MVAPICH Performance on Arm at Scale

Arm HPC User Group Talk (SC '19)

by

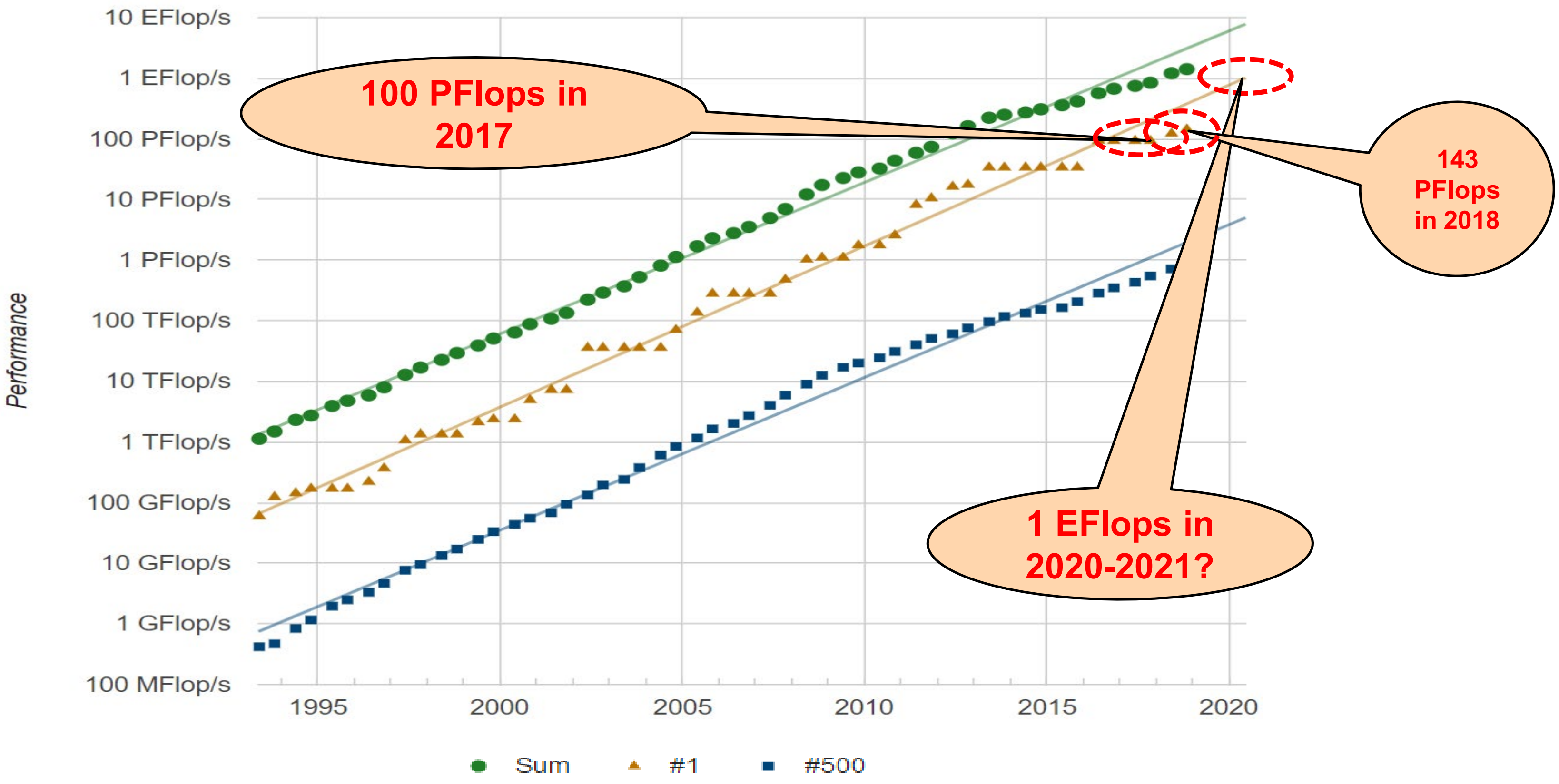
Dhabaleswar K. (DK) Panda

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



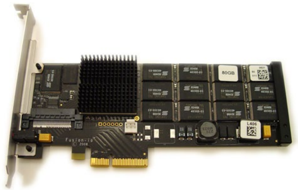
Expected to have an ExaFlop system in 2020-2021!

Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand
<1usec latency, 200Gbps Bandwidth>

Accelerators / Coprocessors
high compute density, high performance/watt
>1 TFlop DP on a chip



SSD, NVMe-SSD, NVRAM

- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)
- Available on HPC Clouds, e.g., Amazon EC2, NSF Chameleon, Microsoft Azure, etc.

Summit

Sierra

Sunway TaihuLight

K - Computer

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

Application Kernels/Applications (HPC and DL)

Middleware

Programming Models

MPI, PGAS (UPC, Global Arrays, OpenSHMEM), CUDA, OpenMP, OpenACC, Cilk, Hadoop (MapReduce), Spark (RDD, DAG), etc.

Communication Library or Runtime for Programming Models

- Point-to-point Communication
- Collective Communication
- Energy-Awareness
- Synchronization and Locks
- I/O and File Systems
- Fault Tolerance

Networking Technologies (InfiniBand, 40/100/200GigE, Slingshot, and Omni-Path)

Multi-/Many-core Architectures

Accelerators (GPU and FPGA)

Co-Design Opportunities and Challenges across Various Layers

- Performance
- Scalability
- Resilience

Designing (MPI+X) at Exascale

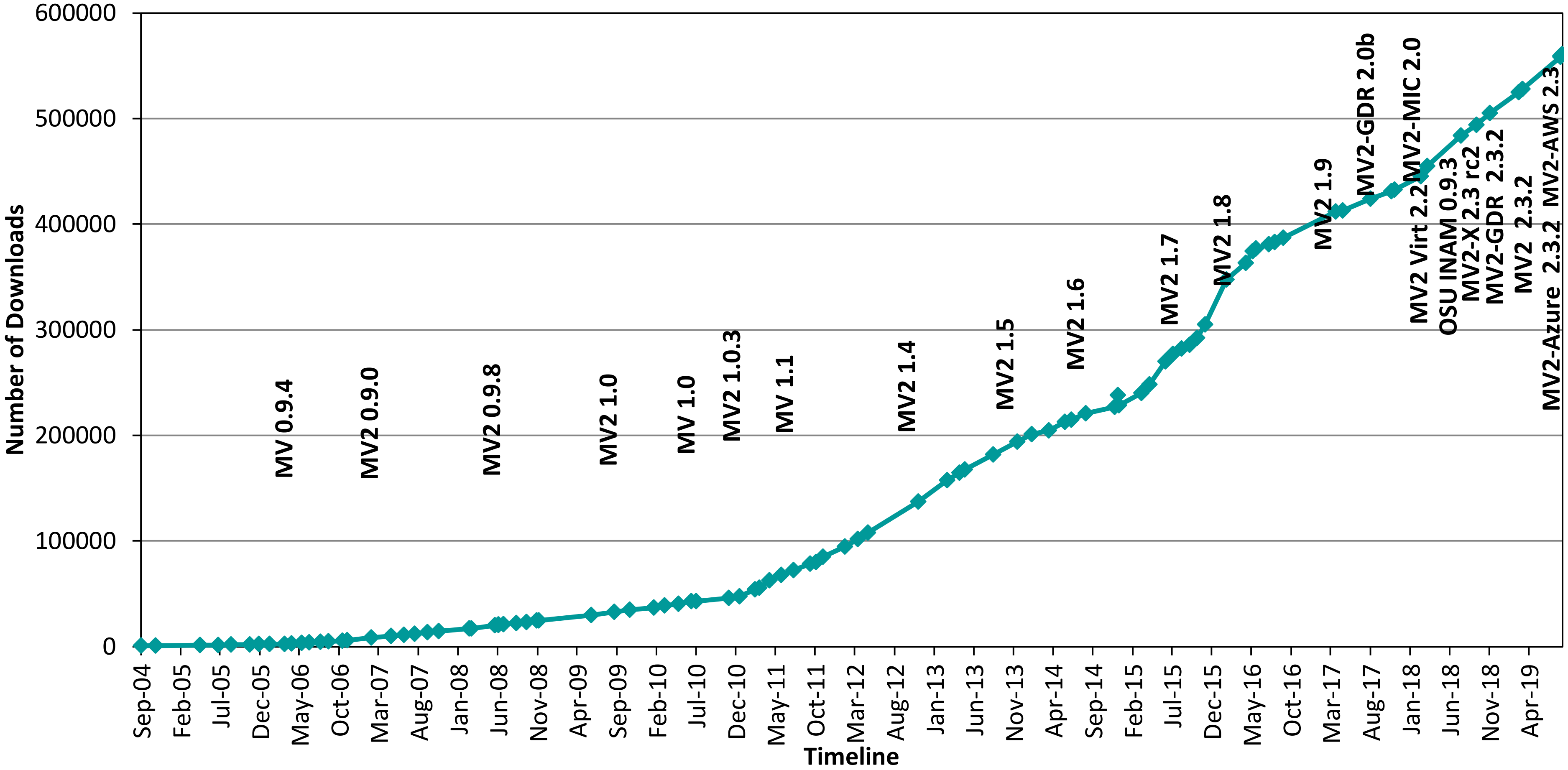
- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Scalable job start-up
  - Low memory footprint
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Balancing intra-node 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 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 3,050 organizations in 89 countries**
  - **More than 615,000 (> 0.6 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Jun '19 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**
- Empowering Top500 systems for over a decade

Partner in the TACC Frontera System

MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family (HPC and DL)

High Performance Parallel Programming Models

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

High Performance and Scalable Communication Runtime

Diverse APIs and Mechanisms

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

Support for Modern Networking Technology
(InfiniBand, iWARP, RoCE, Omni-Path, Elastic Fabric Adapter)

Transport Protocols: RC, SRD, UD, DC

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

Support for Modern Multi-/Many-core Architectures
(Intel-Xeon, OpenPOWER, Xeon-Phi, ARM, NVIDIA GPGPU)

Transport Mechanisms: Shared Memory, CMA, IVSHMEM, XPMEM

Modern Features: Optane*, NVLink, CAPI*

* Upcoming

MVAPICH2 Software Family

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

Features and Improvement in MAVPIACH2-X for ARM

- Enhanced architecture and IB HCA detection for various ARM systems
- Optimization and tuning for
  - Intra-node and inter-node point-to-point operations
  - Intra-node shared memory communication protocols
  - Collective operations for different message sizes and job/system sizes using the existing collective algorithms in MVAPICH2-X
- Optimizations to job startup performance to achieve scalable job startup when running large-scale jobs on ARM systems
- Support for latest GCC and ARM compilers

Performance Evaluation of Optimized MVAPICH2-X

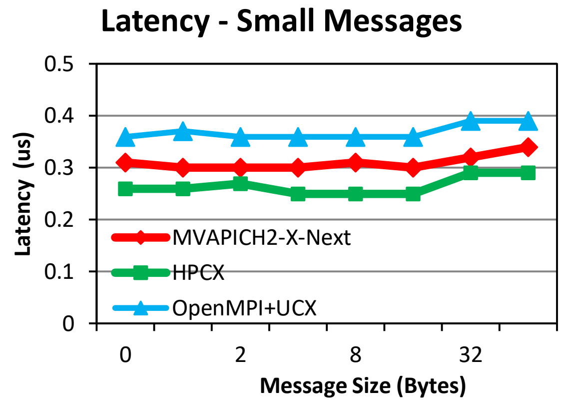
- **EPCC Fulhame Cluster**
  - Nodes: 16 x ARM ThunderX2
  - Processor: 2x 32 core ARM ThunderX2
  - Network: EDR 100Gbps MT4119
  - Operating System: Linux 4.12.14-23-default
  - MPI and Communication Libraries
    - MVAPICH2-X (latest)
    - HPCX-v2.4.0-gcc-MLNX_OFED_LINUX-4.6-1.0.1.1-suse15.0-aarch64
    - OpenMPI-4.0.2 w/ latest UCX
  - OSU-Microbenchmarks-v5.6.2

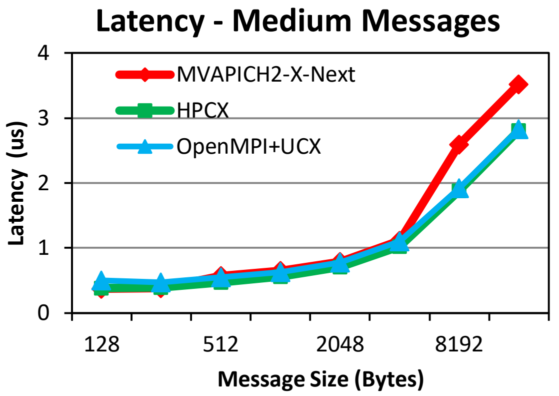
- **Mayer Cluster**
  - Nodes: 14 x ARM ThunderX2
  - Processor: 2x 28 core ARM ThunderX2
  - Network: EDR 100Gbps MT4119
  - Operating System: Linux 4.14.0-115.13
  - MPI and Communication Libraries
    - MVAPICH2-X (latest)
    - OpenMPI 4.0.1
    - UCX 1.5.2
  - OSU-Microbenchmarks-v5.6.2

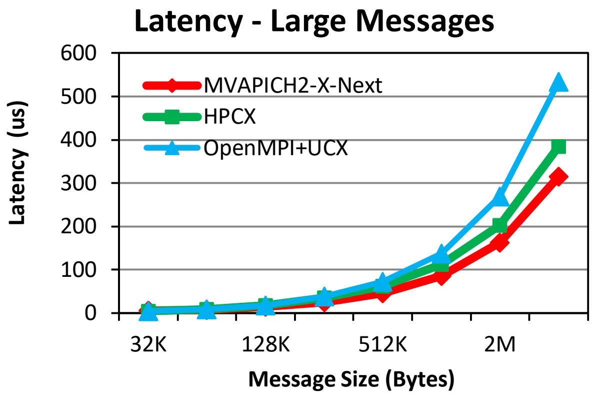
Evaluation of Point-to-point on EPCC ARM System

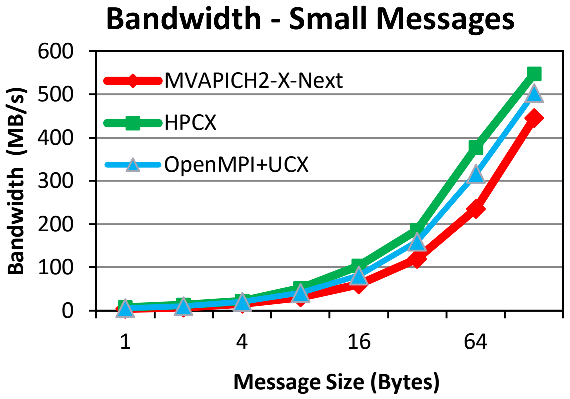
- EPCC Fulhame ARM cluster with up to 16 dual-socket 32-core ThunderX2 nodes
- Comparison among MVAPICH2X (Next), OpenMPI+UCX, and HPCX communication libraries
- OSU Micro-benchmark Suite (OMB) v5.6.2
- Measure the MPI-level communication performance of latency, bandwidth, bi-directional bandwidth, and message rate
- Three different configurations
  - Intra-socket
  - Inter-socket
  - Inter-node

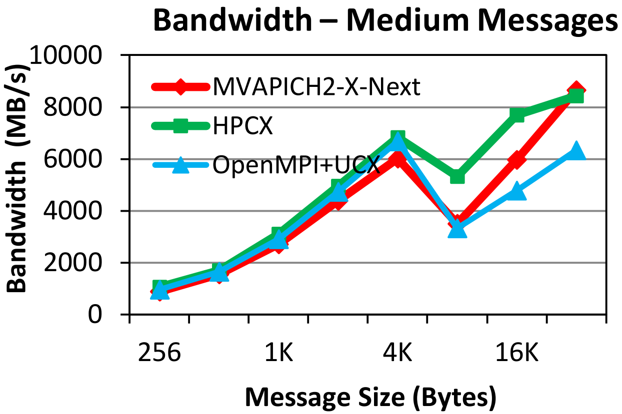
Point-to-point: Latency & Bandwidth (Intra-socket)

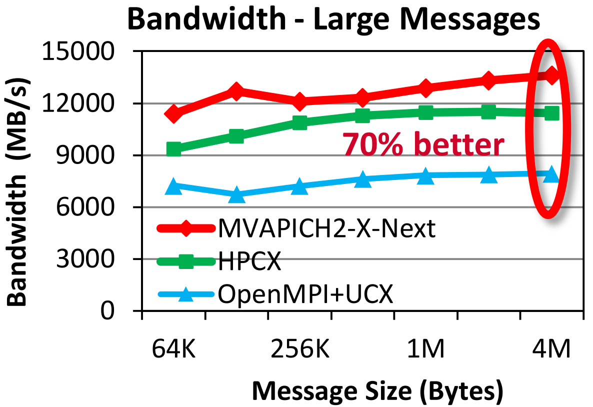




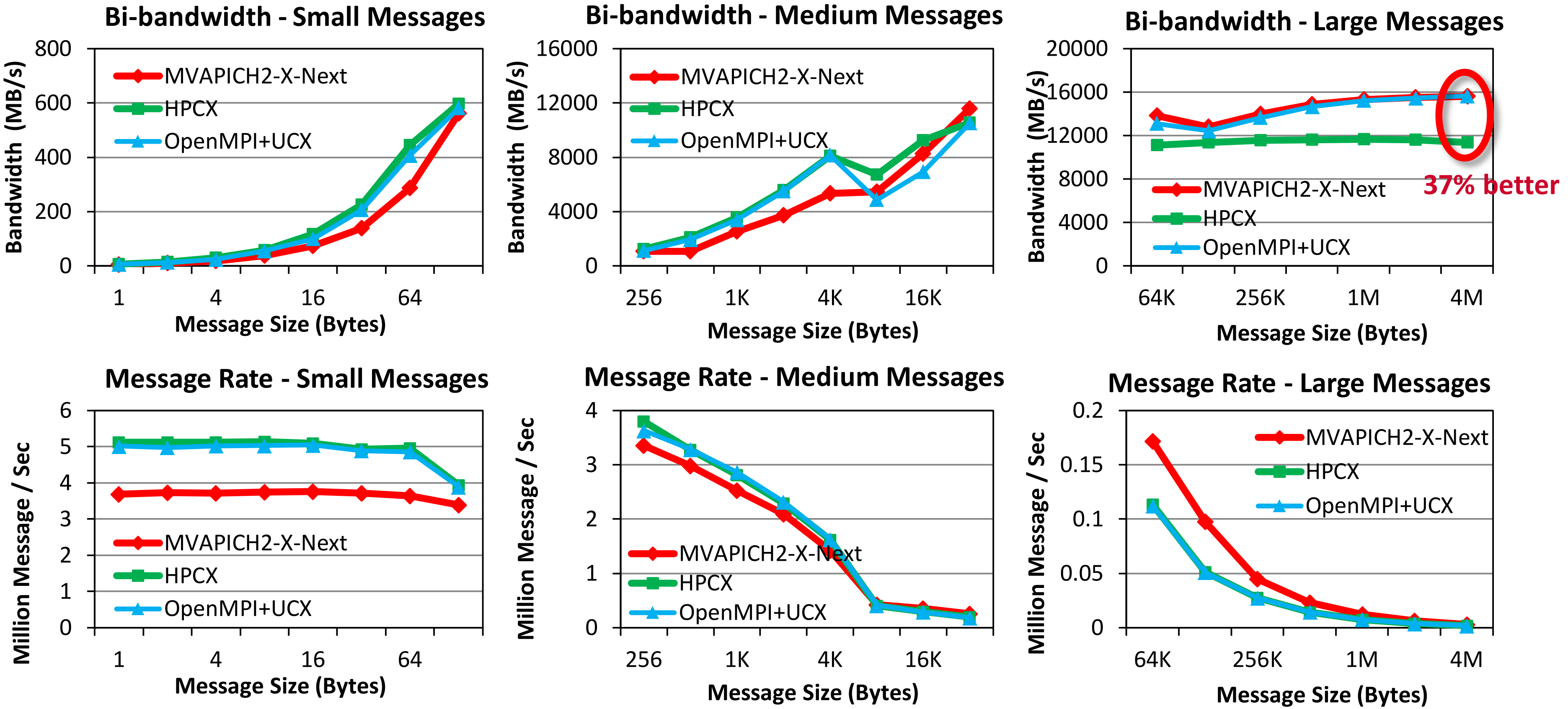




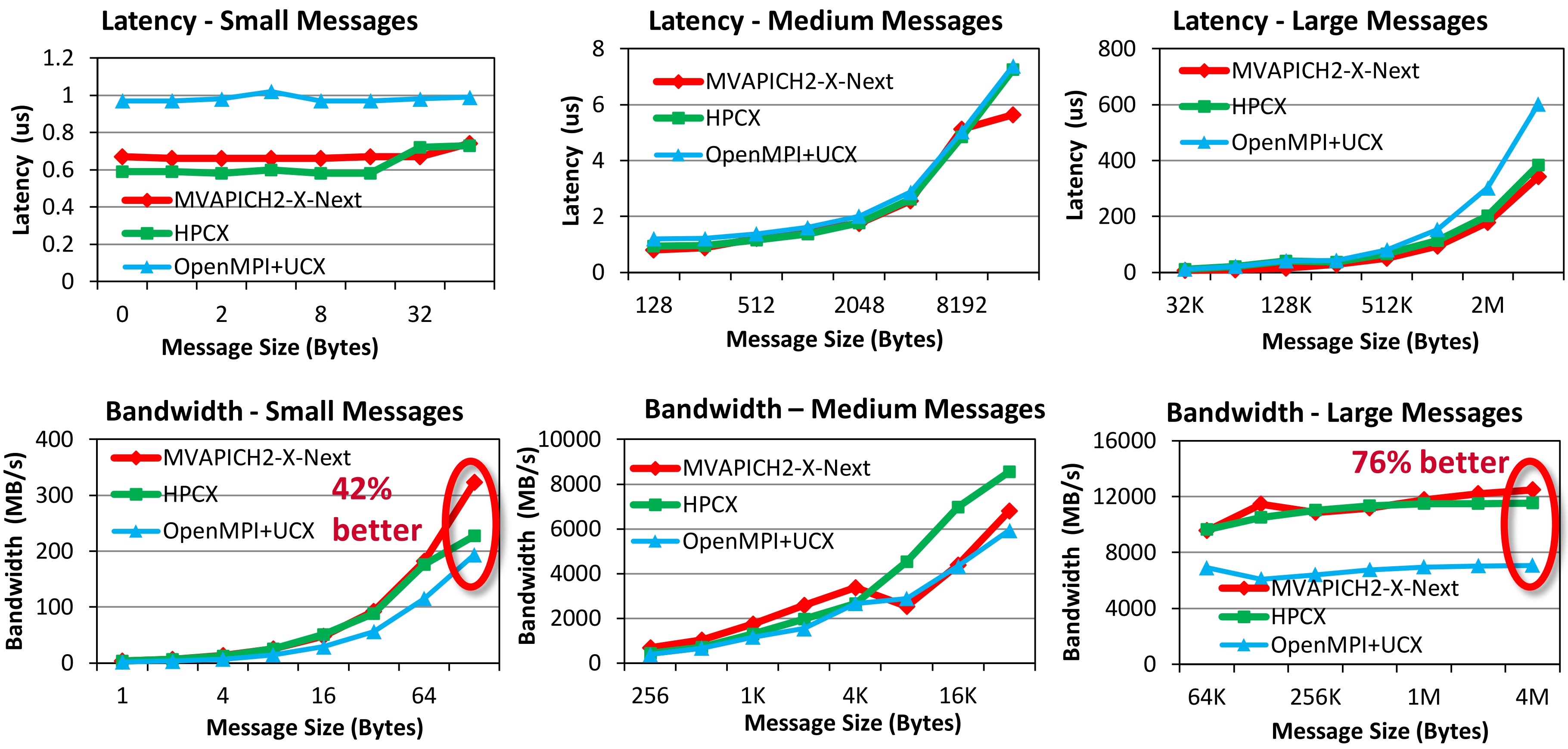




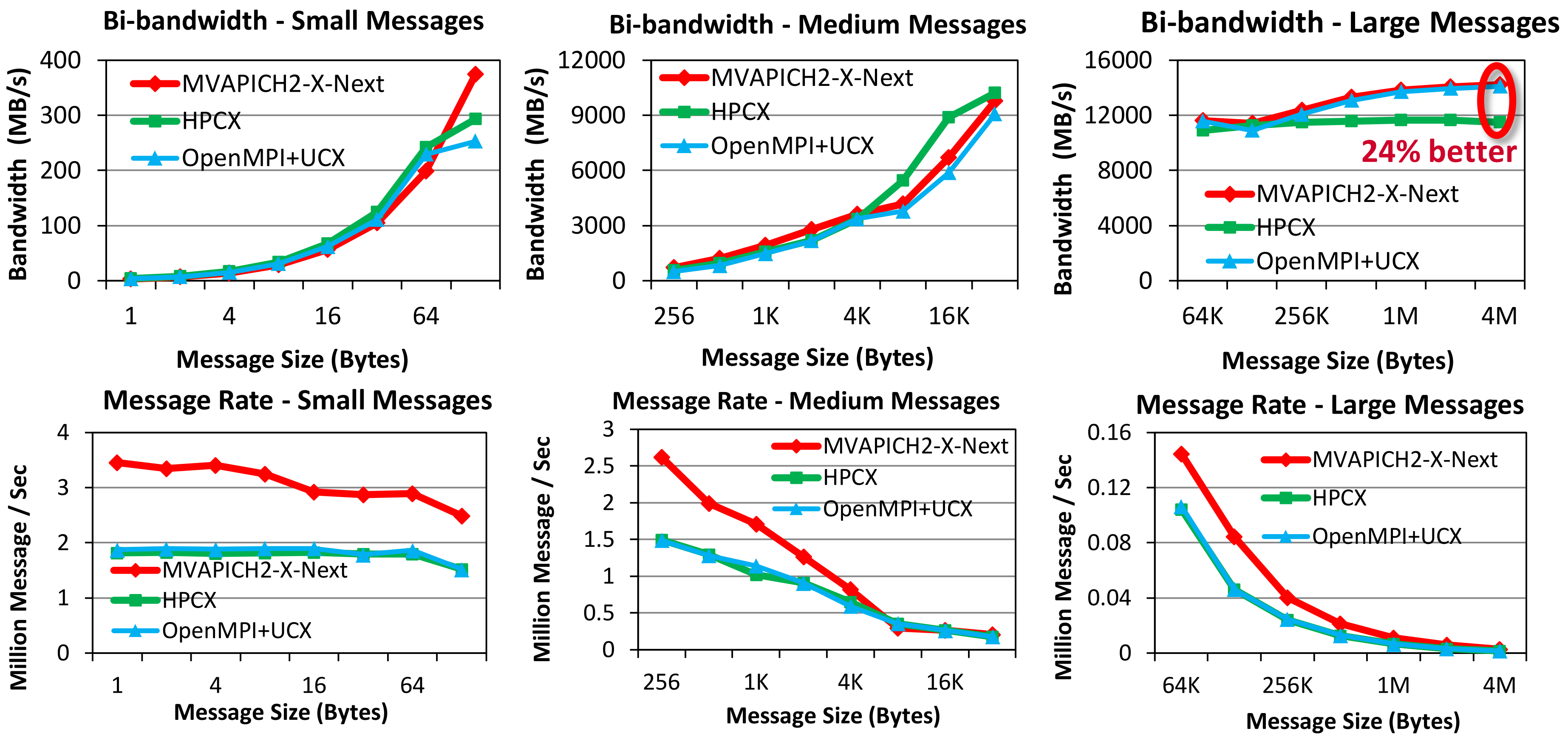
Point-to-point: Bi-Bandwidth (Intra-socket)



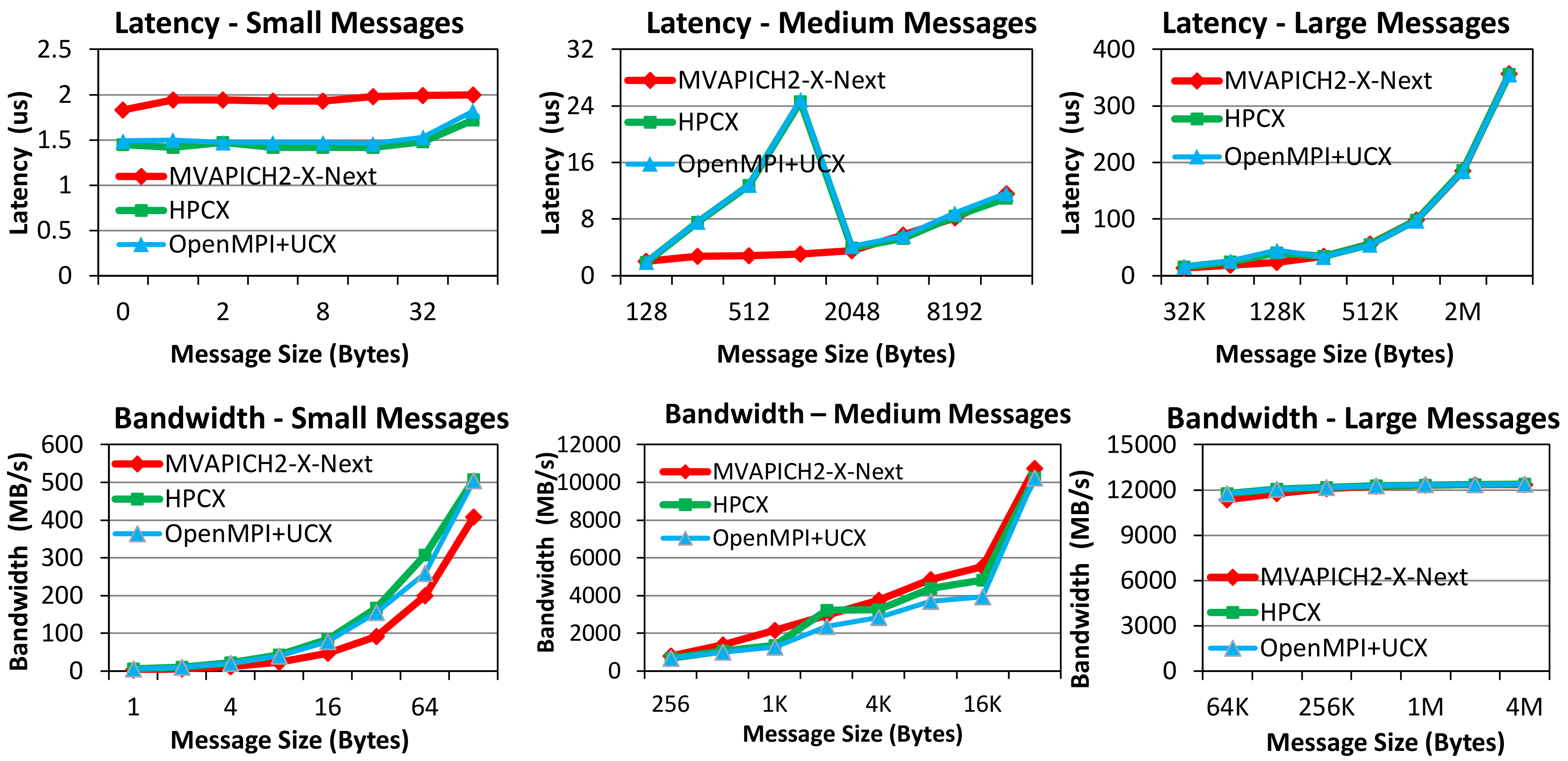
Point-to-point: Latency & Bandwidth (Inter-socket)



Point-to-point: Bi-Bandwidth (Inter-socket)



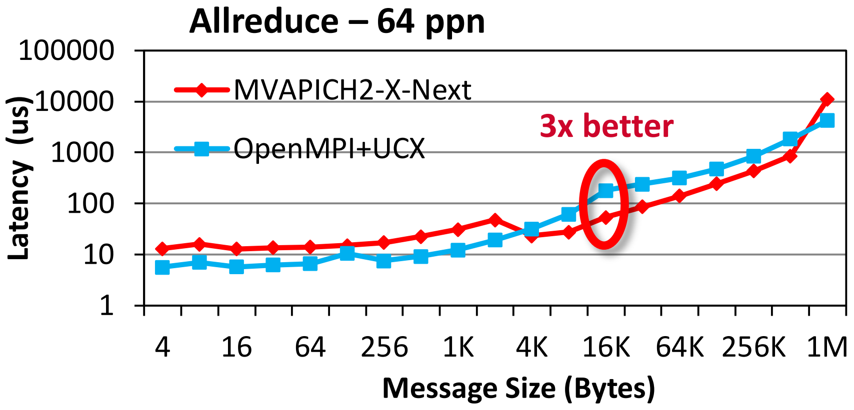
Point-to-point: Latency & Bandwidth (Inter-Node)

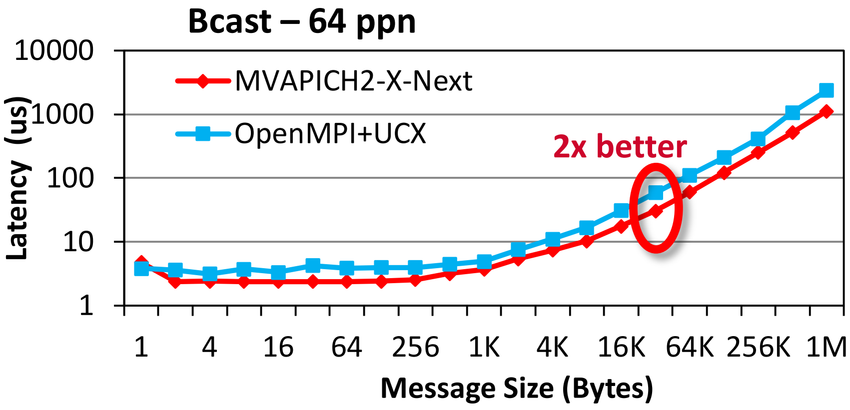


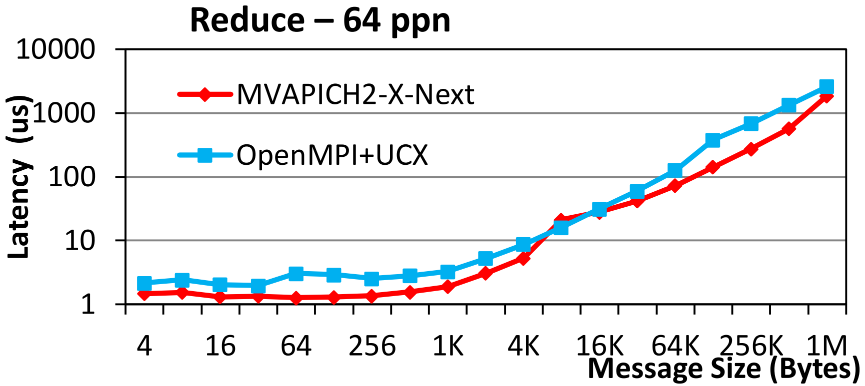
Evaluation of Collectives Communication on EPCC ARM System

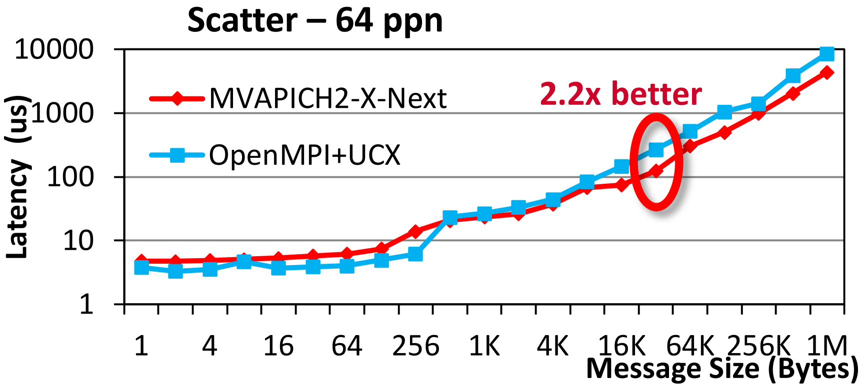
- Fulhame cluster with up to 16 dual-socket 32-core ThunderX2 nodes
- Comparison among MVAPICH2X (Next), OpenMPI+UCX, and HPCX communication libraries
- OSU Micro-benchmark Suite (OMB) 5.6.2
- Measure the MPI-level communication performance of collectives communication latency
- Evaluate single-socket (half-subscription) and dual-socket (full-subscription) scenarios on varying scale

Collectives: Single Node (64-ppn)

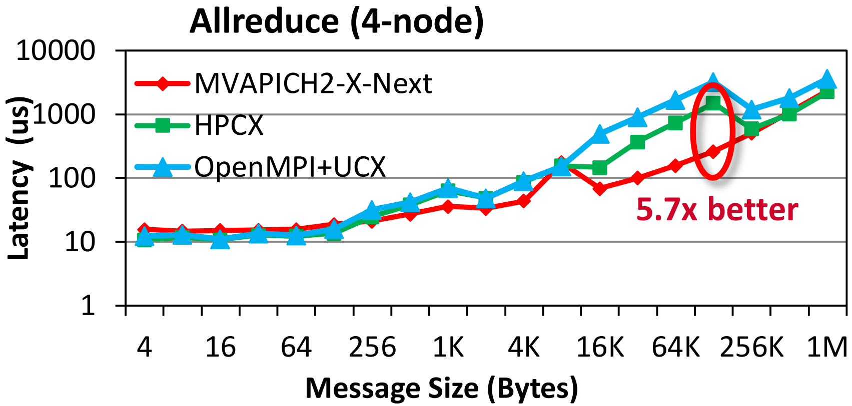


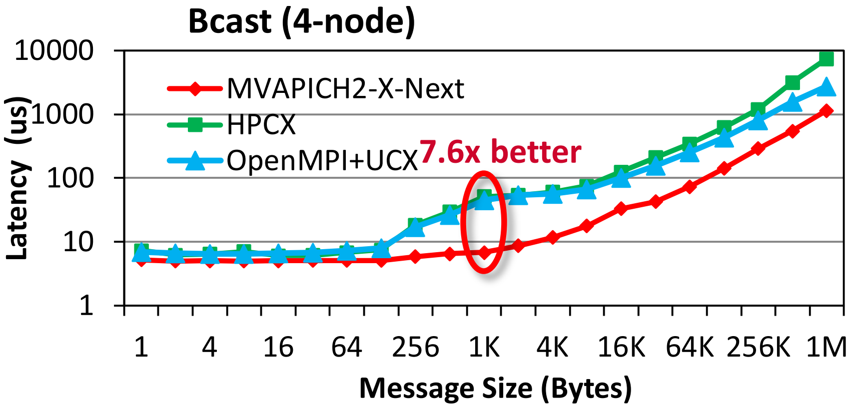


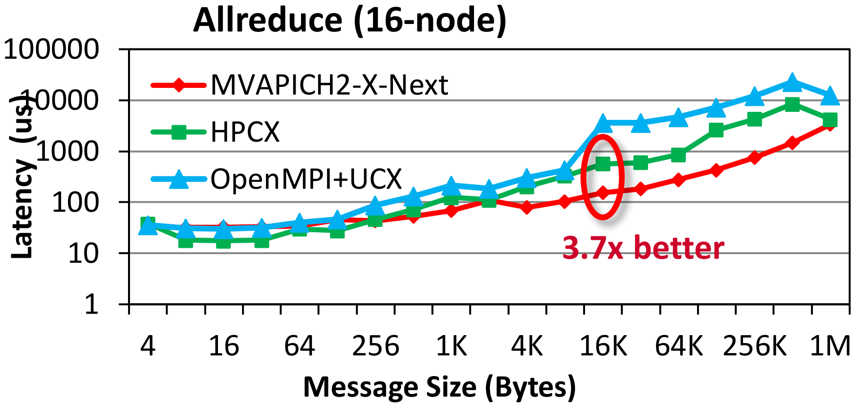


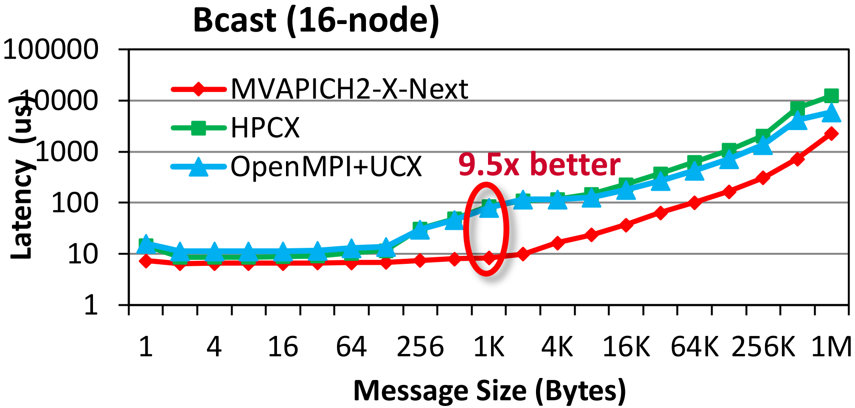


Collectives: 4 & 16 Nodes (32-ppn)

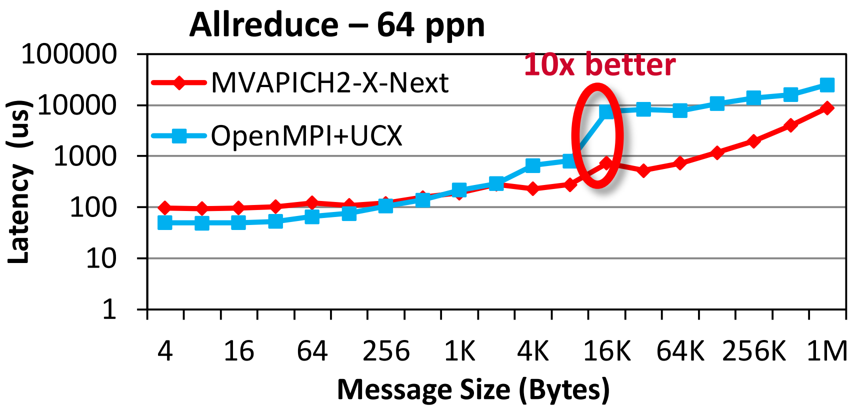


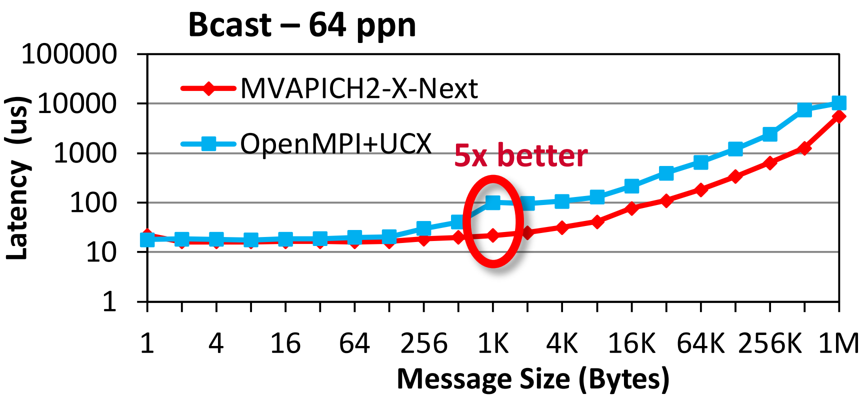


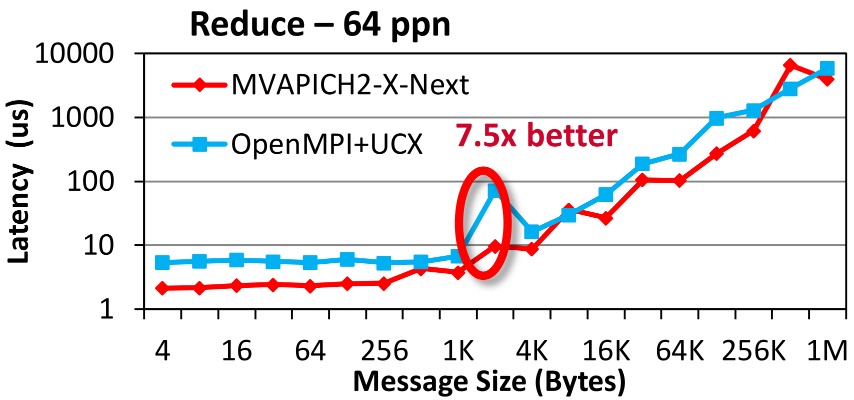


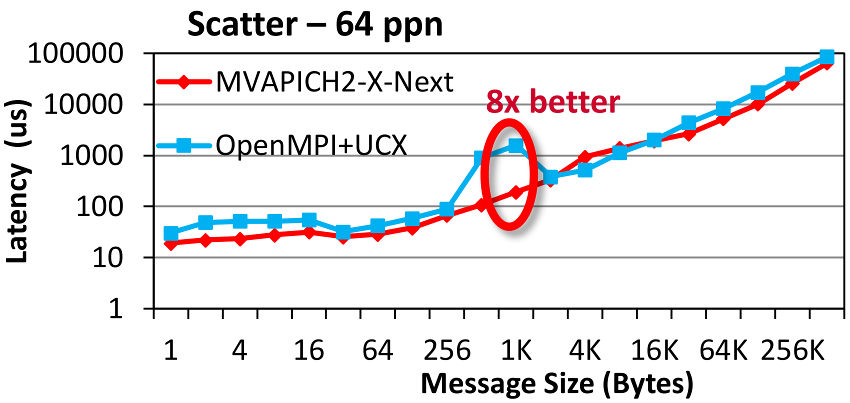


Collectives: 16 Nodes (64-ppn)

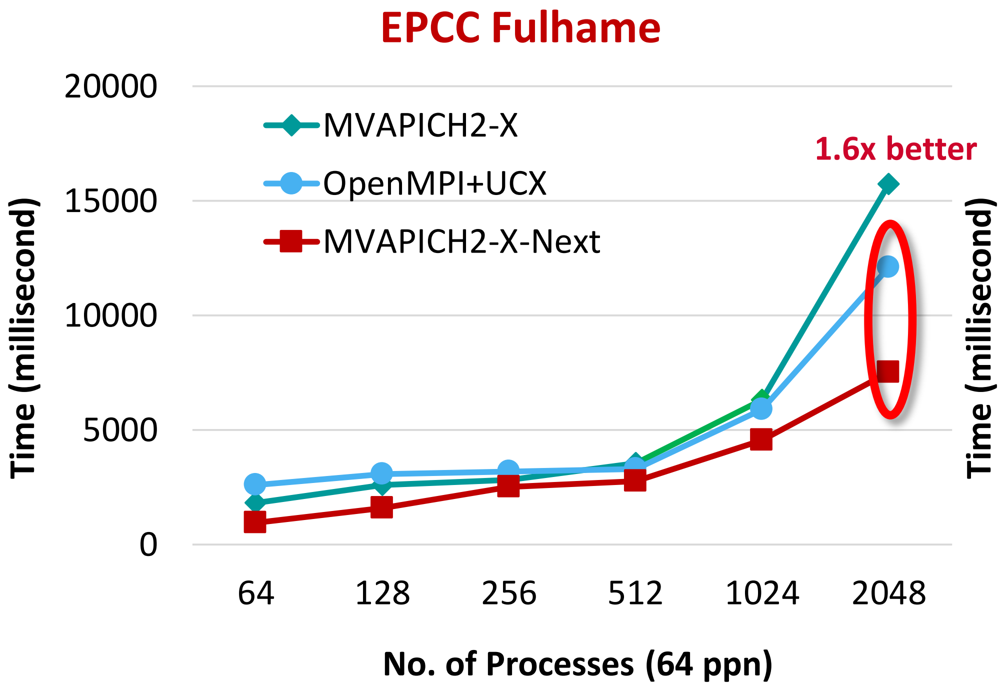


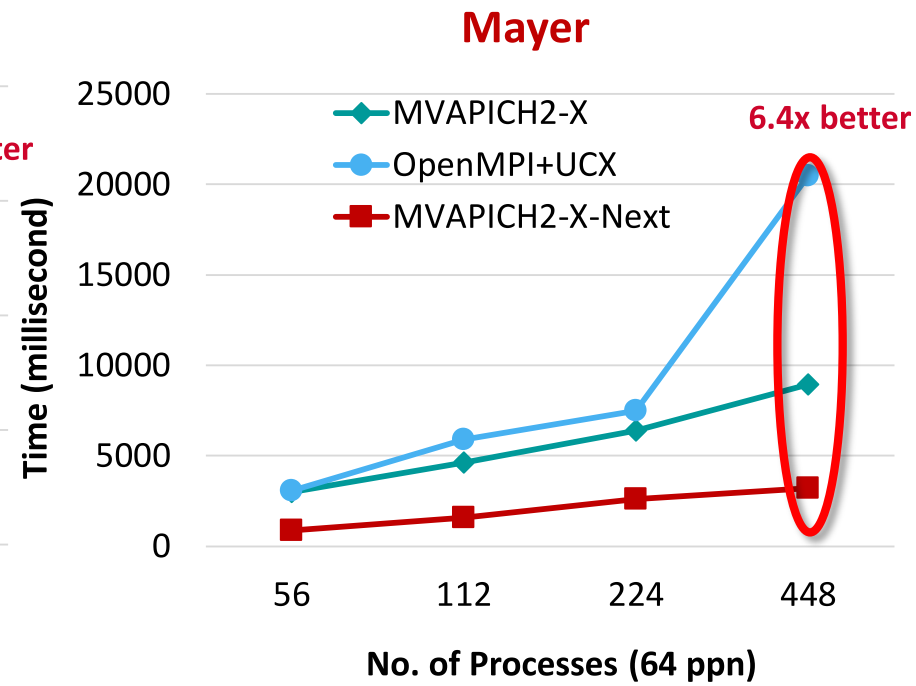






MPI Job Startup Evaluation on different ARM clusters



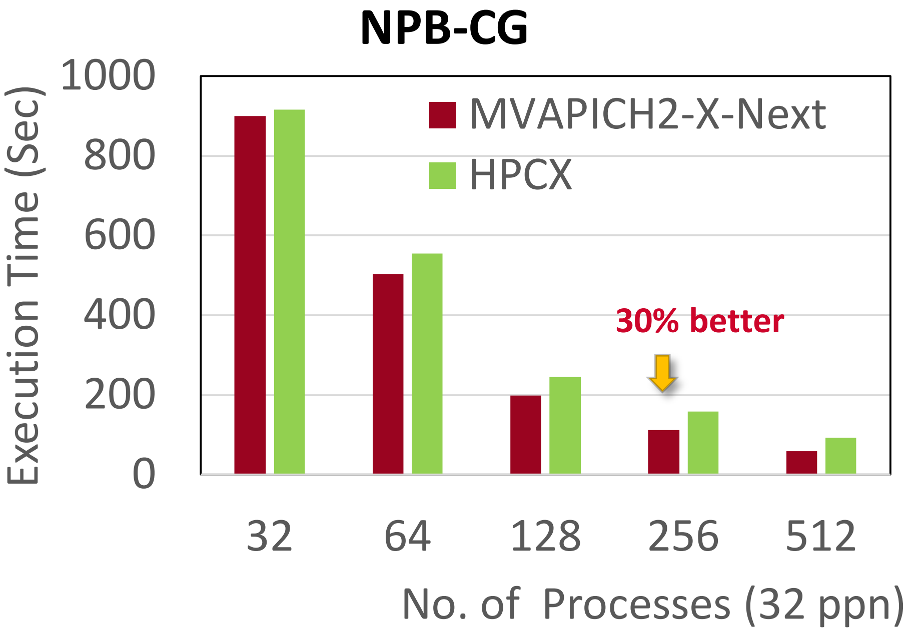


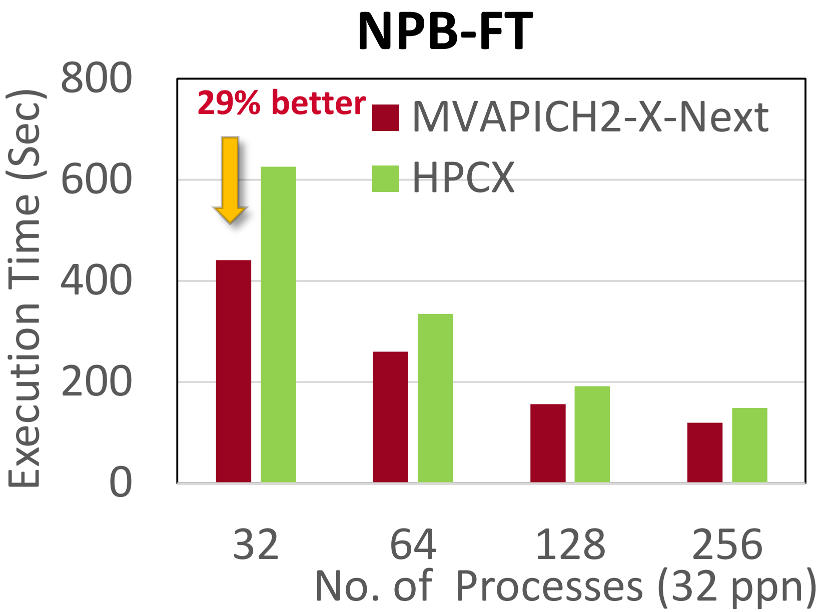
- Up to 1.6x speedup over OpenMPI w/UCX on Catalyst Fulhame system
- Up to 6.4x speedup over OpenMPI w/ UCX on Mayer system

Evaluation of Application Kernels

- Evaluation of NAS Parallel Benchmarks, MiniAMR, and Cloverleaf kernels
- Comparison among MVAPICH2-X (Next), OpenMPI+UCX, and HPCX communication libraries
- Measure the application communication performance at varying scales with full-subscription scenarios on up to 1,024 processes
- Significant performance improvement is observed when using MVAPICH2-X

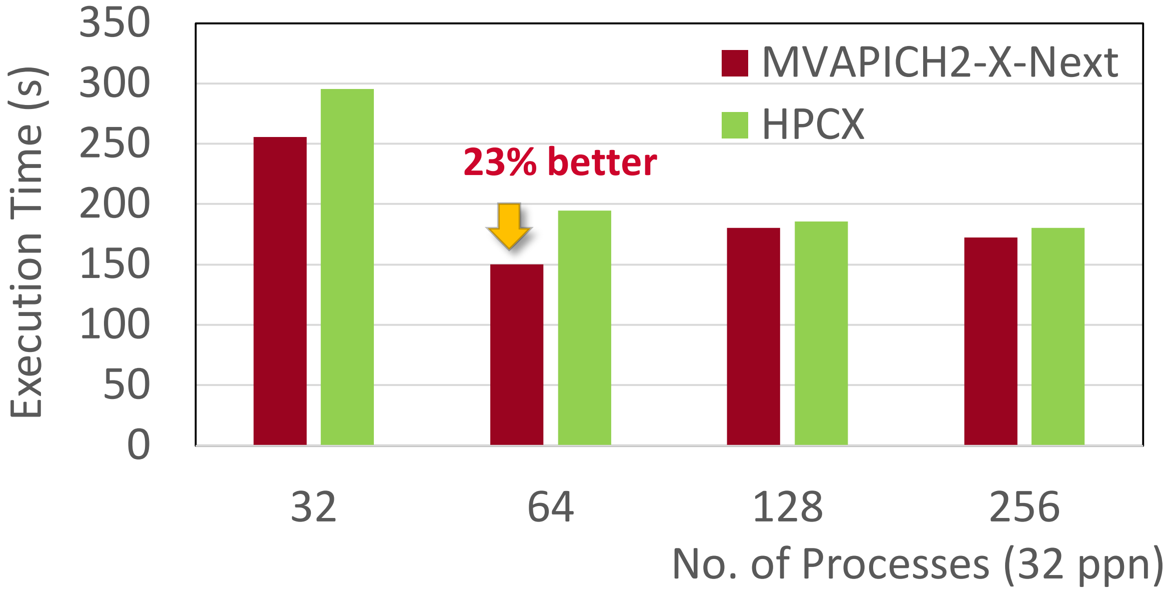
Application Evaluation – (NAS Parallel Benchmarks)





- NPB-3.4 Class-D comparing MVAPICH2-X (upcoming) and HPCX on EPCC Fulhame
- Up to 30% and 29% improvement over HPCX for CG and FT kernels.

Application Evaluation – (MiniAMR)



- MiniAMR kernel comparing MVAPICH2-X (upcoming) and HPCX on EPCC Fulhame
- Up to 23% improvement over HPCX is observed.

Input Parameters: --percent_sum 0 --num_vars 10 --stencil 21 --report_diffusion 0 --report_perf 2 --num_tsteps 100 --num_spikes 1

Conclusions

- ARM has emerged as a new platform for HPC systems
- Requires high-performance middleware designs while exploiting modern interconnects (InfiniBand)
- Provided the approaches being taken care of by the MVAPICH2 project to provide MPI support with high-performance
- Will continue to optimize and tune the MVAPICH2 stack for higher performance and scalability on ARM platforms

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

Multiple Events at SC ‘19

- Presentations at OSU and X-Scale Booth (#2094)
  - Members of the MVAPICH, HiBD and HiDL members
  - External speakers
- Presentations at SC main program (Tutorials, Workshops, BoFs, Posters, and Doctoral Showcase)
- Presentation at many other booths (Mellanox, Intel, Microsoft, and AWS) and satellite events
- Complete details available at

  http://mvapich.cse.ohio-state.edu/conference/752/talks/

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 Students (Undergraduate)

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

Current Post-doc

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

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

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/