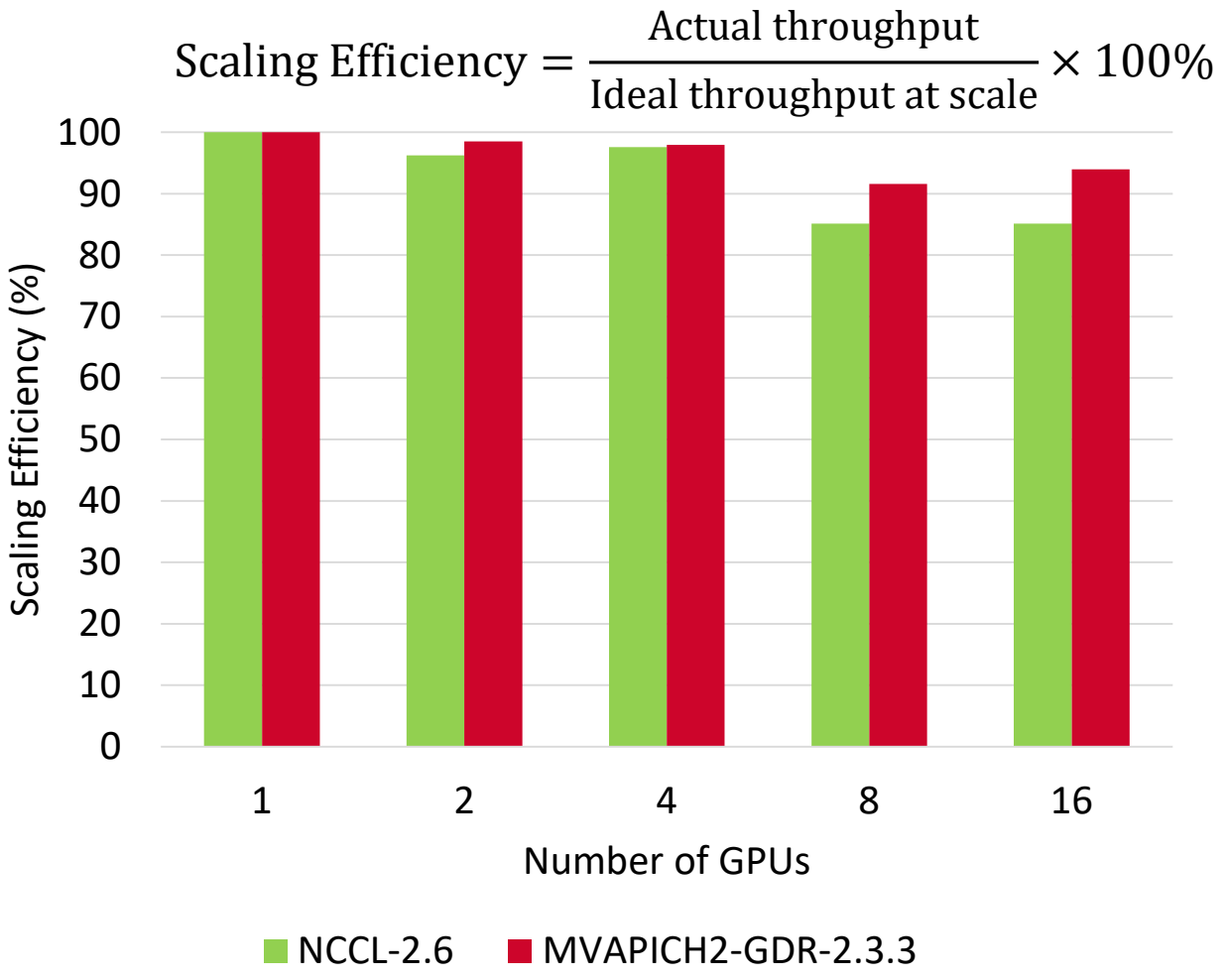
A. A. Awan, A. Jain, Q. Anthony, H. Subramoni, and DK Panda, "HyPar-Flow: Exploiting MPI and Keras for Hybrid Parallel Training of TensorFlow models", ISC '20, https://arxiv.org/pdf/1911.05146.pdf

More details in Jain's Talk (later Today)

Scalable TensorFlow using Horovod and MVAPICH2-GDR

- **ResNet-50 Training using TensorFlow benchmark on 1 DGX-2 node (16 Volta GPUs)**

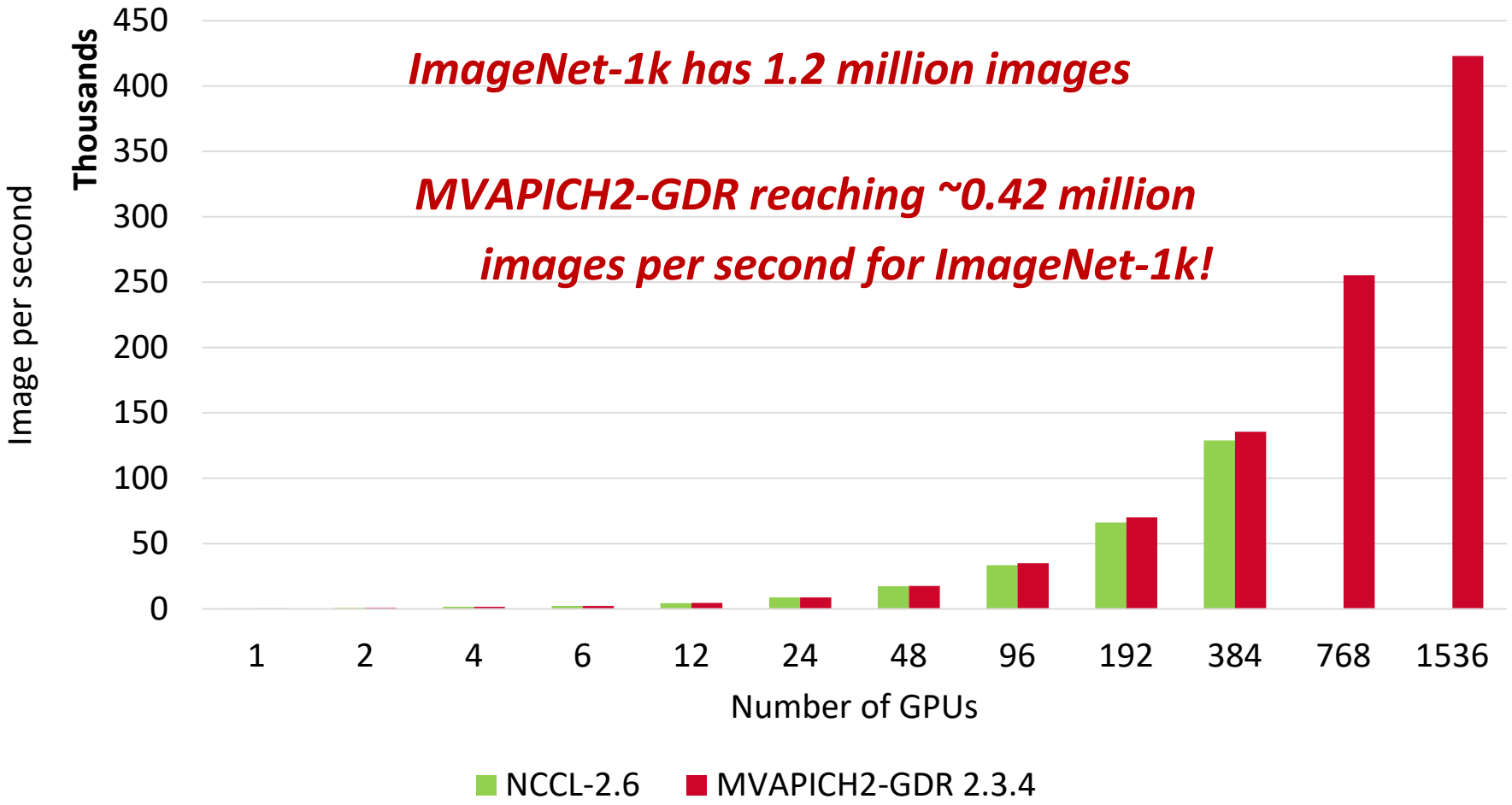




C.-H. Chu, P. Kousha, A. Awan, K. S. Khorassani, H. Subramoni and D. K. Panda, "NV-Group: Link-Efficient Reductions for Distributed Deep Learning on Modern Dense GPU Systems, " ICS-2020, June-July 2020.

Distributed TensorFlow on ORNL Summit (1,536 GPUs)

- ResNet-50 Training using TensorFlow benchmark on SUMMIT -- 1536 Volta GPUs!
- 1,281,167 (1.2 mil.) images
- Time/epoch = 3 seconds
- Total Time (90 epochs) = 3 x 90 = 270 seconds = **4.5 minutes!**



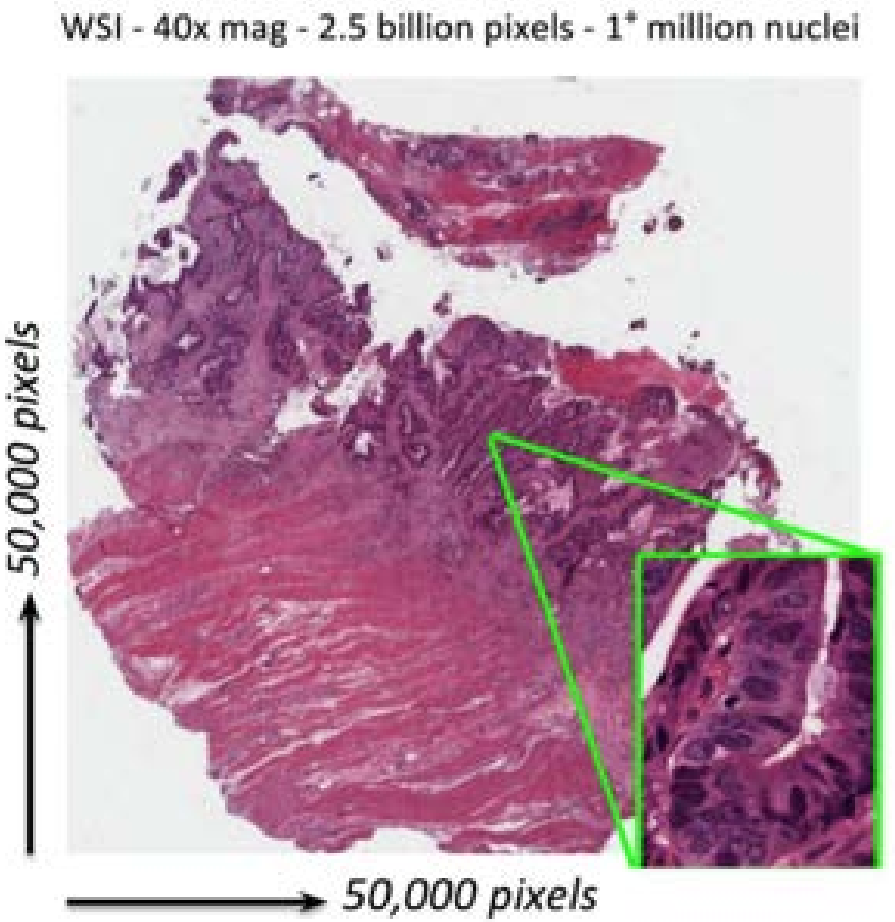
*We observed issues for NCCL2 beyond 384 GPUs

Platform: The Summit Supercomputer (#2 on Top500.org) – 6 NVIDIA Volta GPUs per node connected with NVLink, CUDA 10.1

Exploiting Model Parallelism in AI-Driven Digital Pathology

- Pathology whole slide image (WSI)
  - Each WSI = 100,000 x 100,000 pixels
  - Can not fit in a single GPU memory
  - Tiles are extracted to make training possible
- Two main problems with tiles
  - Restricted tile size because of GPU memory limitation
  - Smaller tiles loose structural information
- Can we use Model Parallelism to train on larger tiles to get better accuracy and diagnosis?
- Reduced training time significantly on OpenPOWER + NVIDIA V100 GPUs
  - **32 hours (1 node, 1 GPU) -> 7.25 hours (1 node, 4 GPUs) -> 27 mins (32 nodes, 128 GPUs)**

More details in Jain's Talk (Tomorrow)



Courtesy: https://blog.kitware.com/digital-slide-archive-large-image-and-histomicstk-open-source-informatics-tools-for-management-visualization-and-analysis-of-digital-histopathology-data/

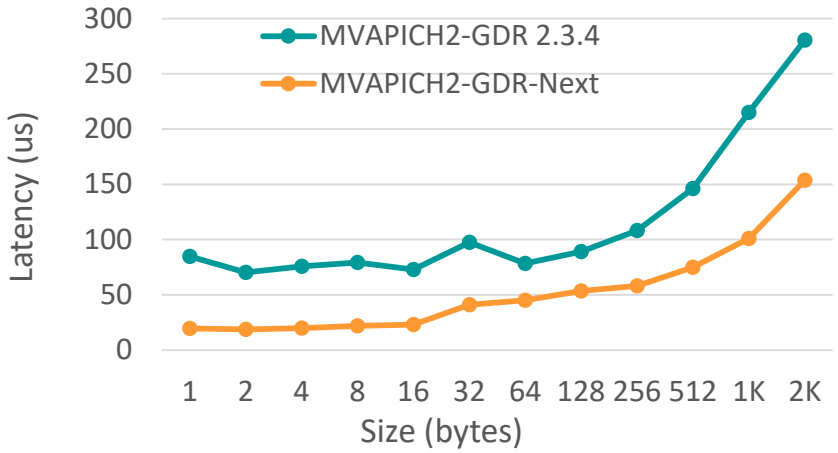
A. Jain, A. Awan, A. Aljuhani, J. Hashmi, Q. Anthony, H. Subramoni, D. K. Panda, R. Machiraju, and A. Parwani, "GEMS: GPU Enabled Memory Aware Model Parallelism System for Distributed DNN Training", Supercomputing (SC '20), Accepted to be Presented.

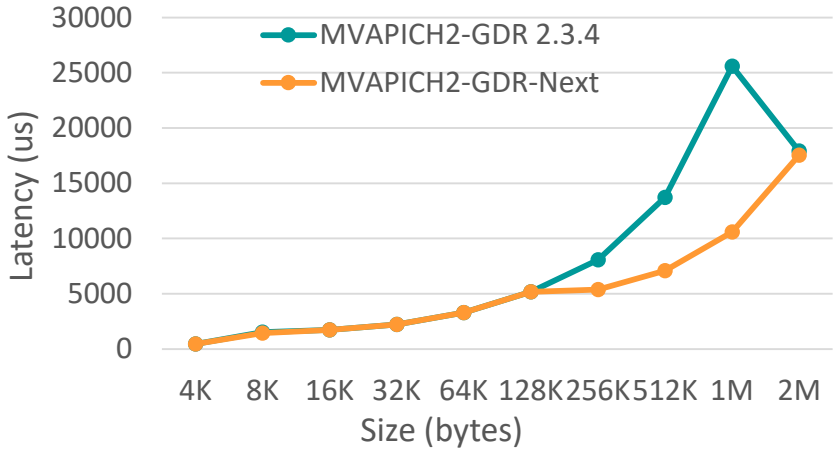
MVAPICH2-GDR Upcoming Features for HPC and DL

- Optimized Alltoall and Allgather Collectives
- Support for AMD GPU
- Optimization for Distributed PyTorch, Horovod, and DeepSpeed
- On-the-fly Message Compression
- MVAPICH2-GDR Support for Dask-based applications using mpi4Dask
- MVAPICH2-GDR Support for cuML applications using mpi4Py

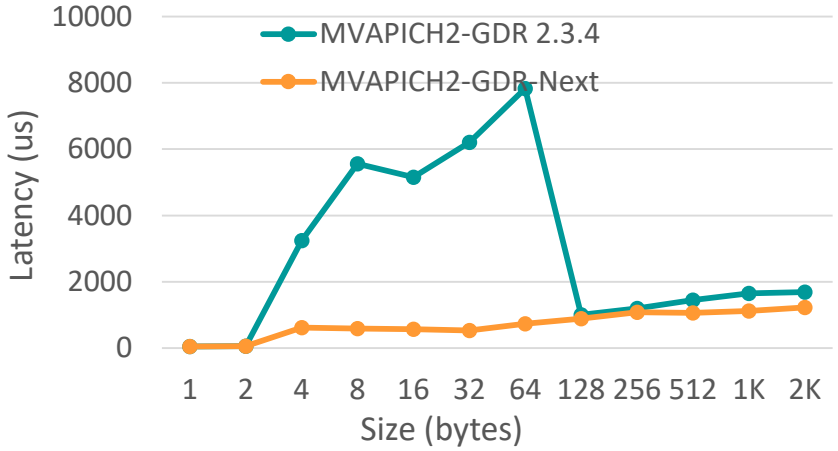
Optimized Alltoall and Allgather Performance with MVAPICH2-GDR

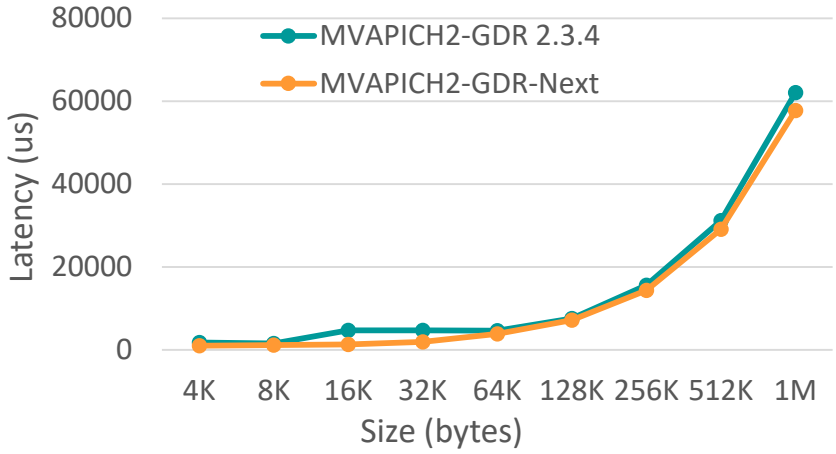
MPI_Allgather - 32 Nodes, 128 GPUs





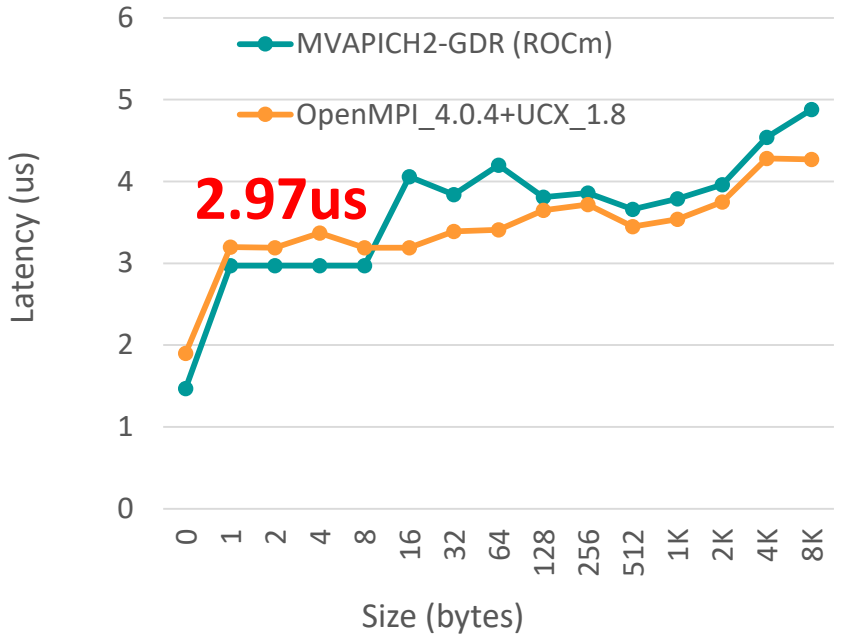
MPI_Alltoall - 32 Nodes, 128 GPUs

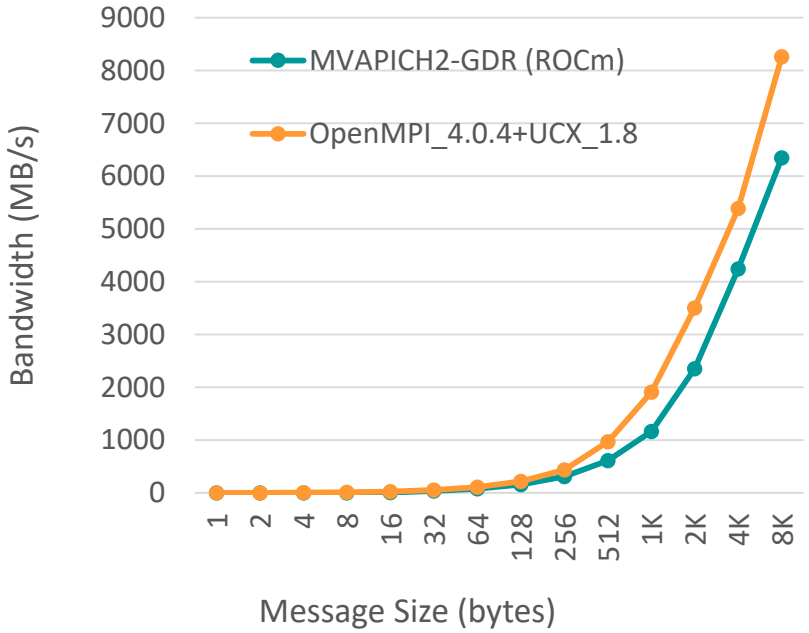


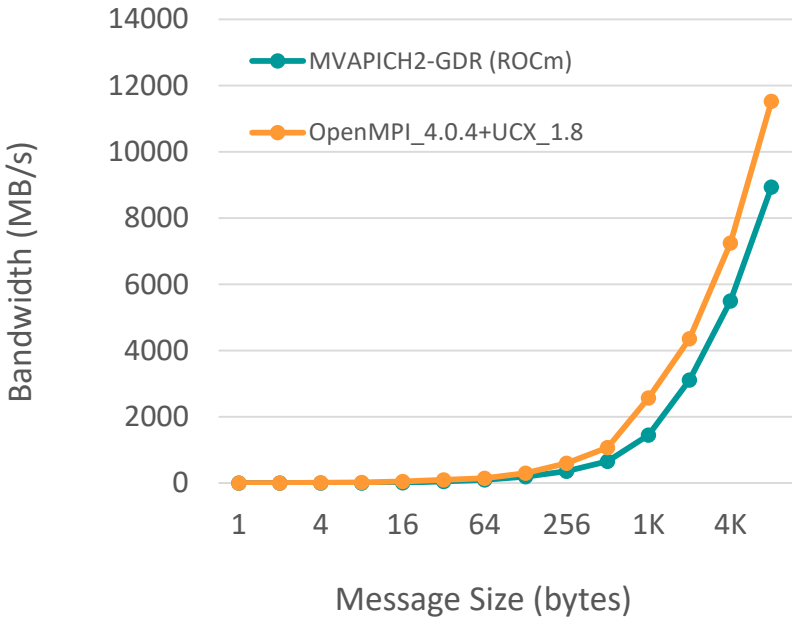


Platform: OpenPOWER (POWER9-ppc64le) nodes equipped with a dual-socket CPU, 4 Volta V100 GPUs, and 2port EDR InfiniBand Interconnect

Inter-node Pt-to-Pt Performance with AMD GPU (Device-to-Device)







More details in Hashmi's Talk (Tomorrow)

- Inter-node (1x Mi50 per node) latency over HDR interconnect
- Latency, Bandwidth, and Bi-directional Bandwidth comparison
- Exploit ROCm PeerDirect (similar to NVIDIA's GPUDirect) feature for short messages

PyTorch, Horovod and DeepSpeed at Scale: Training ResNet-50 on 256 V100 GPUs

- Training performance for 256 V100 GPUs on LLNL Lassen
  - **~10,000 Images/sec faster** than NCCL training!

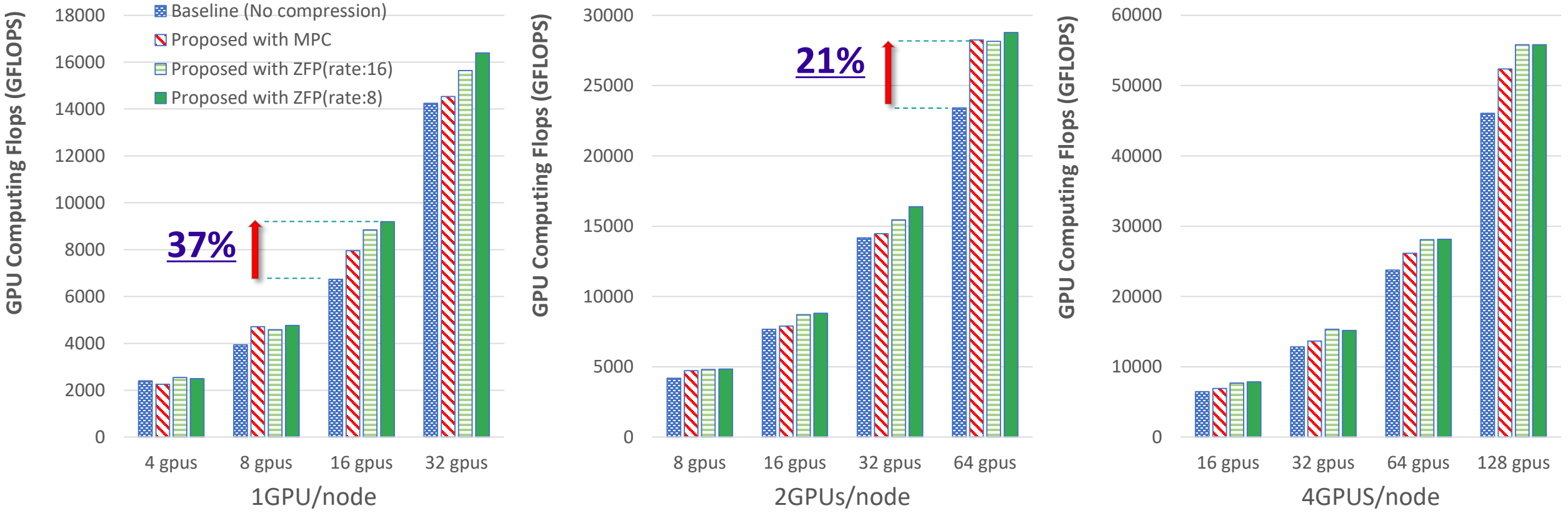
Distributed Framework	Torch.distributed		Horovod		DeepSpeed	
Images/sec on 256 GPUs	61,794	**72,120**	74,063	**84,659**	80,217	**88,873**
Communication Backend	NCCL	MVAPICH2-GDR	NCCL	MVAPICH2-GDR	NCCL	MVAPICH2-GDR

More details in Anthony's Talk (Tomorrow)

Scaling AWP-ODC (Seismic Simulation) with On-the-fly Message Compression

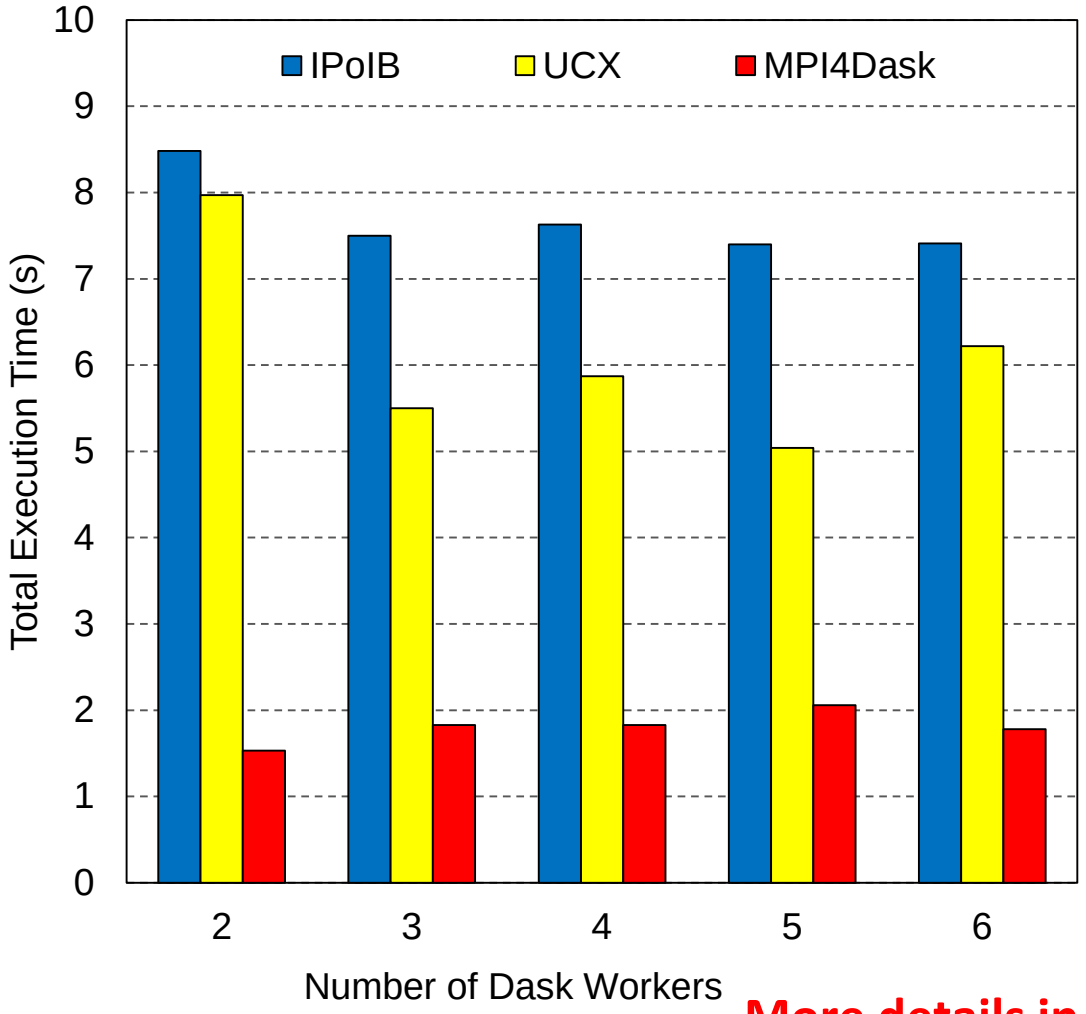
- Weak-Scaling of AWP-ODC on Longhorn-V100
- MPC achieved up to 21% improvement
- ZFP achieved up to 37% improvement with rate=8 (compression ratio=4)

More details in Zhou's Talk (Tomorrow)

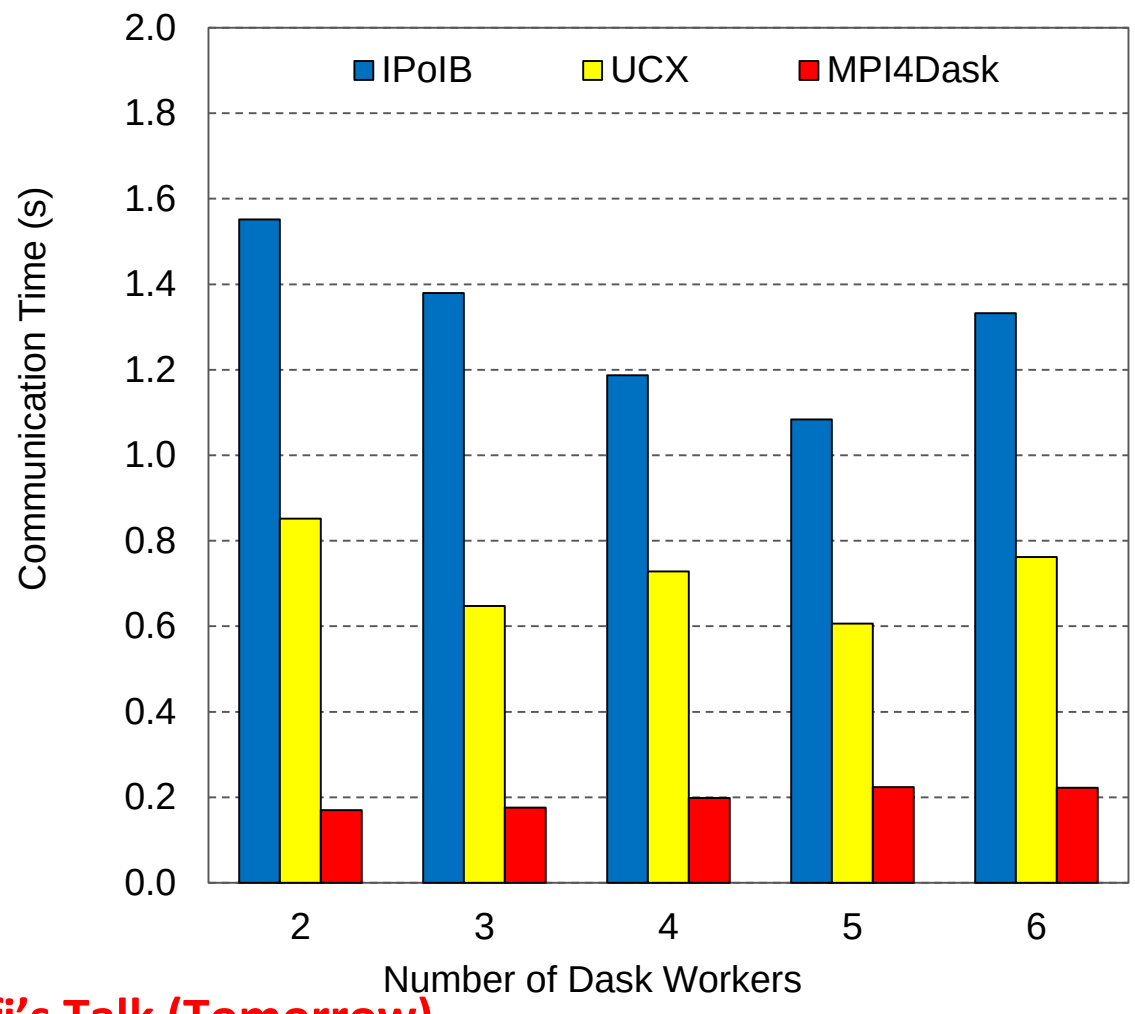


MPI4Dask - Accelerating Dask (Benchmark #1: Sum of cuPy Array and its Transpose)

3.47x better on average



6.92x better on average

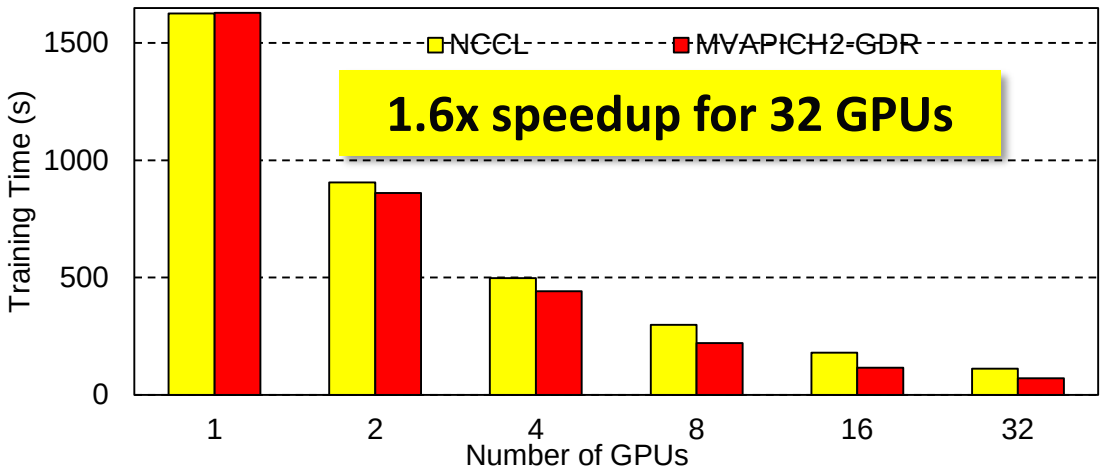


More details in Shafi's Talk (Tomorrow)

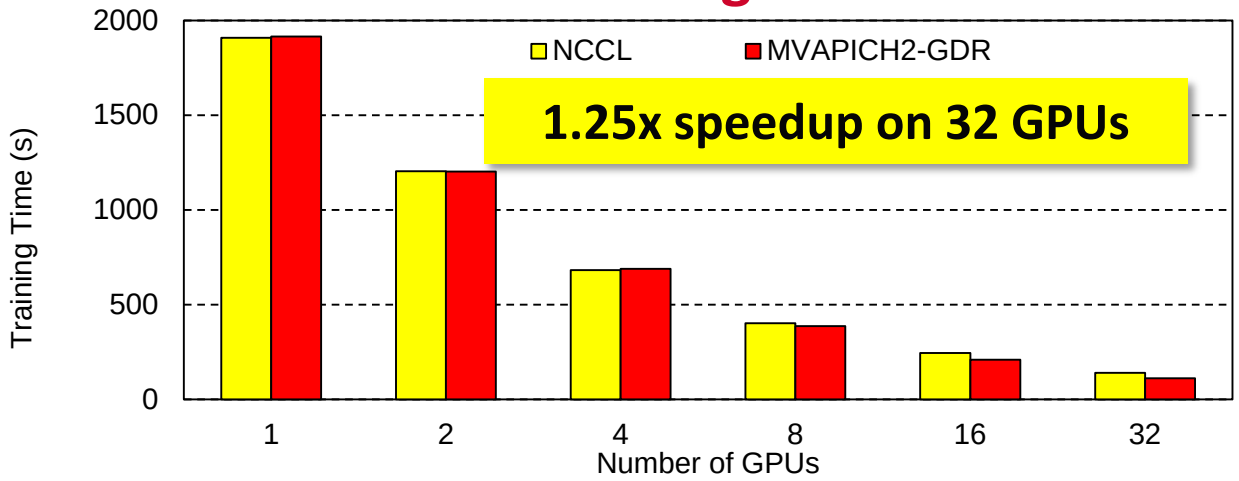
Accelerating Machine Learning Applications with mpi4Py and MVAPICH2-GDR

More details in Shafi's Talk (Tomorrow)

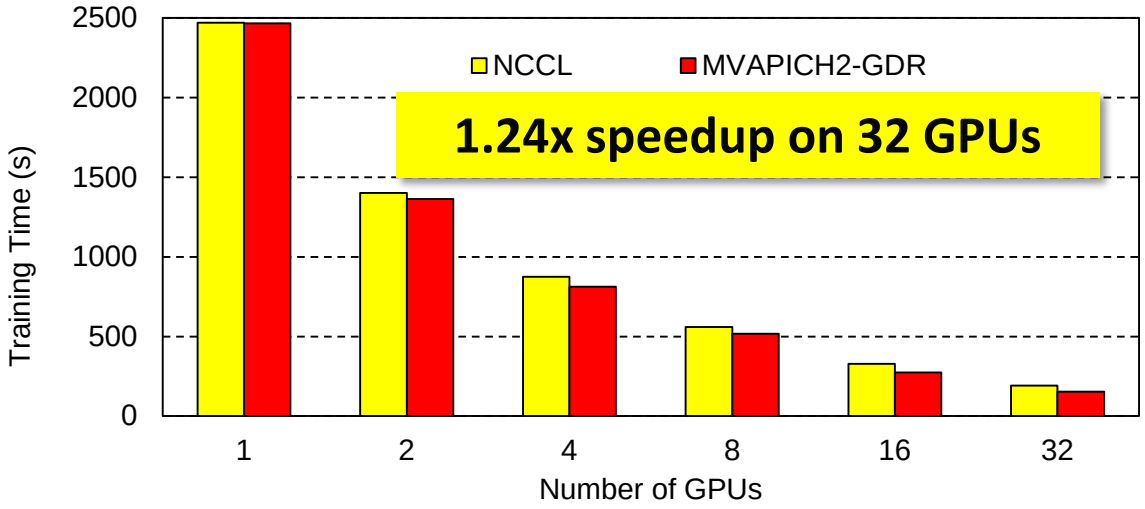
K-Means



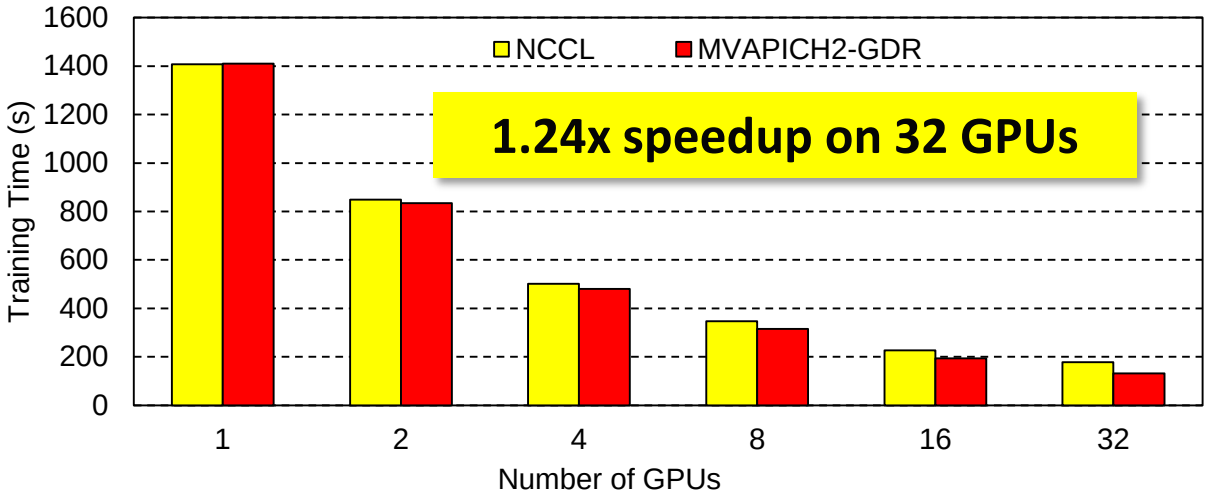
Linear Regression



Nearest Neighbors



Truncated SVD



MVAPICH2 Software Family

Requirements	Library
MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2
Advanced MPI features/support (UMR, ODP, DC, Core-Direct, SHARP, XPMEM), OSU INAM (InfiniBand Network Monitoring and Analysis), MPI+PGAS	MVAPICH2-X
Optimized Support of MVAPICH2-X for Microsoft Azure Platform with InfiniBand	MVAPICH2-X-Azure
Advanced MPI features (SRD and XPMEM) with support for Amazon Elastic Fabric Adapter (EFA)	MVAPICH2-X-AWS
Optimized MPI for clusters with NVIDIA GPUs and for GPU-enabled Deep Learning Applications	MVAPICH2-GDR
Energy-aware MPI with Support for InfiniBand, Omni-Path, Ethernet/iWARP and, RoCE (v1/v2)	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

OSU Microbenchmarks

- Available since 2004
- Suite of microbenchmarks to study communication performance of various programming models
- Benchmarks available for the following programming models
  - Message Passing Interface (MPI)
  - Partitioned Global Address Space (PGAS)
    - Unified Parallel C (UPC), Unified Parallel C++ (UPC++), and OpenSHMEM
- Benchmarks available for multiple accelerator based architectures
  - Compute Unified Device Architecture (CUDA)
  - OpenACC Application Program Interface
- Part of various national resource procurement suites like NERSC-8 / Trinity Benchmarks
- Continuing to add support for newer primitives and features
- ROCm Support for AMD GPUs will be available in upcoming release
- Please visit the following link for more information
  - http://mvapich.cse.ohio-state.edu/benchmarks/

Applications-Level Tuning: Compilation of Best Practices

- MPI runtime has many parameters
- Tuning a set of parameters can help you to extract higher performance
- Compiled a list of such contributions through the MVAPICH Website
  - http://mvapich.cse.ohio-state.edu/best_practices/
- Initial list of applications
  - Amber
  - HoomDBlue
  - HPCG
  - Lulesh
  - MILC
  - Neuron
  - SMG2000
  - Cloverleaf
  - SPEC (LAMMPS, POP2, TERA_TF, WRF2)
- Soliciting additional contributions, send your results to mvapich-help at cse.ohio-state.edu.
- We will link these results with credits to you.

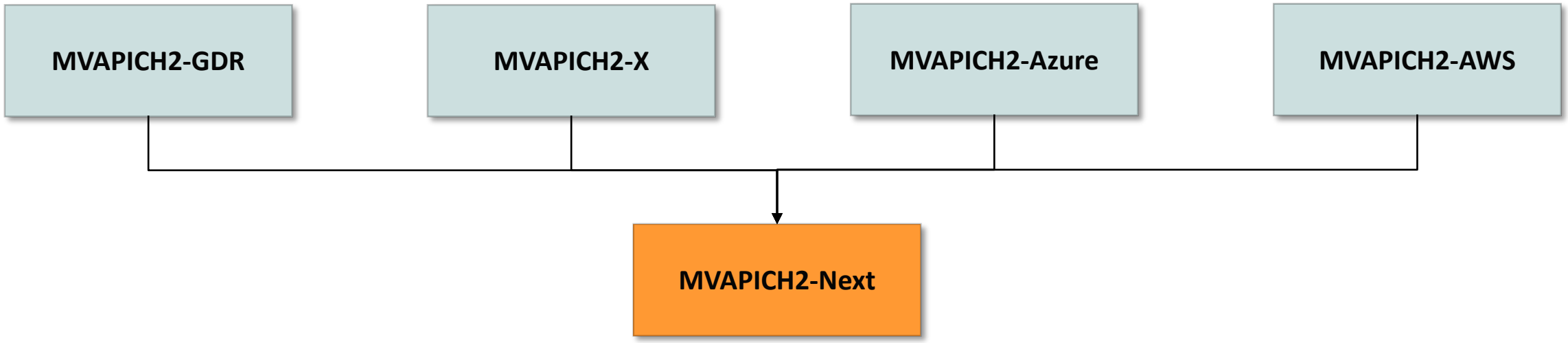
MVAPICH2 Libraries available through Spack

- Released on 08/23/20
- Easy Installation of MVAPICH2 Libraries through Spack
  - MVAPICH2
  - MVAPICH2-X
  - MVAPICH2-GDR
- Detailed Spack-based Installation User Guide is available:
  http://mvapich.cse.ohio-state.edu/userguide/userguide_spack/

Enabling Auto detection of Underlying features with dlopen

- Use dlopen to identify support available in underlying system
  - Removes dependencies on libraries like
    - ibverbs
    - ibmad
    - ibumad
    - rdmacm
    - xpmem
  - Simplifies installation and deployment
    - No need for multiple configure options
    - One binary can have all necessary support

MVAPICH2-Next: One Stack to Conquer all Architectures and Interconnects



- Various libraries will be merged into one release
- Simplifies installation and deployment
- Enables utilizing the best advanced features for all architectures

MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 1-10M cores
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - MPI + Task*
- Enhanced Optimization for GPU Support and Accelerators
- Taking advantage of advanced features of Mellanox InfiniBand
  - Tag Matching*
  - Adapter Memory*
- Enhanced communication schemes for upcoming architectures
  - Intel Optane*
  - BlueField*
  - 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

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 being provided to Lawrence Livermore National Laboratory (LLNL) and KISTI, Korea

X-ScaleSolutions

Value-Added Products with Support

- Silver ISV member of the OpenPOWER Consortium
- 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 2019 OpenPOWER North America Summit)
  - X-ScaleHPC
  - X-ScaleAI

More details in Dai's Talk (Tomorrow)

X-ScaleSolutions

Funding Acknowledgments

Funding Support by

Equipment Support by

Acknowledgments to all the Heroes (Past/Current Students and Staffs)

Current Students (Graduate)

- Q. Anthony (Ph.D.)
- M. Bayatpour (Ph.D.)
- C.-H. Chu (Ph.D.)
- A. Jain (Ph.D.)
- M. Kedia (M.S.)
- K. S. Khorassani (Ph.D.)
- P. Kousha (Ph.D.)
- N. S. Kumar (M.S.)
- B. Ramesh (Ph.D.)
- K. K. Suresh (Ph.D.)
- N. Sarkauskas (Ph.D.)
- S. Srivastava (M.S.)
- S. Xu (Ph.D.)
- Q. Zhou (Ph.D.)

Current Research Scientists

- A. Shafi
- H. Subramoni

Current Senior Research Associate

- J. Hashmi

Current Software Engineer

- A. Reifsteck

Current Post-docs

- M. S. Ghazimeersaeed
- K. Manian

Current Research Specialist

- J. Smith

Past Students

- A. Awan (Ph.D.)
- 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.)
- J. Hashmi (Ph.D.)
- 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.)
- K. Raj (M.S.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- N. Sarkauskas (B.S.)
- 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 Post-Docs

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

Past Research Scientists

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

Past Programmers

- D. Bureddy
- J. Perkins

Past Research Specialist

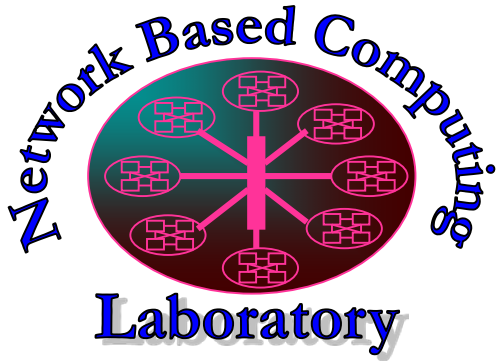
- M. Arnold

Multiple Positions Available in MVAPICH2, BigData and DL/ML Projects in my Group

- Looking for Bright and Enthusiastic Personnel to join as
  - PhD Students
  - Post-Doctoral Researchers
  - MPI Programmer/Software Engineer
  - Hadoop/Spark/Big Data Programmer/Software Engineer
  - Deep Learning and Cloud Programmer/Software Engineer
- If interested, please send an e-mail to panda@cse.ohio-state.edu

Thank You!

panda@cse.ohio-state.edu



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

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/
Follow us on Twitter: @mvapich

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

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