MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

HiBD

High-Performance Big Data

HiDL

High-Performance Deep Learning

Accelerating HPC Applications on HPC Systems with Intel Omni-Path: The MVAPICH Approach

Omni-Path User Group (OPUG) Meeting at ISC'19

by

Dhabaleswar K. (DK) Panda

The Ohio State University

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

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

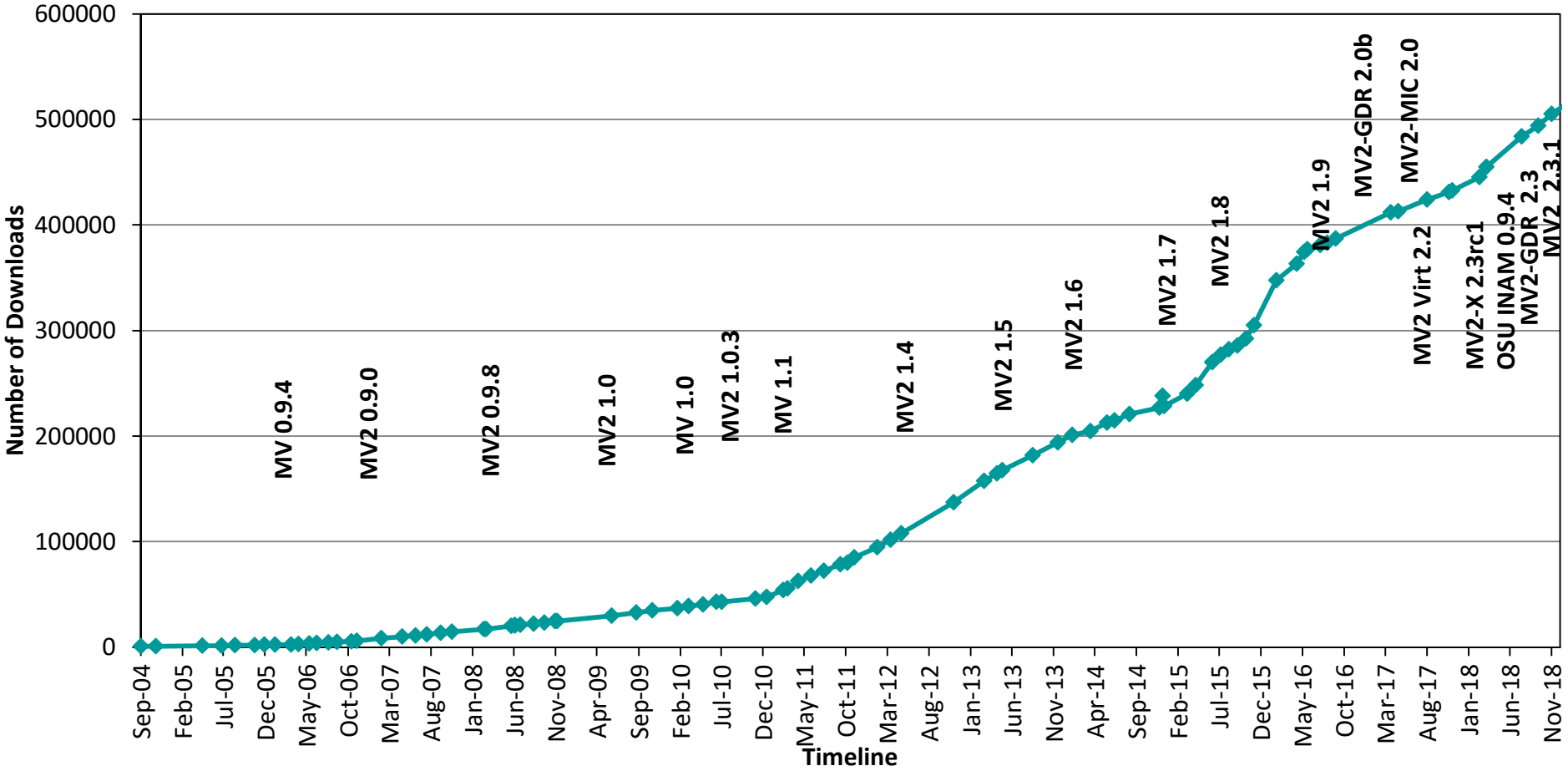
Overview of the MVAPICH2 Project

- 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 3,000 organizations in 89 countries**
  - **More than 549,000 (> 0.5 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '18 ranking)
    - 3rd ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China
    - 16th, 556,104 cores (Oakforest-PACS) in Japan
    - 19th, 367,024 cores (Stampede2) at TACC
    - 31st, 241,108-core (Pleiades) at NASA 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

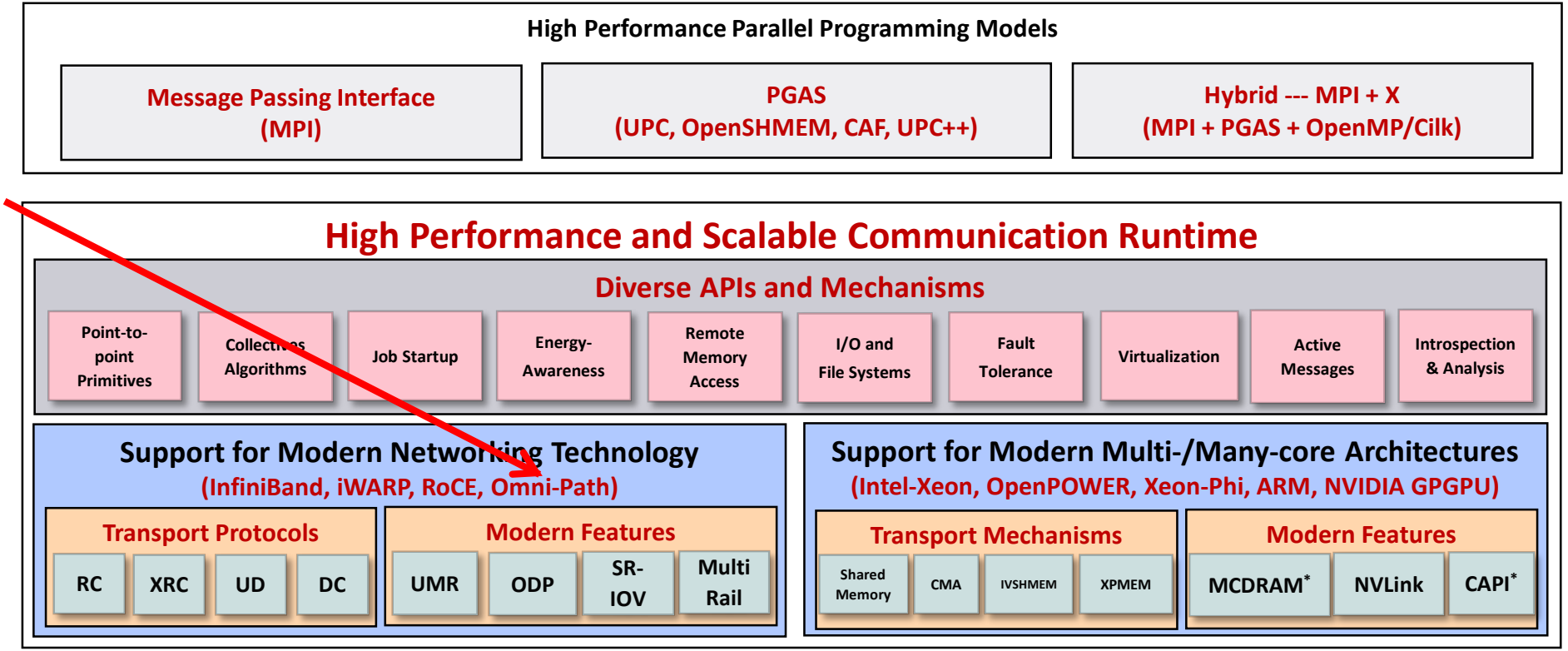
MVAPICH2 Release Timeline and Downloads



History of Support for Omni-Path in MVAPICH2

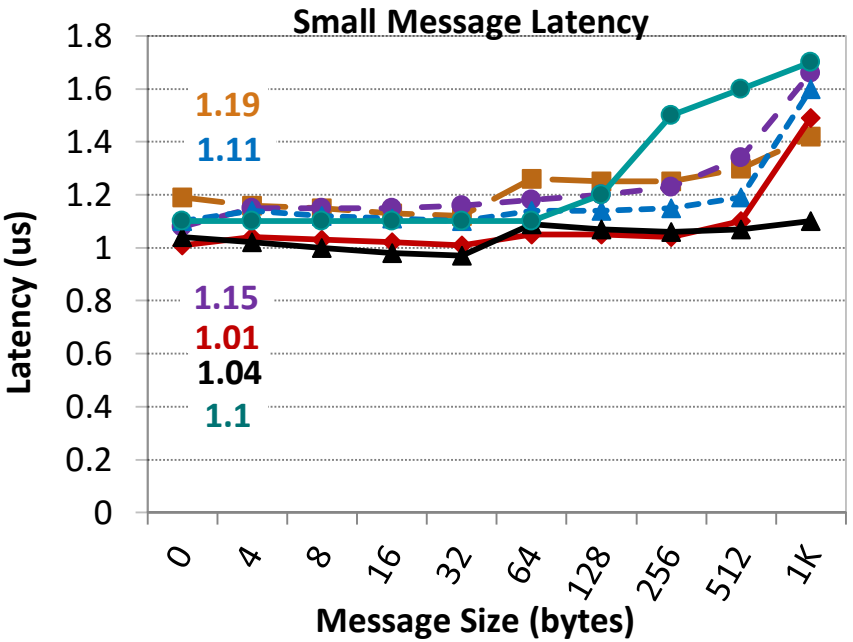
- Initial designs for the Performance Scaled Messaging (PSM) interface for QLogic cards added in 2008
- Designs later enhanced in collaboration with Intel for Omni-Path by updating to the PSM2 interface
- Code being actively developed, tested and deployed at multiple supercomputing centers including Stampede2@TACC, OakForest-PACS, …

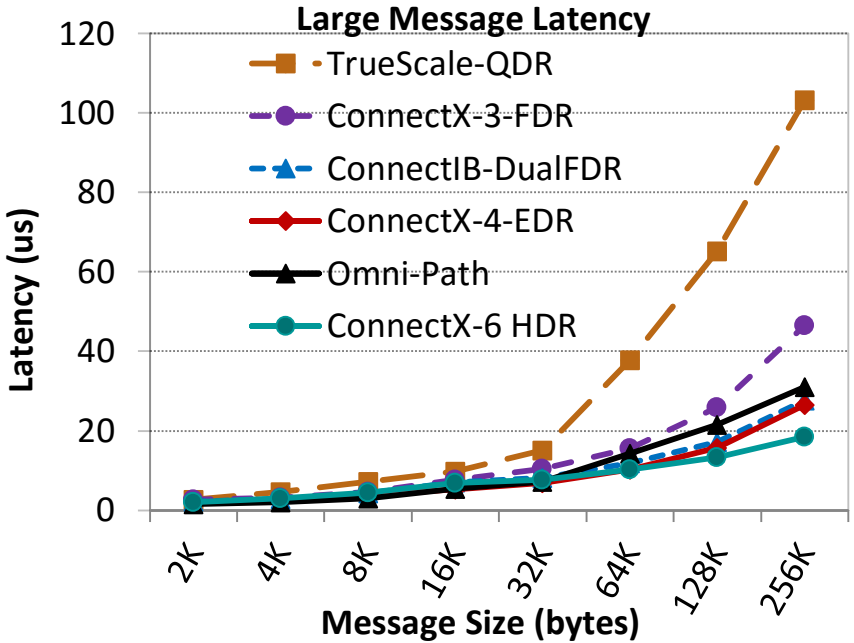
Architecture of MVAPICH2 Software Family



* Upcoming

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

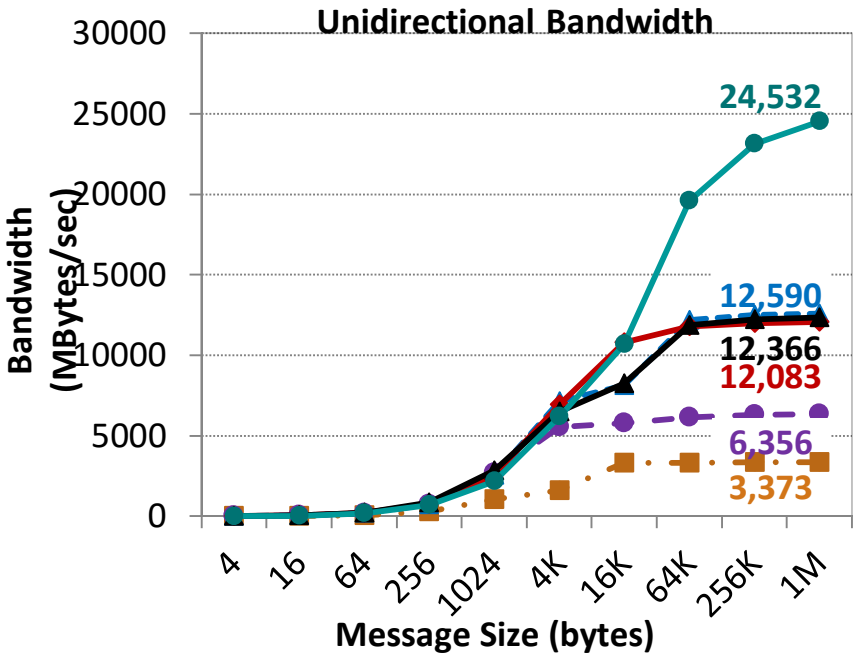
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

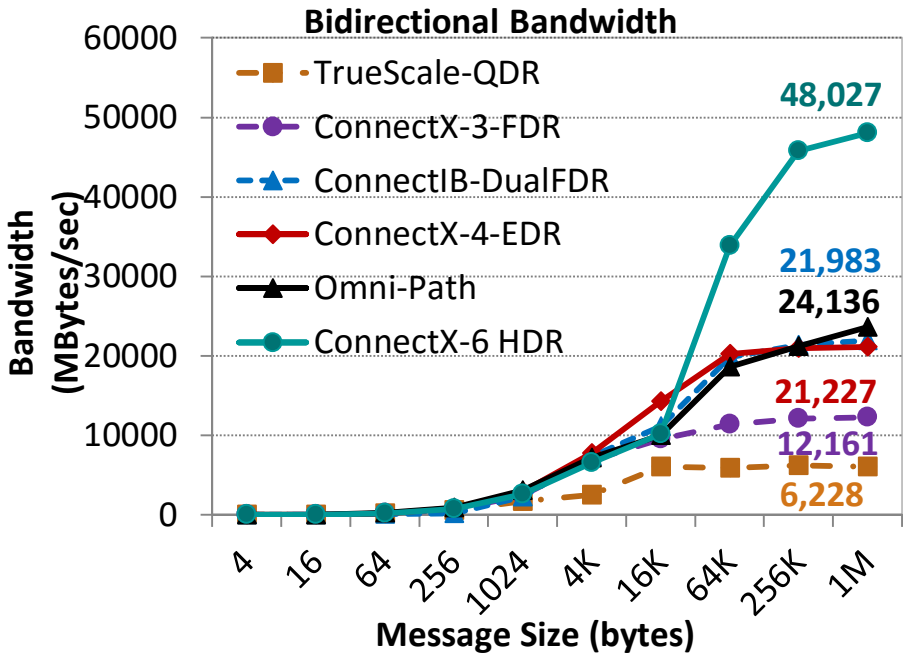
ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

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

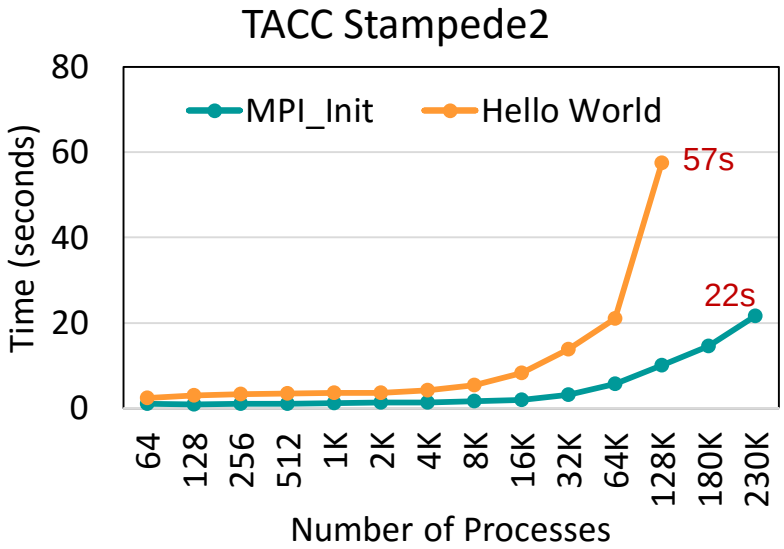
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

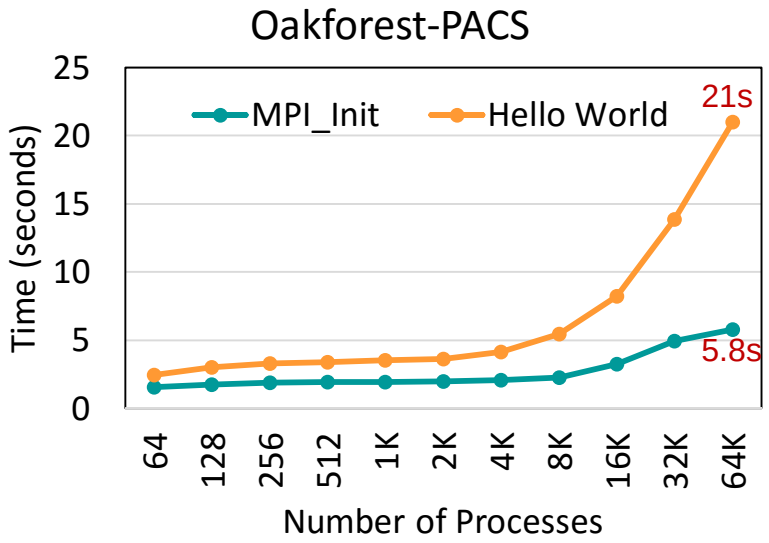
ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

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

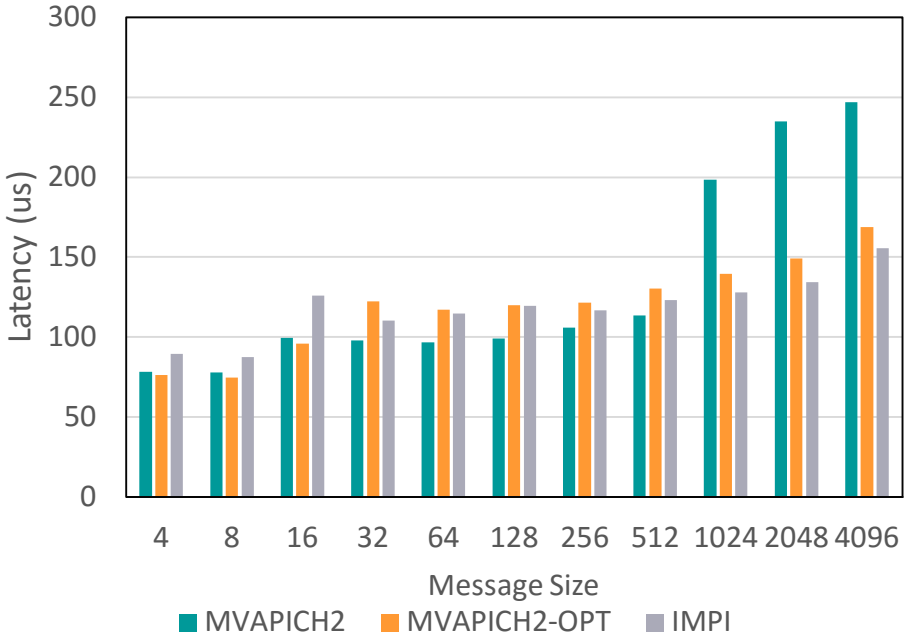
Startup Performance on KNL + Omni-Path

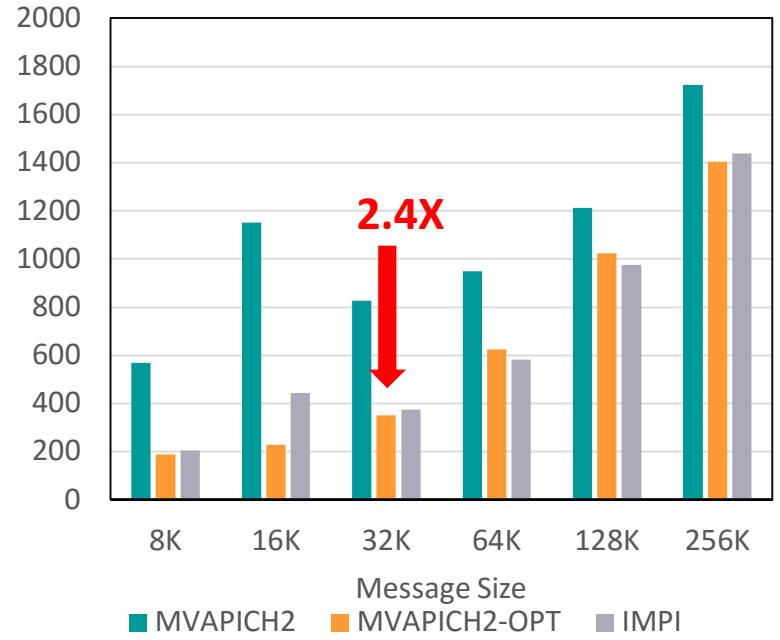




- MPI_Init takes 22 seconds on 231,936 processes on 3,624 KNL nodes (Stampede2 – Full scale)
- At 64K processes, MPI_Init and Hello World takes 5.8s and 21s respectively (Oakforest-PACS)
- All numbers reported with 64 processes per node, MVAPICH2-2.3a
- Designs integrated with mpirun_rsh, available for srun (SLURM launcher) as well

MPI_Allreduce on KNL + Omni-Path (10,240 Processes)





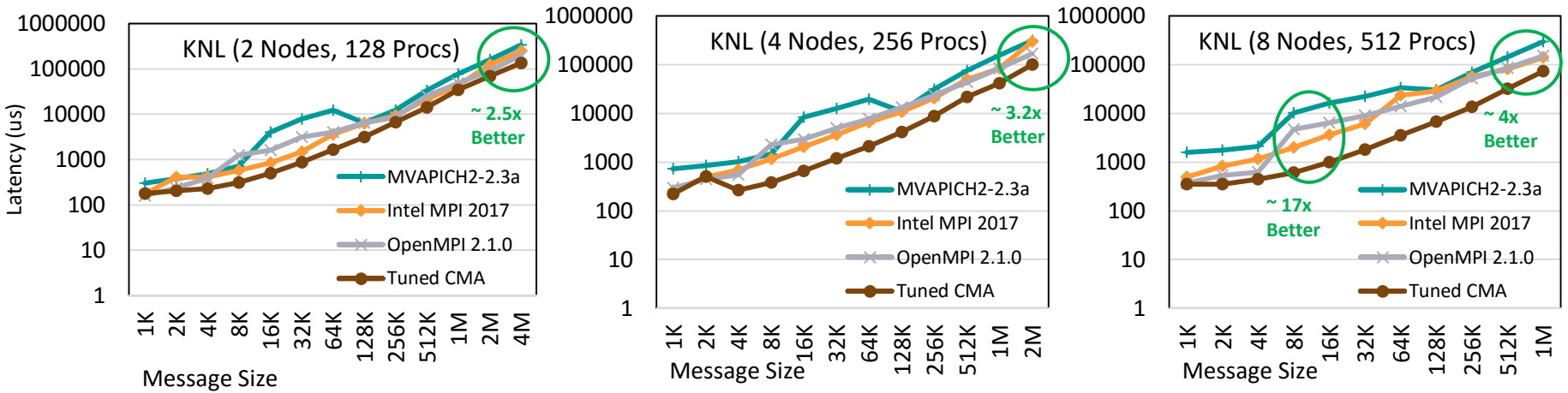
OSU Micro Benchmark 64 PPN

- For MPI_Allreduce latency with 32K bytes, MVAPICH2-OPT can reduce the latency by 2.4X

M. Bayatpour, S. Chakraborty, H. Subramoni, X. Lu, and D. K. Panda, Scalable Reduction Collectives with Data Partitioning-based Multi-Leader Design, SuperComputing '17.

Available since MVAPICH2-X 2.3b

Optimized CMA-based Collectives for Large Messages



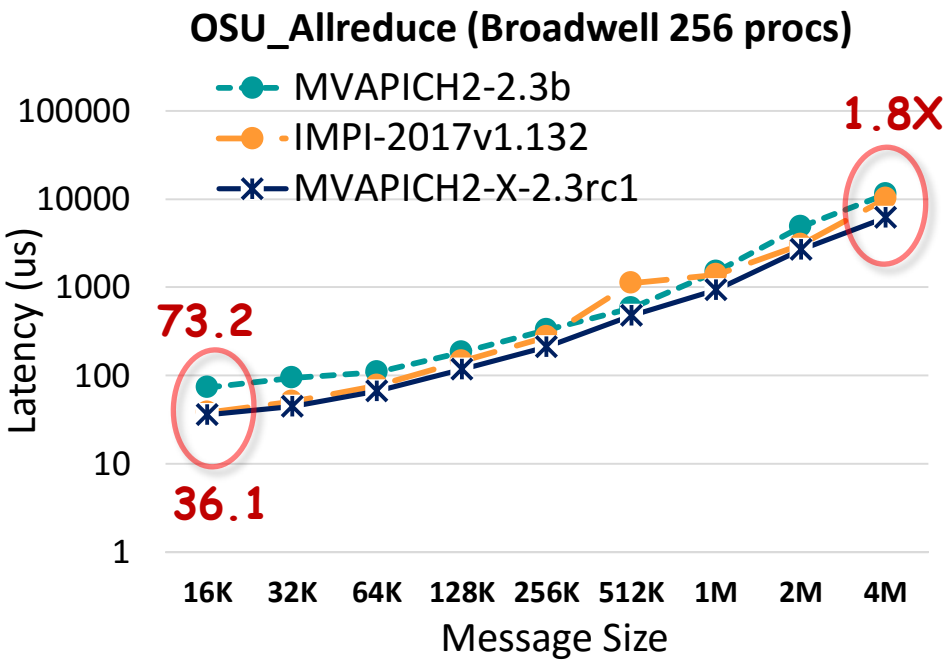
Performance of MPI_Gather on KNL nodes with Omni-Path (64PPN)

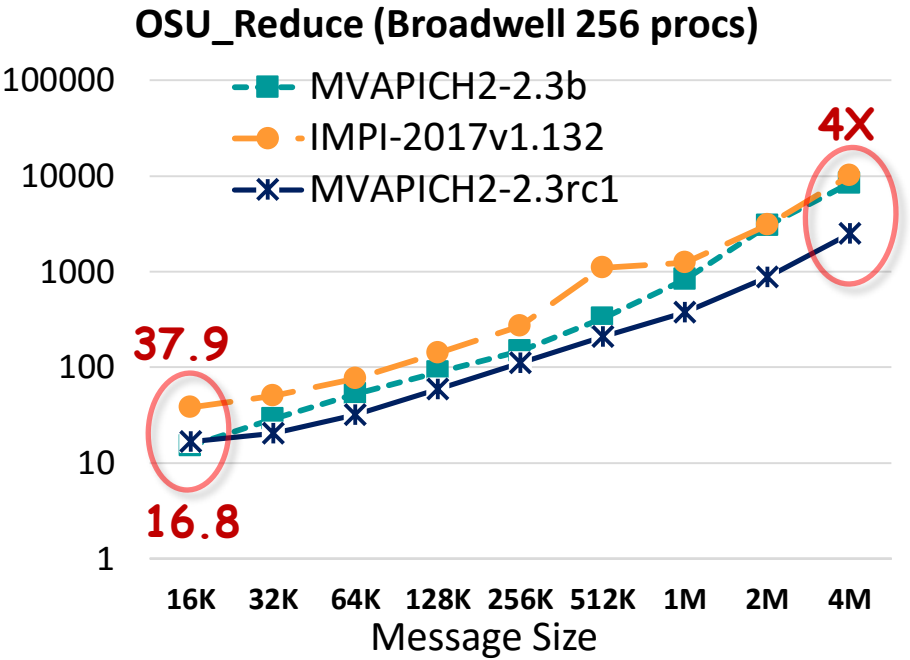
- Significant improvement over existing implementation for Scatter/Gather with 1MB messages (up to 4x on KNL, 2x on Broadwell, 14x on OpenPower)
- New two-level algorithms for better scalability
- Improved performance for other collectives (Bcast, Allgather, and Alltoall)

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

Shared Address Space (XPMEM)-based Collectives Design



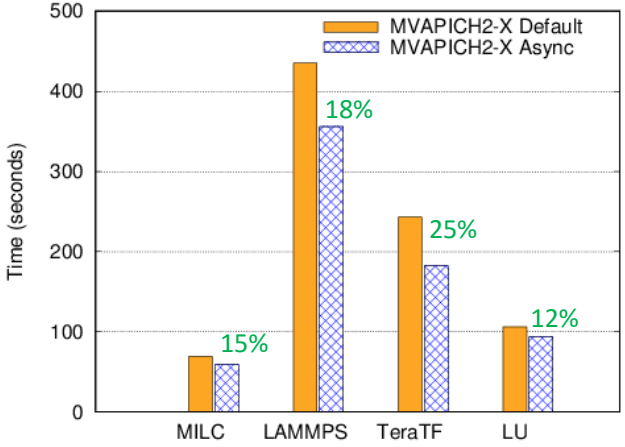


- "*Shared Address Space*"-based true *zero-copy* Reduction collective designs in MVAPICH2
- Offloaded computation/communication to peers ranks in reduction collective operation
- Up to **4X** improvement for 4MB Reduce and up to **1.8X** improvement for 4M AllReduce

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.

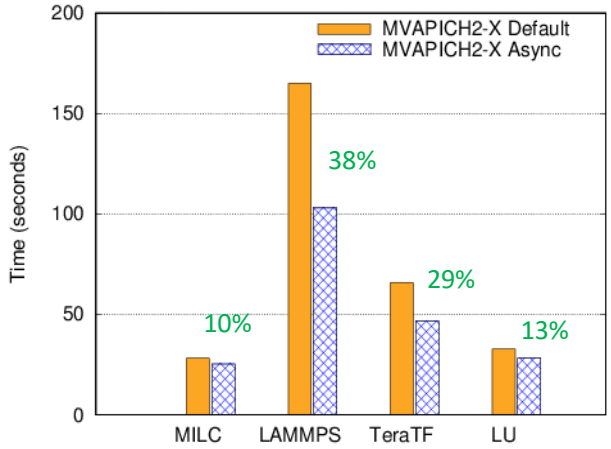
Available since MVAPICH2-X 2.3rc1

Benefits of the New Asynchronous Progress Design



384 Processes (6 Nodes : 64 PPN)

SPEC MPI : KNL + Omni-Path



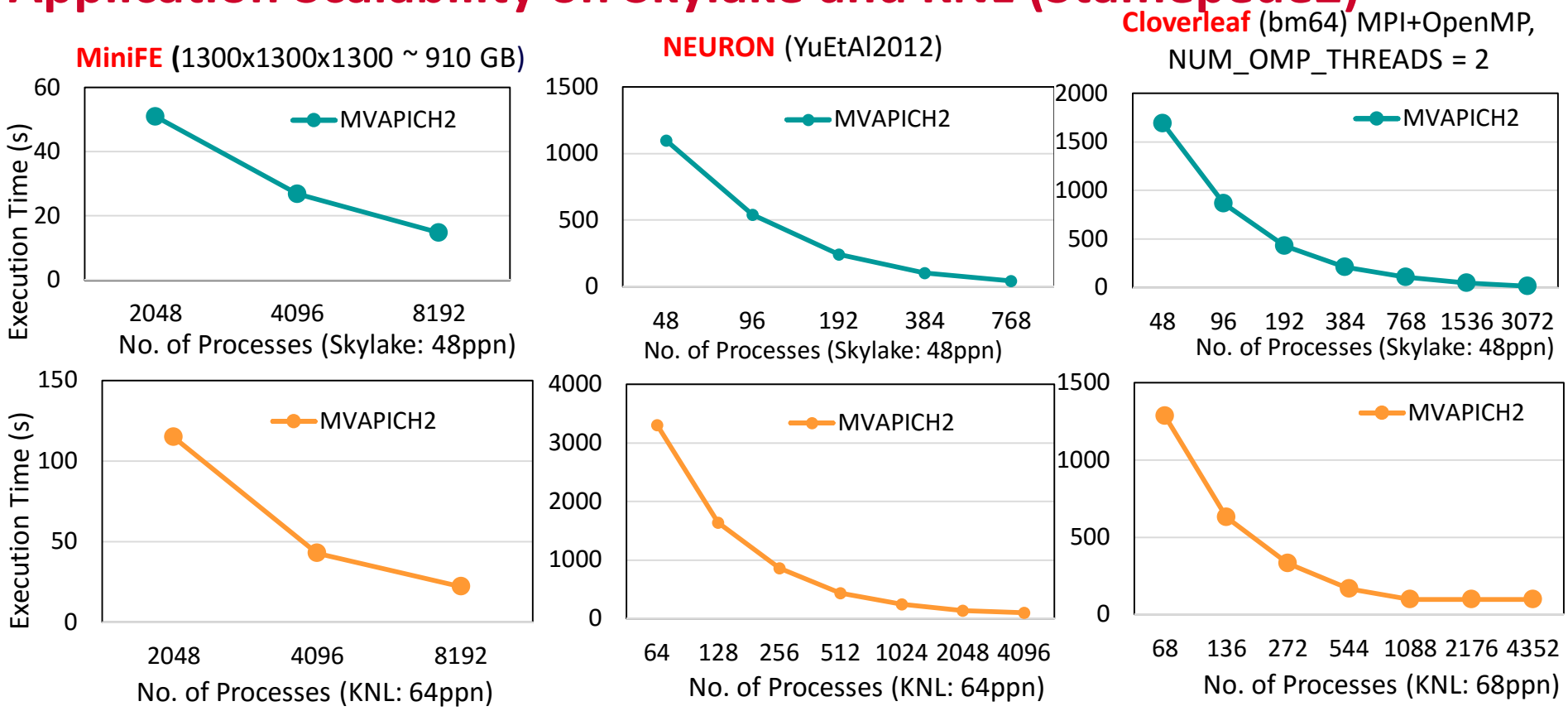
384 Processes (8 Nodes : 48 PPN)

SPEC MPI : Skylake + Omni-Path

Observations :

1. ***Up to 25 % performance improvement for SPECMPI applications on 384 processes with KNL + Omni-Path***
2. ***Up to 38 % performance improvement for SPECMPI applications on 384 processes with Skylake + Omni-Path***

Application Scalability on Skylake and KNL (Stamepede2)



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*

Concluding Remarks

- High-performance interconnects critical to the continued advancement of HPC
- Presented an overview of solutions available in MVAPICH2 for Omni-Path enabled systems
- Highlighted performance impact of proposed designs on HPC applications running on Omni-Path enabled 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

Upcoming 7th Annual MVAPICH User Group (MUG) Meeting

- **August 19-21, 2019; Columbus, Ohio, USA**
- Keynote Talks, Invited Talks, Invited Tutorials by ARM, IBM, Mellanox, Contributed Presentations, Student Poster Presentations, Tutorial on MVAPICH2 Libraries as well as other optimization and tuning hints.

- **Keynote Speakers**
  - Dan Stanzione, Texas Advanced Computing Center (TACC)
  - Robert Harrison, Director of the Institute of Advanced Computational Science (IACS) and Brookhaven Computational Science Center (CSC)
- **Tutorials** (Confirmed so far)
  - ARM
  - IBM
  - Mellanox
  - OSU/MVAPICH2

- **Invited Speakers (Confirmed so far)**
  - Gene Cooperman, Northeastern University
  - Hyon-Wook Jin, Konkuk University (South Korea)
  - Jithin Jose, Microsoft Azure
  - Minsik Kim, KISTI Supercomputing Center (South Korea)
  - Pramod Kumbhar, Blue Brain Project, EPFL (Switzerland)
  - Heechang Na, Ohio Supercomputer Center
  - Vikram Saletore, Intel
  - Jeffrey Salmond, University of Cambridge (United Kingdom)
  - Gilad Shainer, Mellanox
  - Sameer Shende, Paratools and University of Oregon
  - Sayantan Sur, Intel
  - Mahidhar Tatineni, San Diego Supercomputing Center (SDSC)
  - Karen Tomko, Ohio Supercomputer Center

More details at: http://mug.mvapich.cse.ohio-state.edu

Funding Acknowledgments

Funding Support by

Equipment Support by

Personnel Acknowledgments

Current Students (Graduate)

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

Current Students (Undergraduate)

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

Current Research Scientist

- H. Subramoni

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.)
- 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.)
- 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 Post-Docs

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

Past Research Scientist

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

Past Programmers

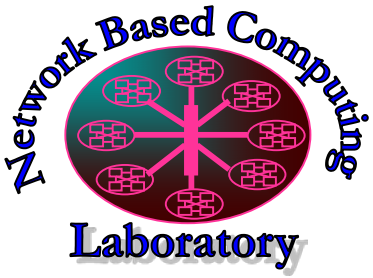
- D. Bureddy
- J. Perkins

Past Research Specialist

- M. Arnold

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/