The MVAPICH2 Project: Latest Developments and Plans Towards Exascale Computing

Presentation at the Mellanox Theater (SC '15)

by

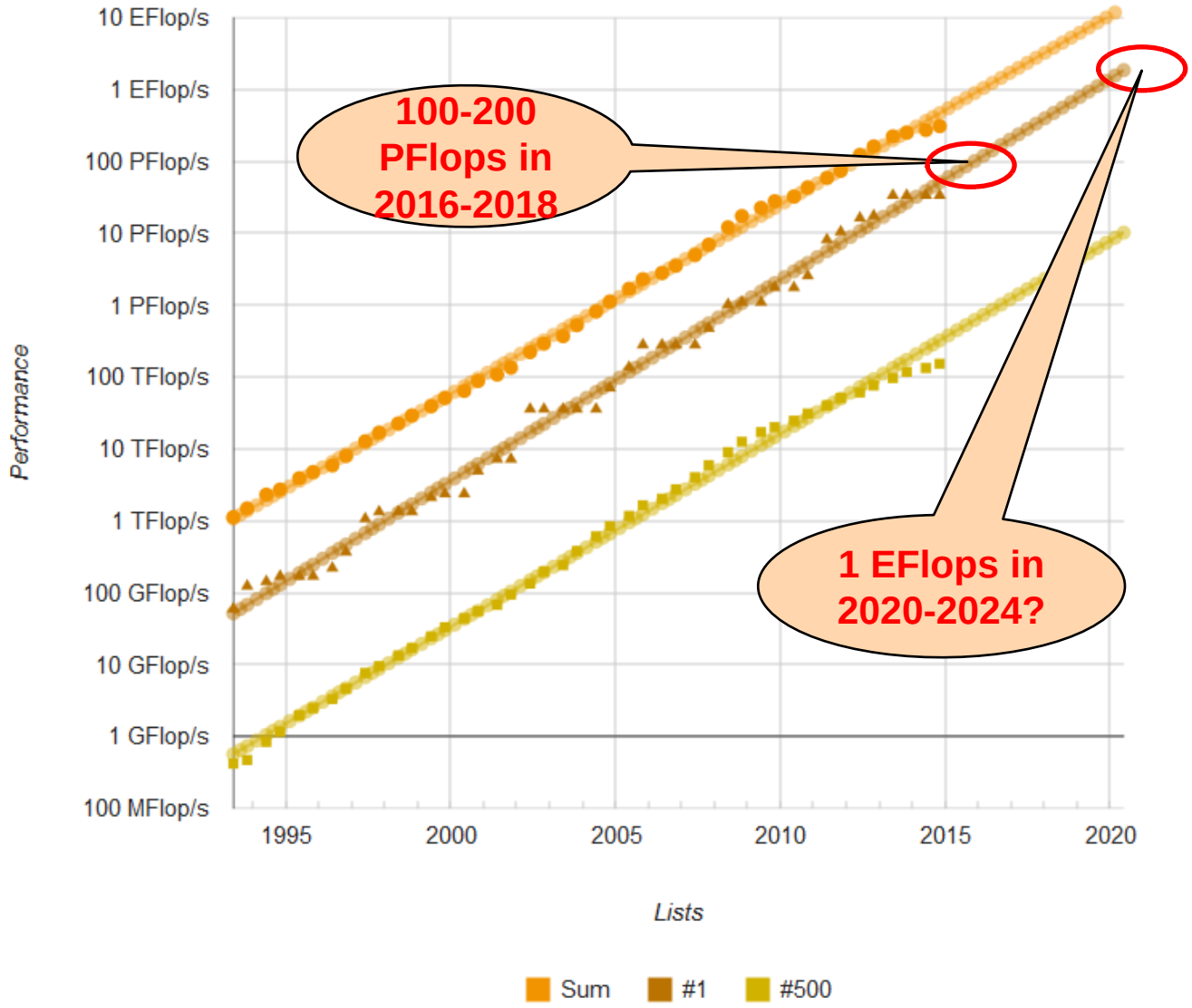
Dhabaleswar K. (DK) Panda

The Ohio State University

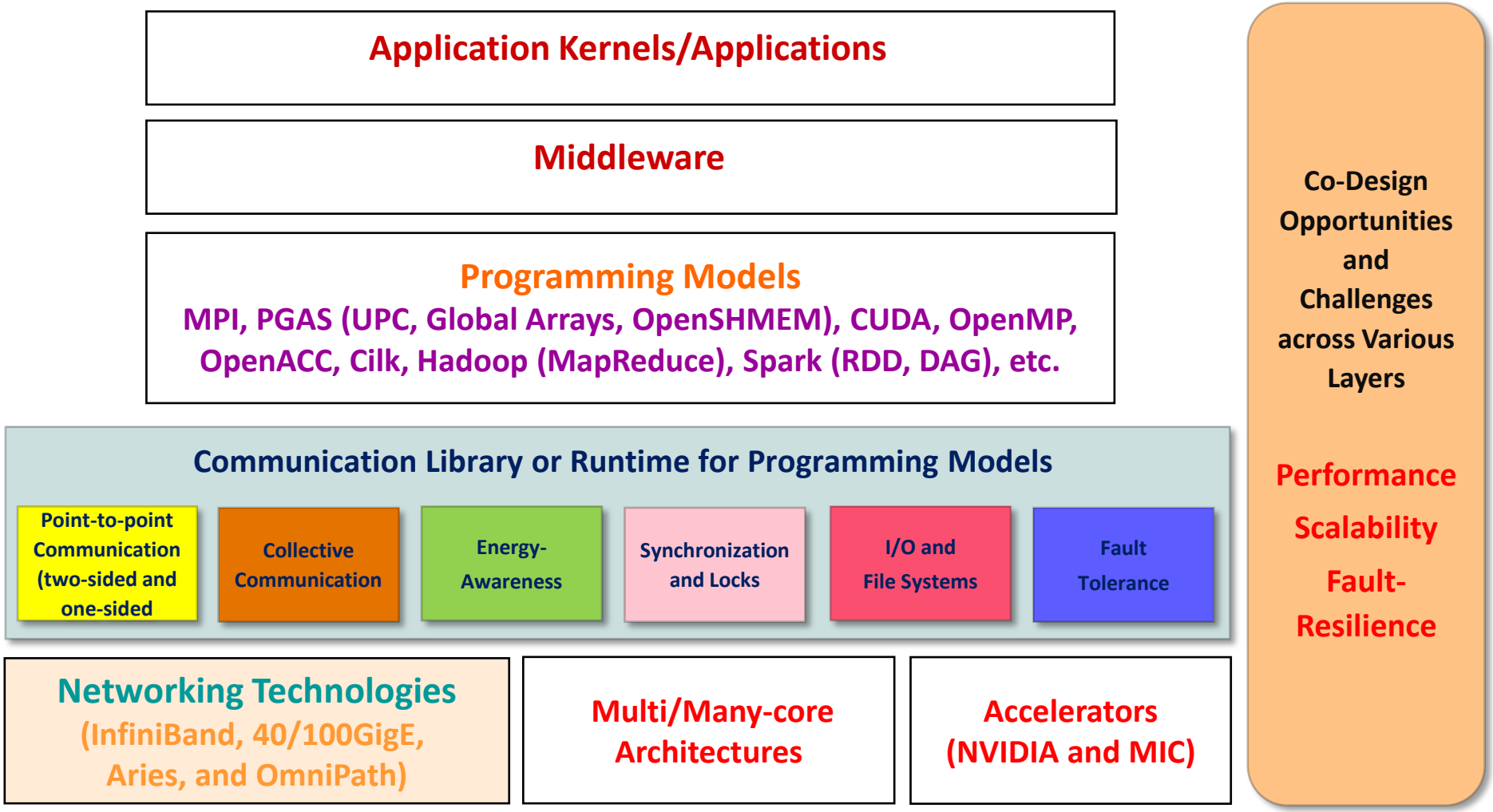
E-mail: panda@cse.ohio-state.edu

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

High-End Computing (HEC): PetaFlop to ExaFlop



Designing Communication Libraries for Multi-Petaflop and Exaflop Systems: Challenges



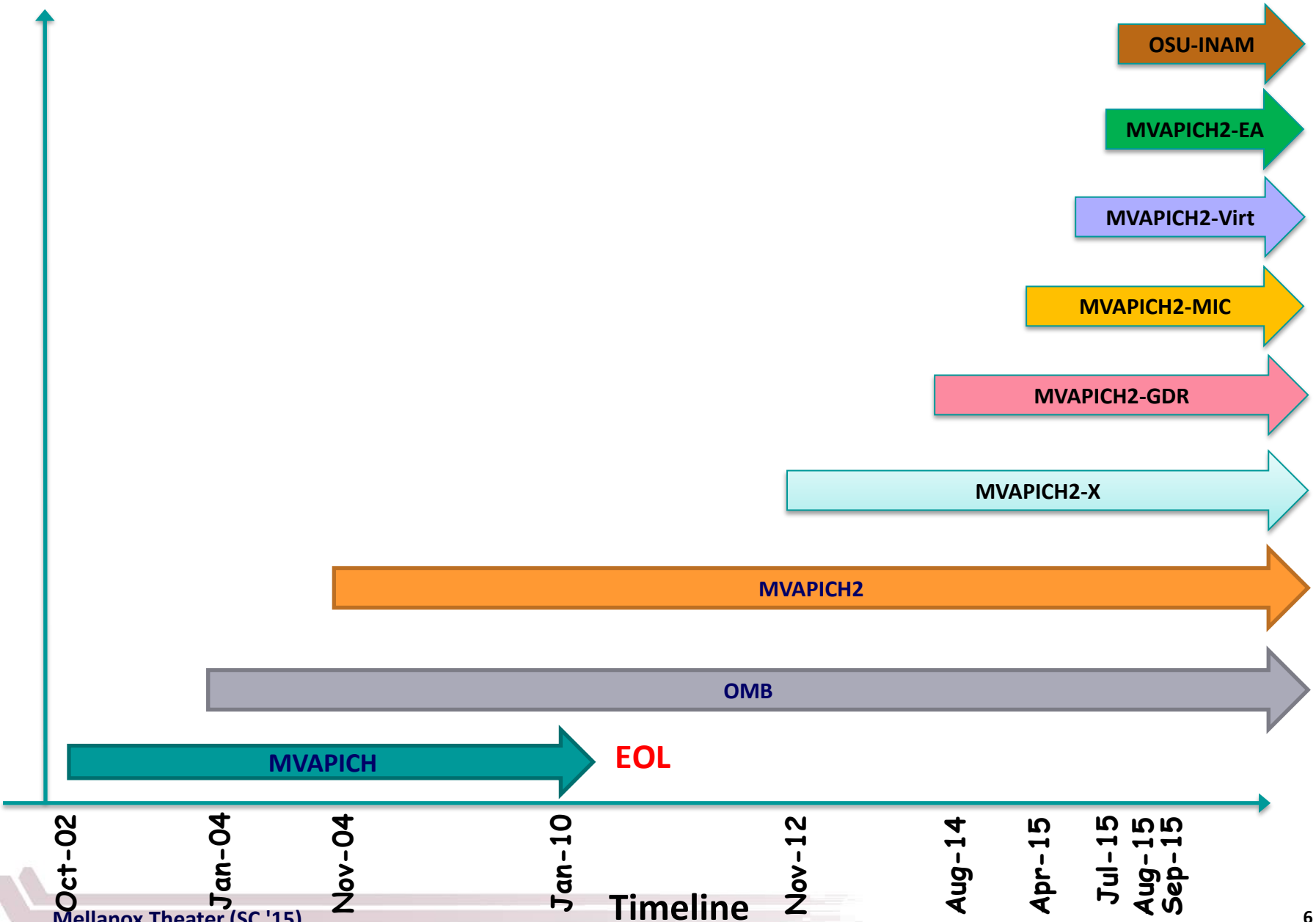
Designing (MPI+X) at Exascale

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

MVAPICH2 Software

- High Performance open-source MPI Library for InfiniBand, 10-40Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 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
  - **Used by more than 2,475 organizations in 76 countries**
  - **More than 307,000 downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '15 ranking)
    - 10th ranked 519,640-core cluster (Stampede) at TACC
    - 13th ranked 185,344-core cluster (Pleiades) at NASA
    - 25th ranked 76,032-core cluster (Tsubame 2.5) at Tokyo Institute of Technology and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- Empowering Top500 systems for over a decade
  - System-X from Virginia Tech (3rd in Nov 2003, 2,200 processors, 12.25 TFlops) ->
  - Stampede at TACC (10th in Nov'15, 519,640 cores, 5.168 Plops)

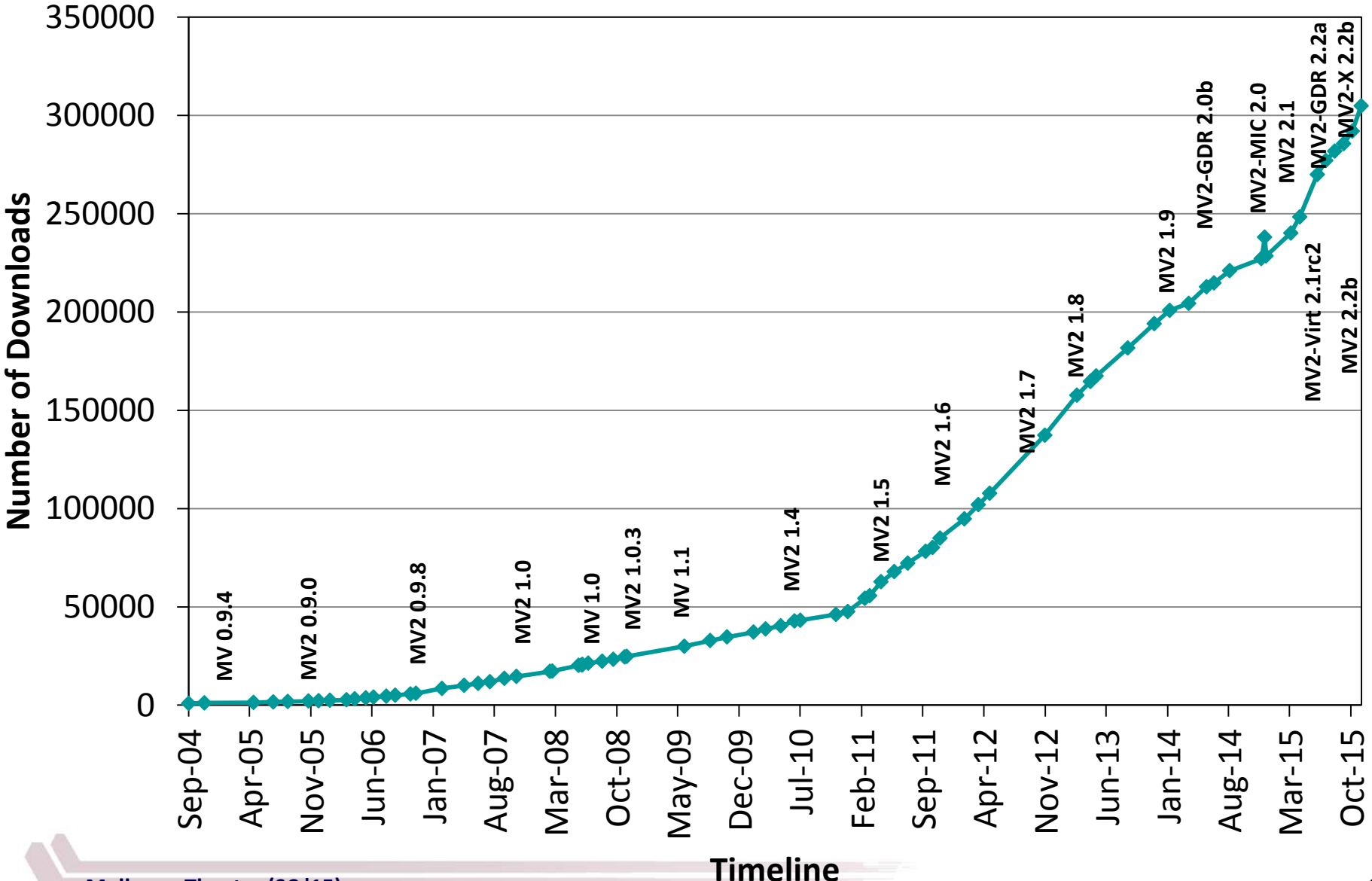
MVAPICH Project Timeline



MVAPICH2 Software Family

Requirements	MVAPICH2 Library to use
MPI with IB, iWARP and RoCE	MVAPICH2
Advanced MPI, OSU INAM, PGAS and MPI+PGAS with IB and RoCE	MVAPICH2-X
MPI with IB & GPU	MVAPICH2-GDR
MPI with IB & MIC	MVAPICH2-MIC
HPC Cloud with MPI & IB	MVAPICH2-Virt
Energy-aware MPI with IB, iWARP and RoCE	MVAPICH2-EA

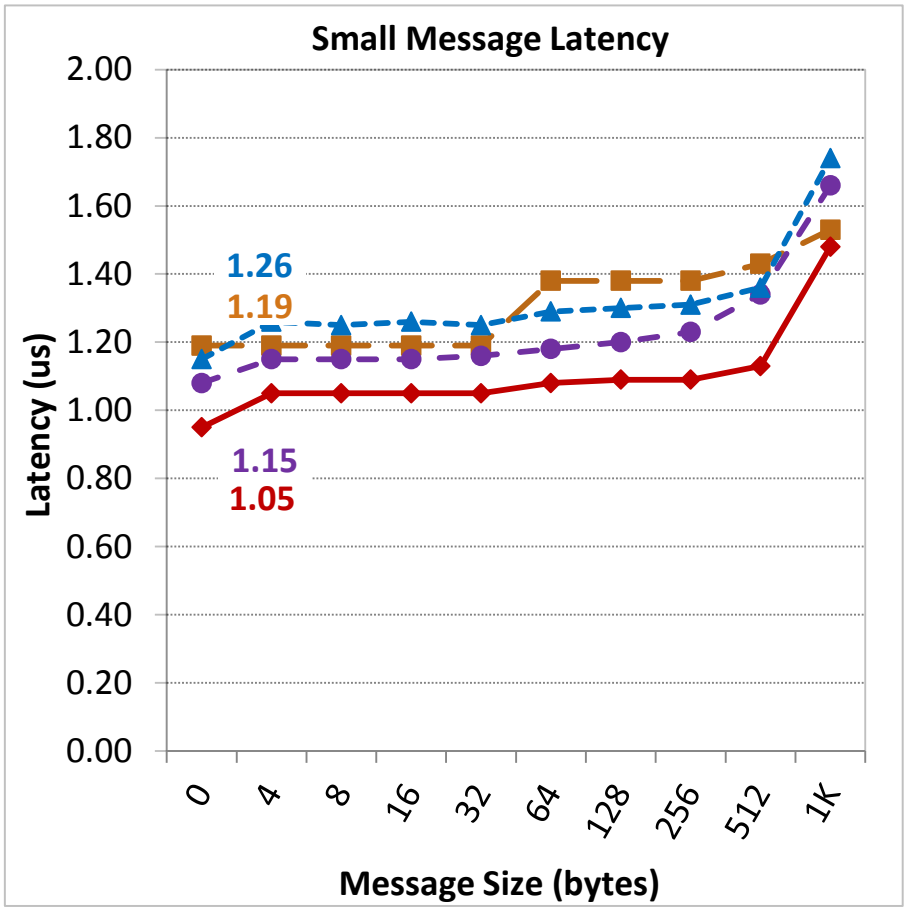
MVAPICH/MVAPICH2 Release Timeline and Downloads

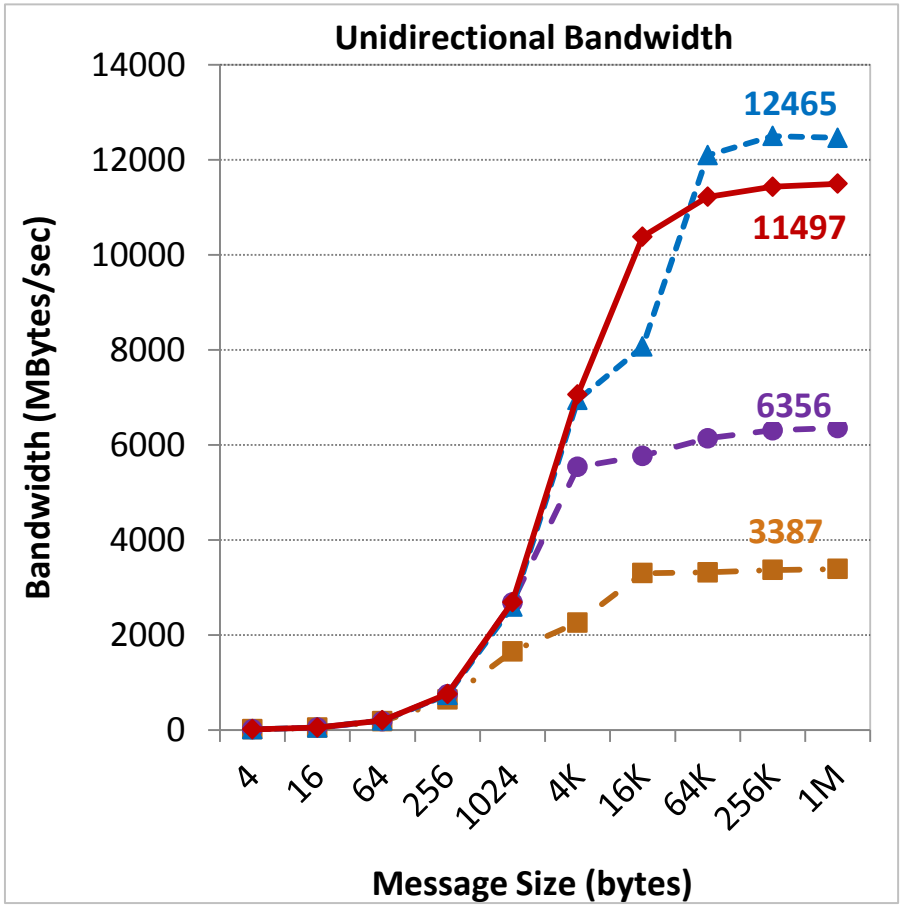


MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPIC2-MIC
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPICH2-GDR (High-Performance for GPUs) will be presented tomorrow (Thursday, Nov 18th from 11:00-11:30am)

Performance of MPI over IB with MVAPICH2





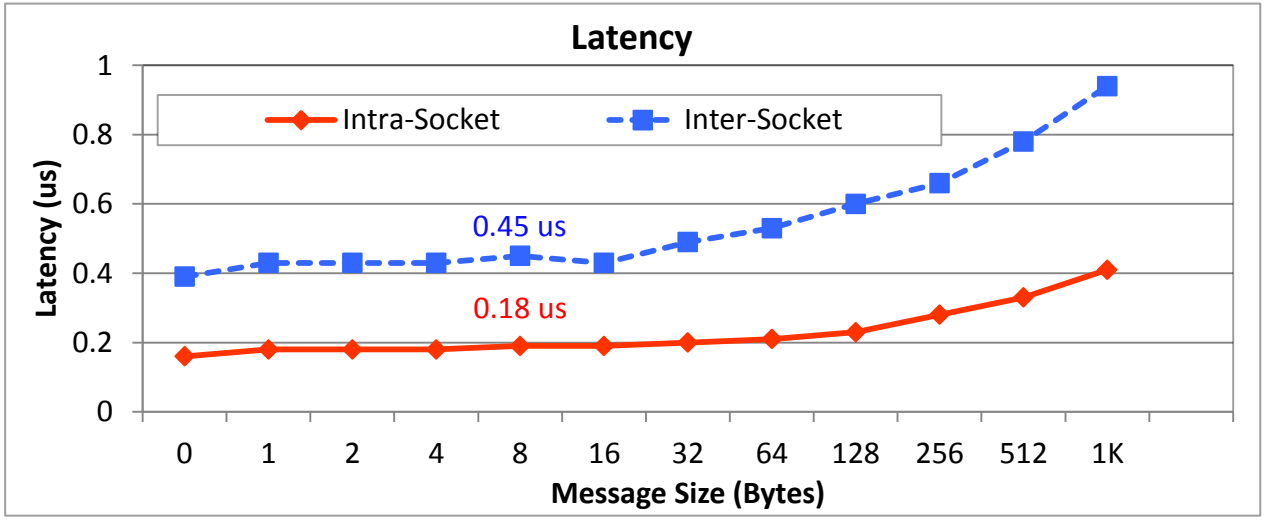
TrueScale-QDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

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

ConnectIB-Dual FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

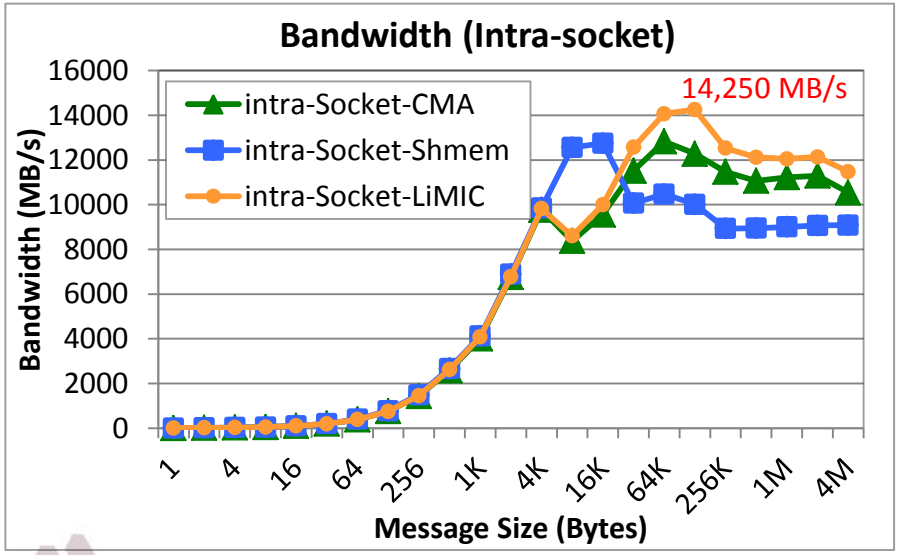
ConnectX-4-EDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 Back-to-back

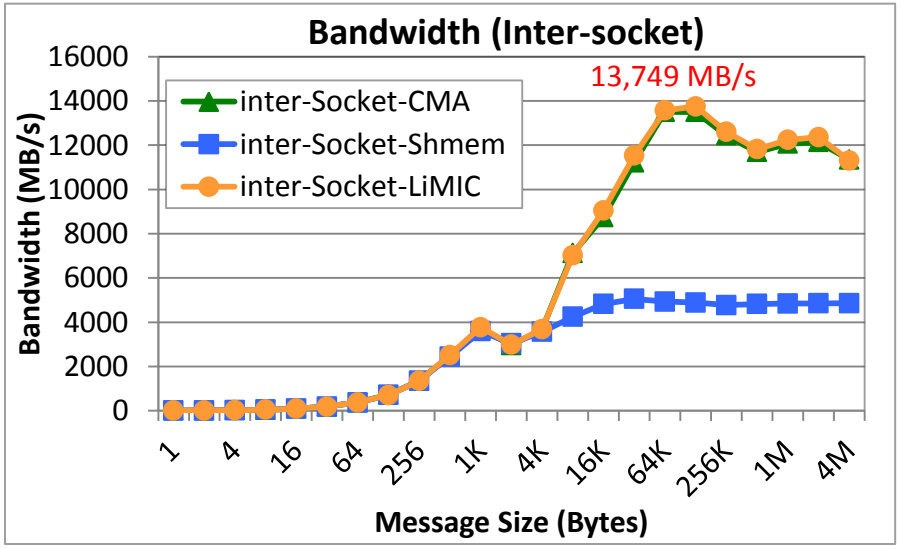
MVAPICH2 Two-Sided Intra-Node Performance
(Shared memory and Kernel-based Zero-copy Support (LiMIC and CMA))



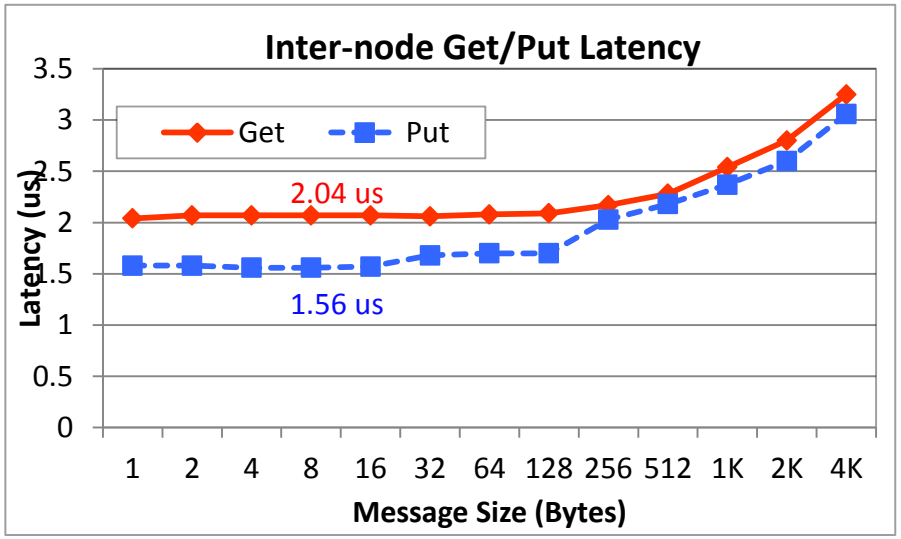
Latest MVAPICH2 2.2b

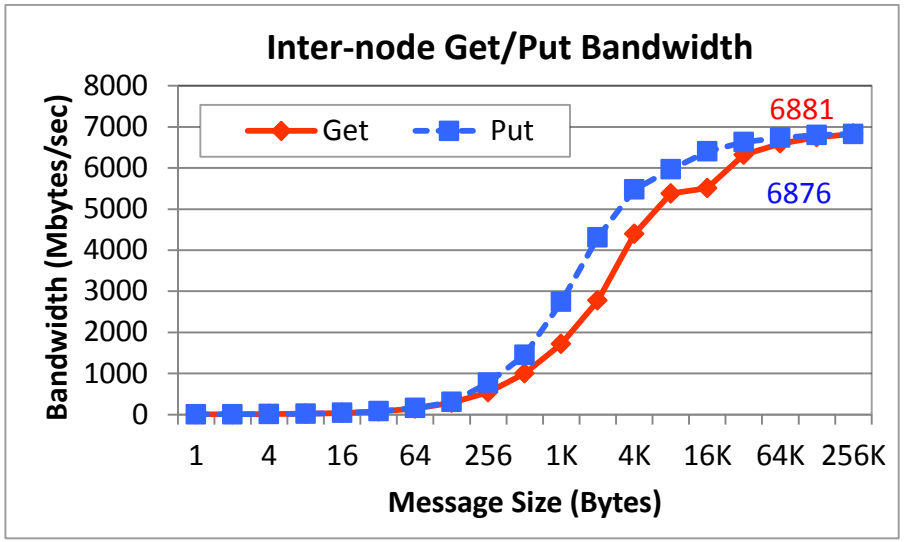
Intel Ivy-bridge

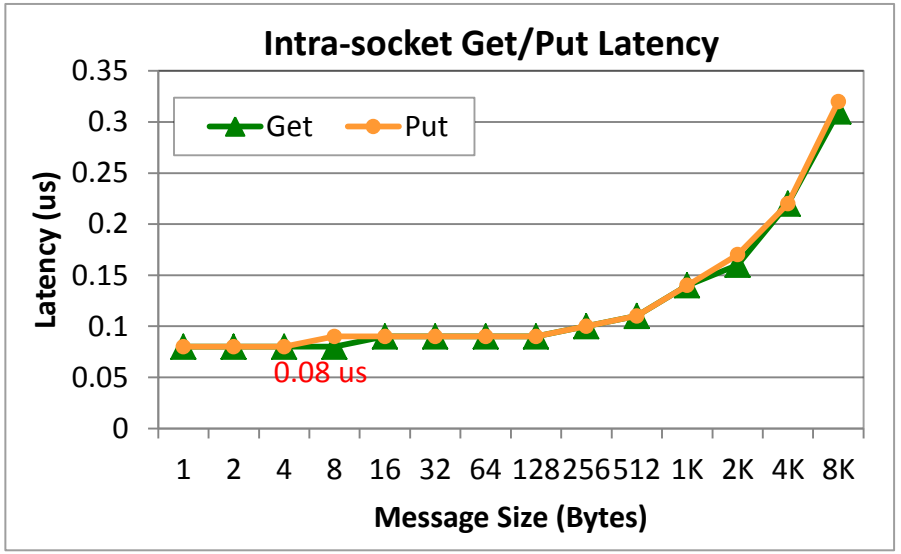


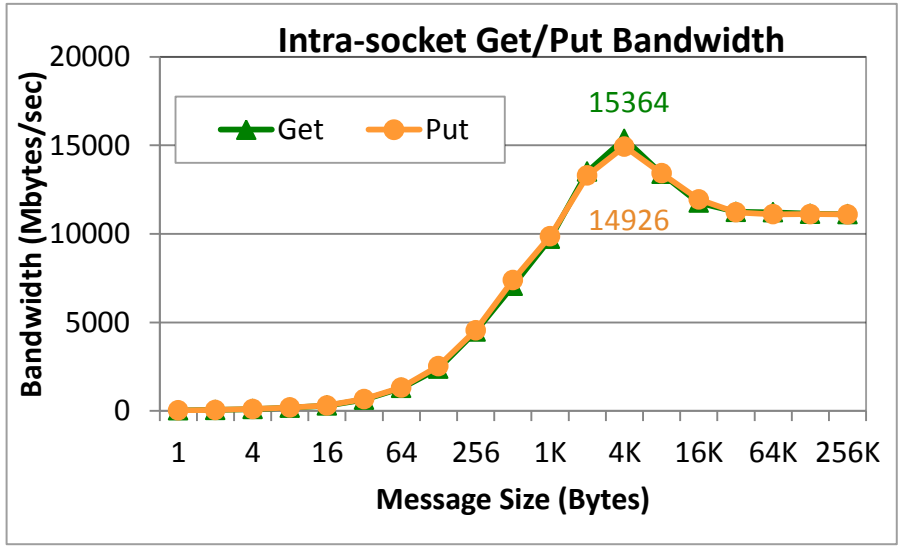


MPI-3 RMA Get/Put with Flush Performance







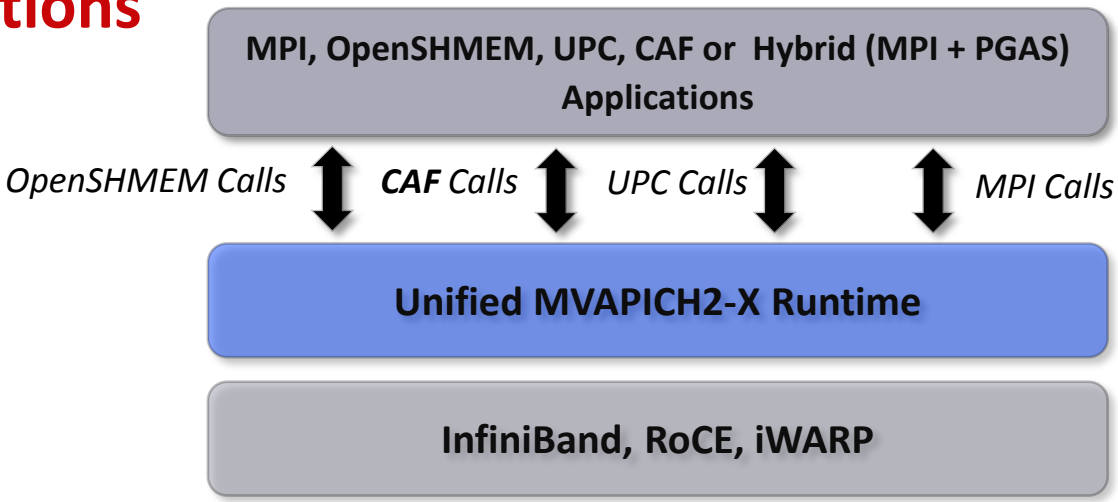


Latest MVAPICH2 2.2b, Intel Sandy-bridge with Connect-IB (single-port)

MVAPICH2 Distributions

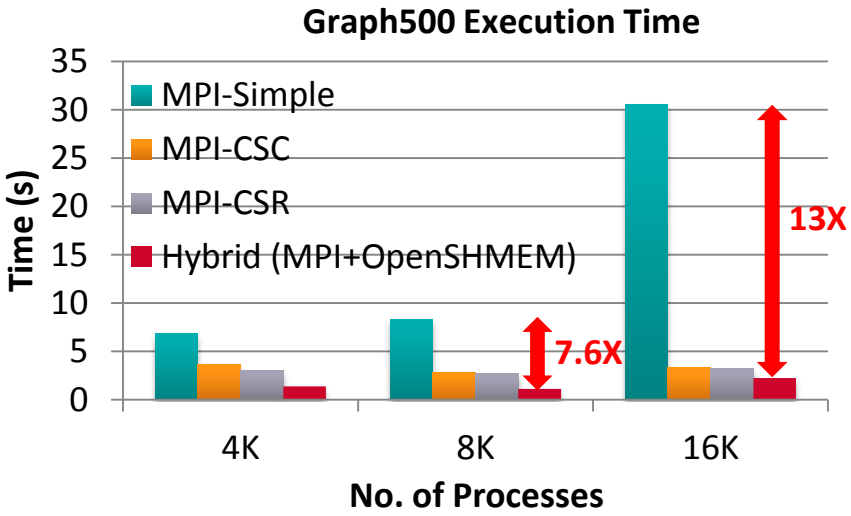
- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPIC2-MIC
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

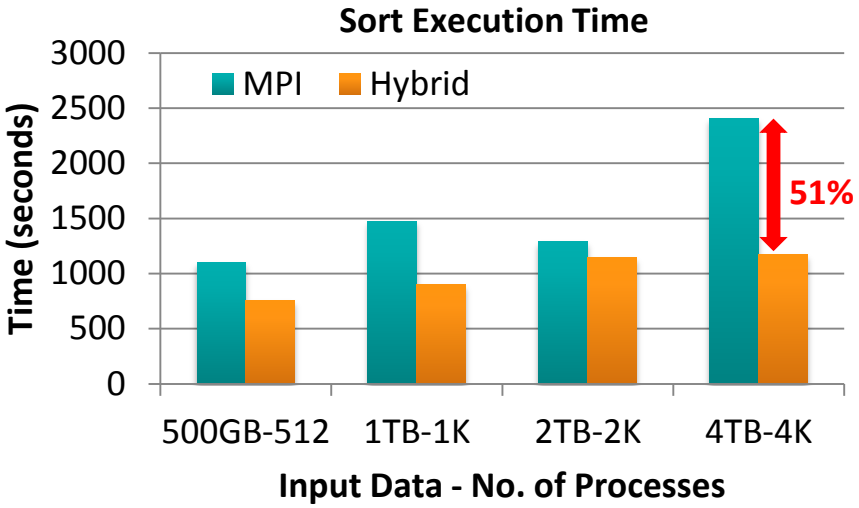
MVAPICH2-X for Advanced MPI and Hybrid MPI + PGAS Applications



- Unified communication runtime for MPI, UPC, OpenSHMEM, CAF available with MVAPICH2-X 1.9 (2012) onwards!
  - http://mvapich.cse.ohio-state.edu
- Feature Highlights
  - Supports MPI(+OpenMP), OpenSHMEM, UPC, CAF, MPI(+OpenMP) + OpenSHMEM, MPI(+OpenMP) + UPC
  - MPI-3 compliant, OpenSHMEM v1.0 standard compliant, UPC v1.2 standard compliant (with initial support for UPC 1.3), CAF 2008 standard (OpenUH)
  - Scalable Inter-node and intra-node communication – point-to-point and collectives

Application Level Performance with Graph500 and Sort





- Performance of Hybrid (MPI+ OpenSHMEM) Graph500 Design
  - 8,192 processes
    - **2.4X** improvement over MPI-CSR
    - **7.6X** improvement over MPI-Simple
  - 16,384 processes
    - **1.5X** improvement over MPI-CSR
    - **13X** improvement over MPI-Simple

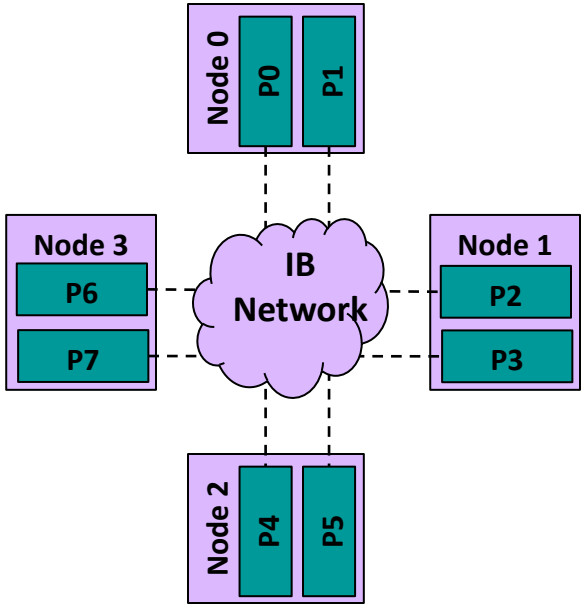
- Performance of Hybrid (MPI+OpenSHMEM) Sort Application
  - 4,096 processes, 4 TB Input Size
    - MPI – 2408 sec; 0.16 TB/min
    - Hybrid – 1172 sec; 0.36 TB/min
    - **51%** improvement over MPI-design

J. Jose, K. Kandalla, S. Potluri, J. Zhang and D. K. Panda, Optimizing Collective Communication in OpenSHMEM, Int'l Conference on Partitioned Global Address Space Programming Models (PGAS '13), October 2013.

J. Jose, S. Potluri, K. Tomko and D. K. Panda, Designing Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models, International Supercomputing Conference (ISC'13), June 2013

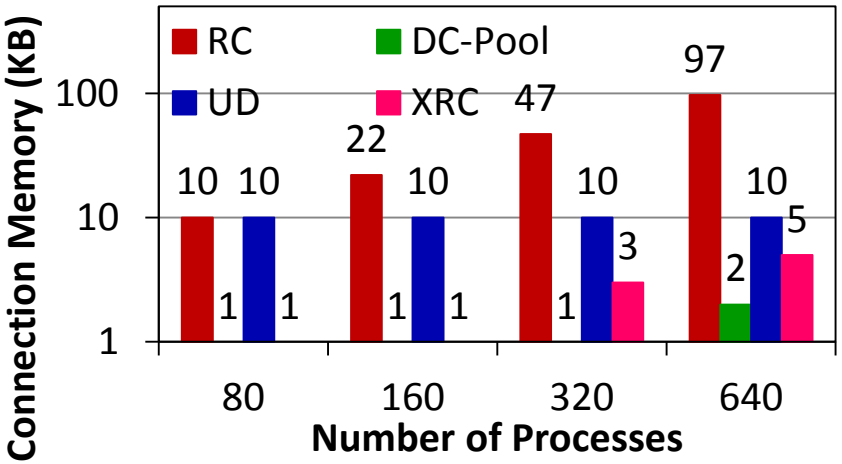
J. Jose, K. Kandalla, M. Luo and D. K. Panda, Supporting Hybrid MPI and OpenSHMEM over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP '12), September 2012

Minimizing Memory Footprint further by DC Transport

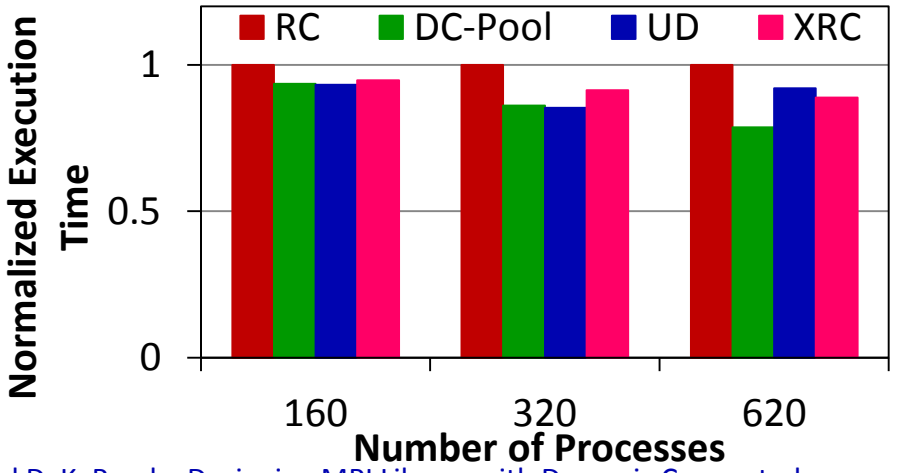


- Constant connection cost (*One QP for any peer*)
- Full Feature Set (RDMA, Atomics etc)
- Separate objects for send (DC Initiator) and receive (DC Target)
  - DC Target identified by "DCT Number"
  - Messages routed with (DCT Number, LID)
  - Requires same "DC Key" to enable communication
- Available with MVAPICH2-X 2.2a

Memory Footprint for Alltoall



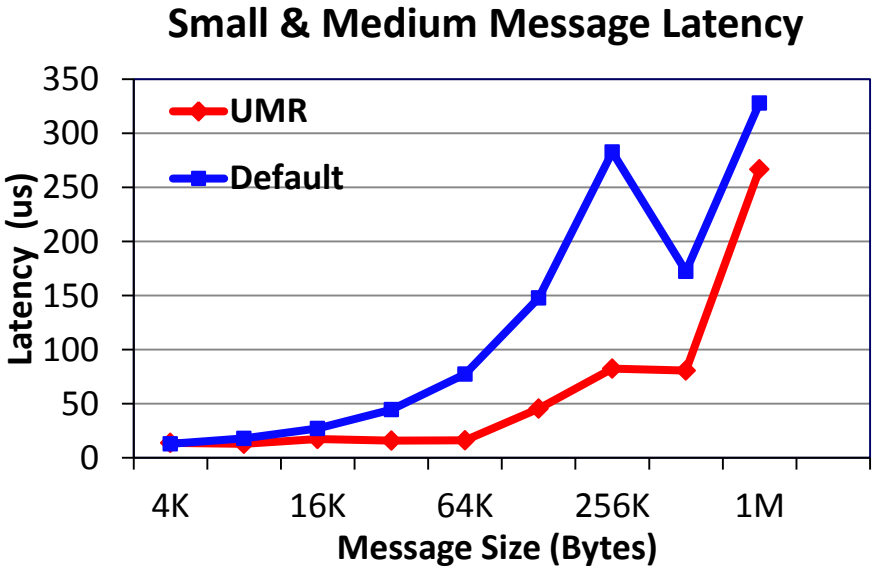
NAMD - Apoa1: Large data set

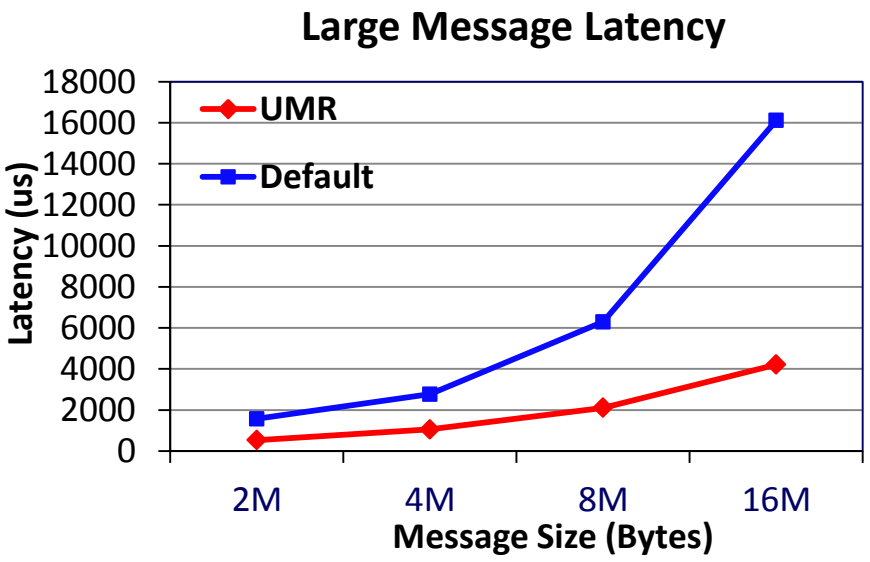


H. Subramoni, K. Hamidouche, A. Venkatesh, S. Chakraborty and D. K. Panda, Designing MPI Library with Dynamic Connected Transport (DCT) of InfiniBand : Early Experiences. IEEE International Supercomputing Conference (ISC '14).

User-mode Memory Registration (UMR)

- Introduced by Mellanox to support direct local and remote noncontiguous memory access
- Avoid packing at sender and unpacking at receiver
- Available in MVAPICH2-X 2.2b





Connect-IB (54 Gbps): 2.8 GHz Dual Ten-core (IvyBridge) Intel PCI Gen3 with Mellanox IB FDR switch

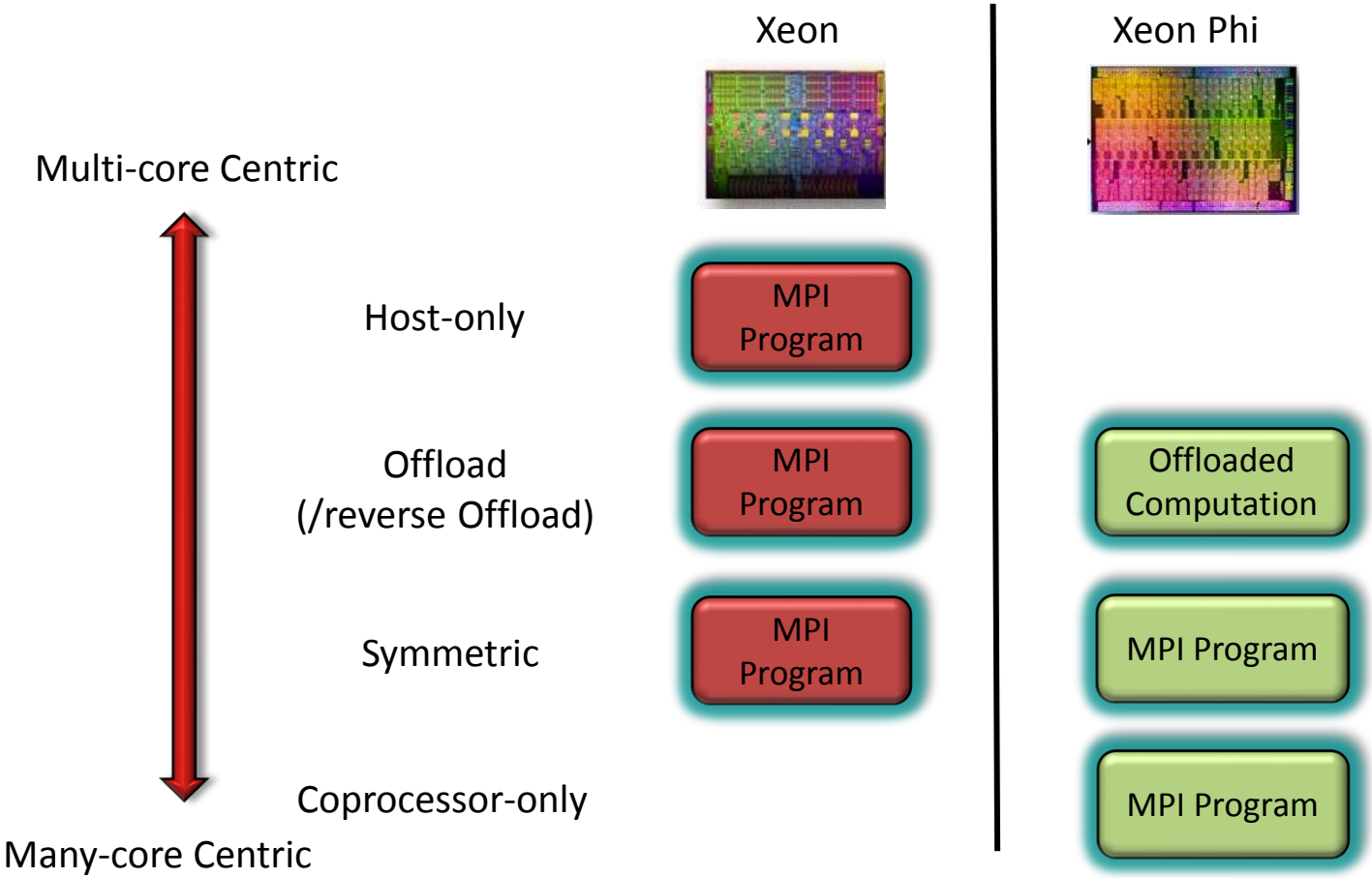
M. Li, H. Subramoni, K. Hamidouche, X. Lu and D. K. Panda, "High Performance MPI Datatype Support with User-mode Memory Registration: Challenges, Designs and Benefits", CLUSTER, 2015

MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPIC2-MIC
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

MPI Applications on MIC Clusters

- Flexibility in launching MPI jobs on clusters with Xeon Phi

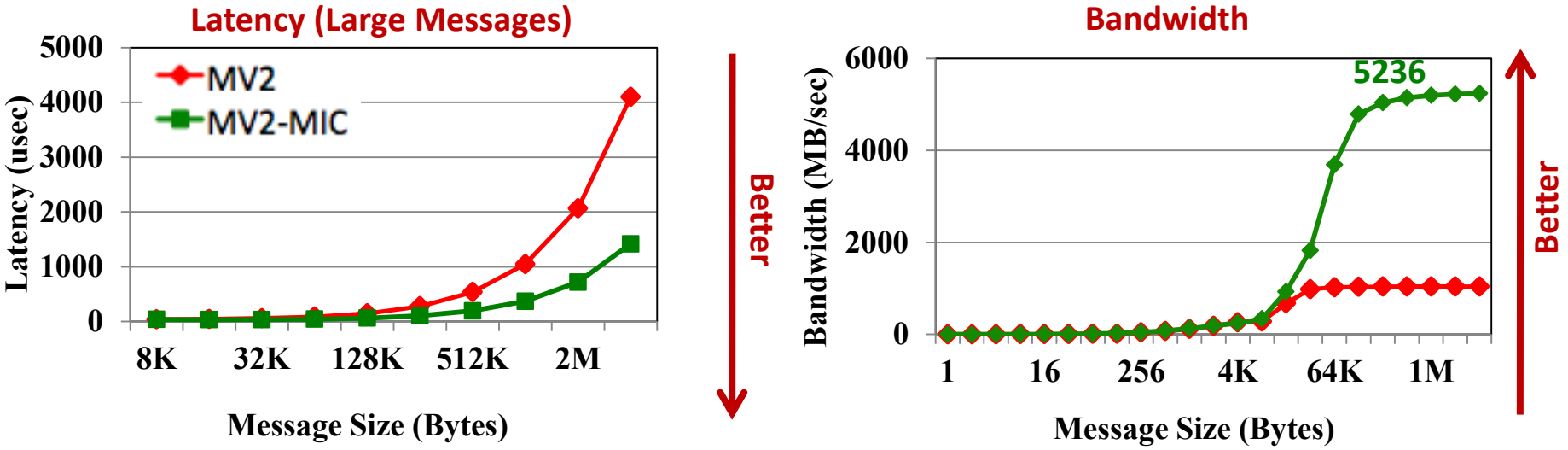


MVAPICH2-MIC 2.0 Design for Clusters with IB and MIC

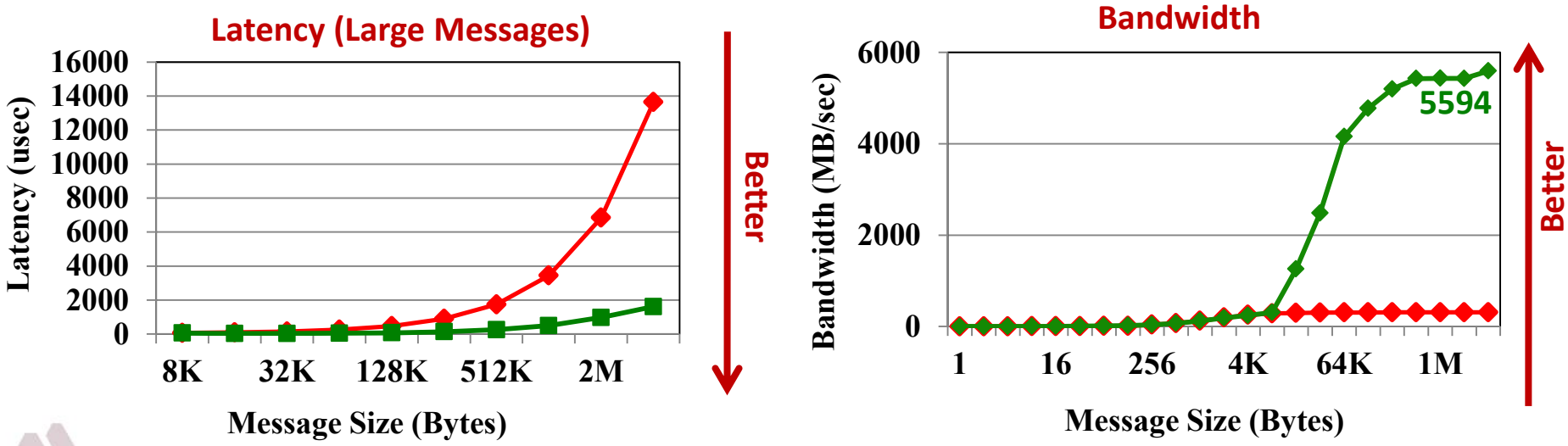
- Offload Mode
- Intranode Communication
  - Coprocessor-only and Symmetric Mode
- Internode Communication
  - Coprocessors-only and Symmetric Mode
- Multi-MIC Node Configurations
- Running on three major systems
  - Stampede, Blueridge (Virginia Tech) and Beacon (UTK)

MIC-Remote-MIC P2P Communication with Proxy-based Communication

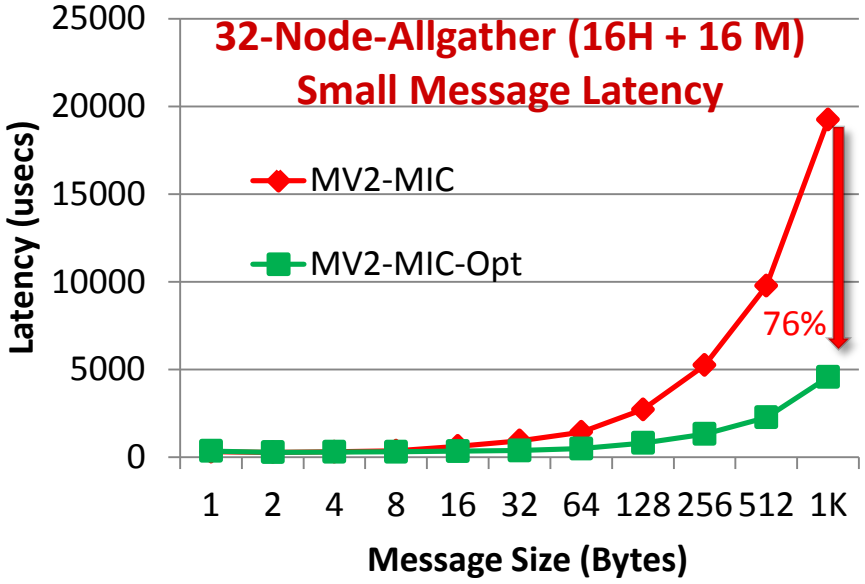
Intra-socket P2P

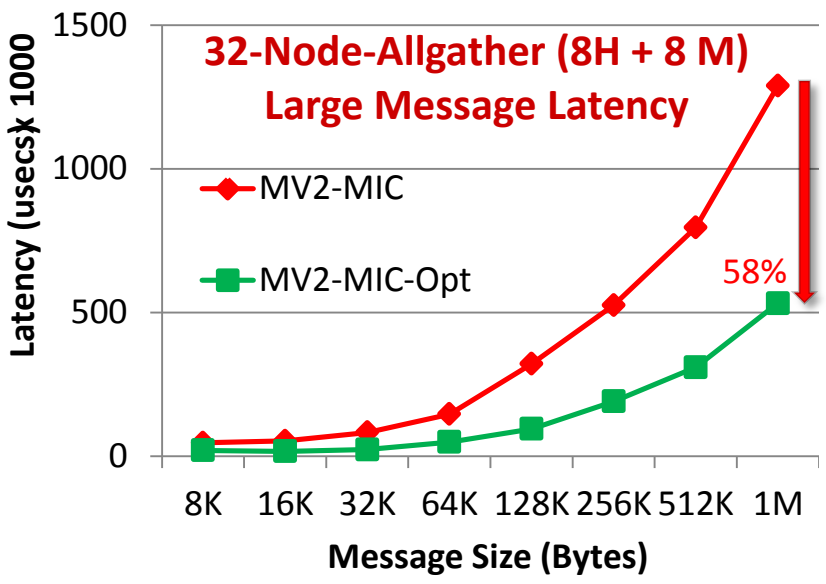


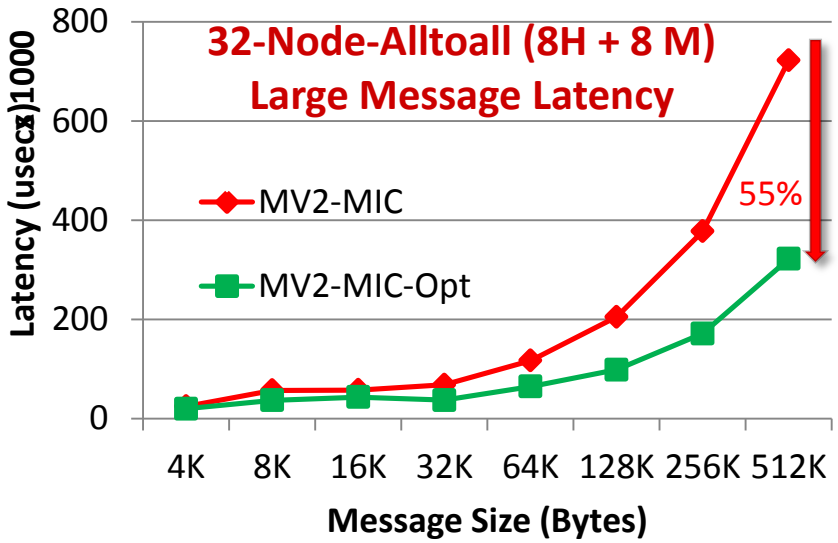
Inter-socket P2P

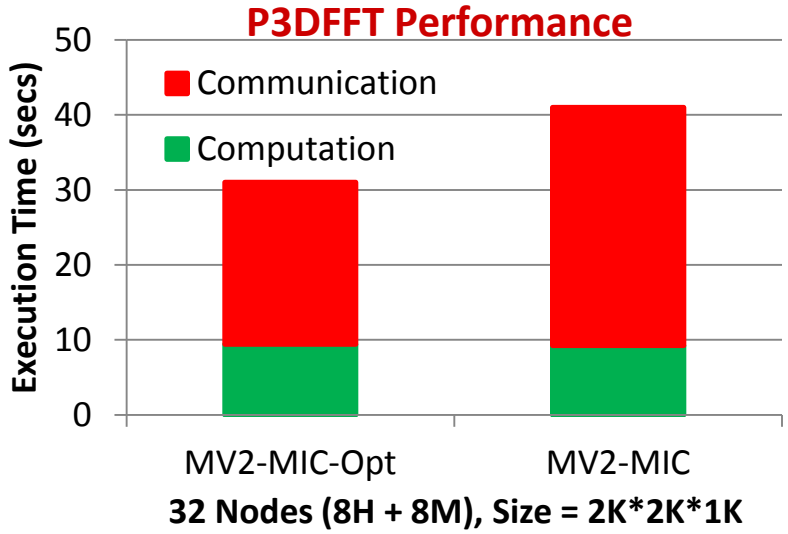


Optimized MPI Collectives for MIC Clusters (Allgather & Alltoall)









A. Venkatesh, S. Potluri, R. Rajachandrasekar, M. Luo, K. Hamidouche and D. K. Panda - High Performance Alltoall and Allgather designs for InfiniBand MIC Clusters; IPDPS'14, May 2014

MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPIC2-MIC
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

Can HPC and Virtualization be Combined?

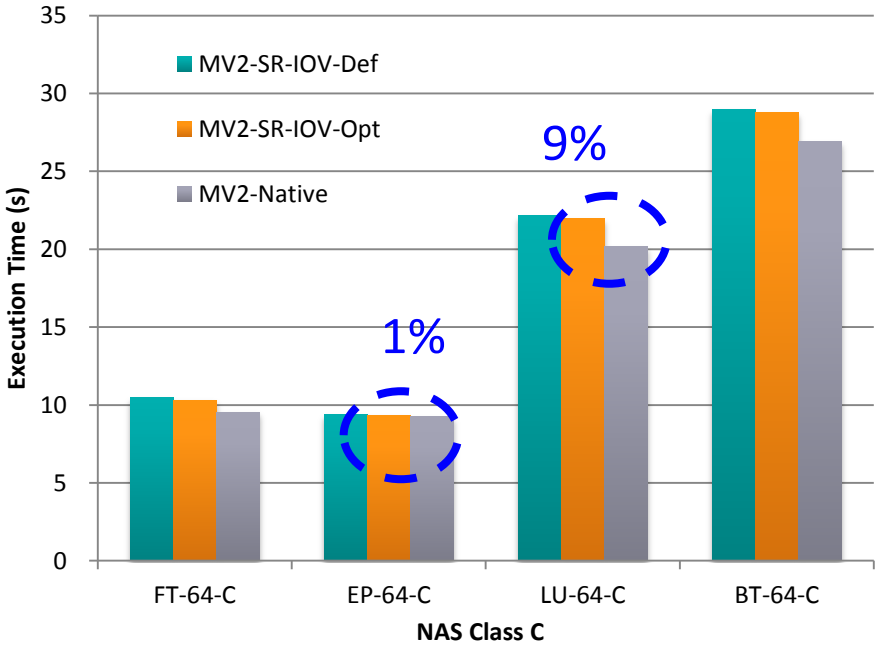
- Virtualization has many benefits
  - Fault-tolerance
  - Job migration
  - Compaction
- Have not been very popular in HPC due to overhead associated with Virtualization
- New SR-IOV (Single Root – IO Virtualization) support available with Mellanox InfiniBand adapters changes the field
- Enhanced MVAPICH2 support for SR-IOV
- MVAPICH2-Virt 2.1 (with and without OpenStack)

J. Zhang, X. Lu, J. Jose, R. Shi and D. K. Panda, Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? EuroPar'14

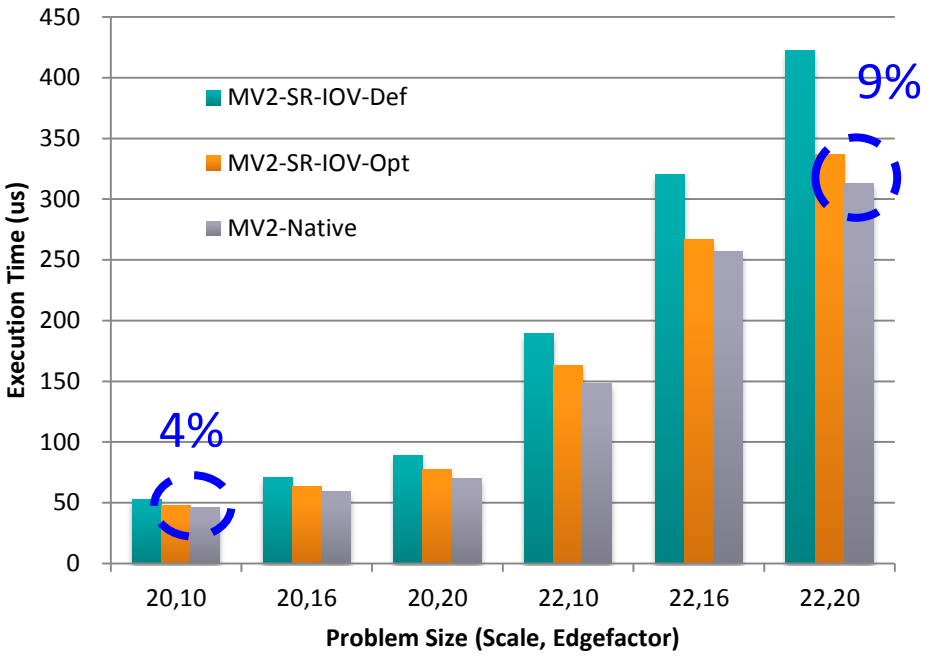
J. Zhang, X. Lu, J. Jose, M. Li, R. Shi and D.K. Panda, High Performance MPI Libray over SR-IOV enabled InfiniBand Clysters, HiPC'14

J. Zhang, X .Lu, M. Arnold and D. K. Panda, MVAPICH2 Over OpenStack with SR-IOV: an Efficient Approach to build HPC Clouds, CCGrid'15

Application-Level Performance (8 VM * 8 Core/VM)



NAS



Graph500

- Compared to Native, 1-9% overhead for NAS
- Compared to Native, 4-9% overhead for Graph500

NSF Chameleon Cloud: A Powerful and Flexible Experimental Instrument

- Large-scale instrument
  - Targeting Big Data, Big Compute, Big Instrument research
  - ~650 nodes (~14,500 cores), 5 PB disk over two sites, 2 sites connected with 100G network
- Reconfigurable instrument
  - Bare metal reconfiguration, operated as single instrument, graduated approach for ease-of-use
- Connected instrument
  - Workload and Trace Archive
  - Partnerships with production clouds: CERN, OSDC, Rackspace, Google, and others
  - Partnerships with users
- Complementary instrument
  - Complementing GENI, Grid'5000, and other testbeds
- Sustainable instrument
  - Industry connections

http://www.chameleoncloud.org/

TACC iCAIR

UTSA

MVAPICH2 Distributions

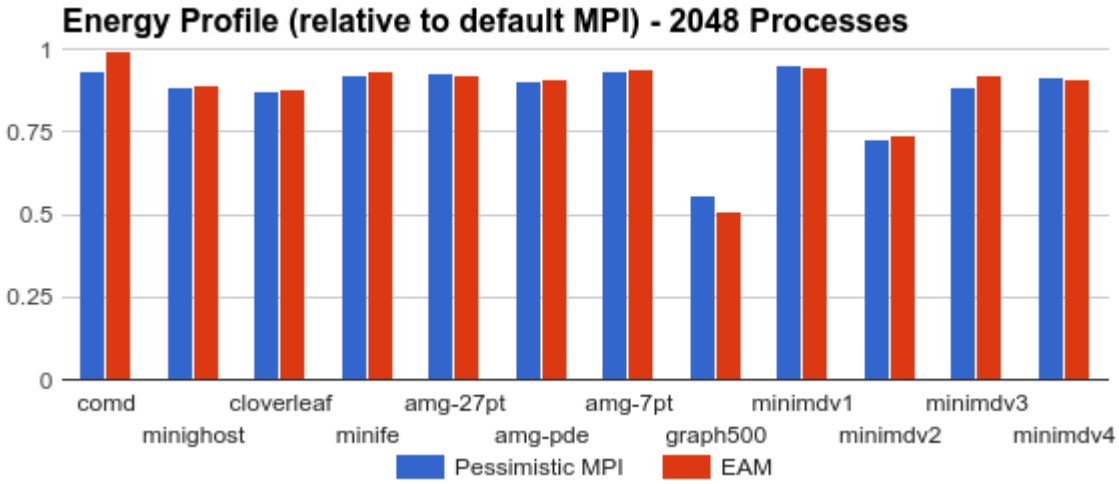
- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPIC2-MIC
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

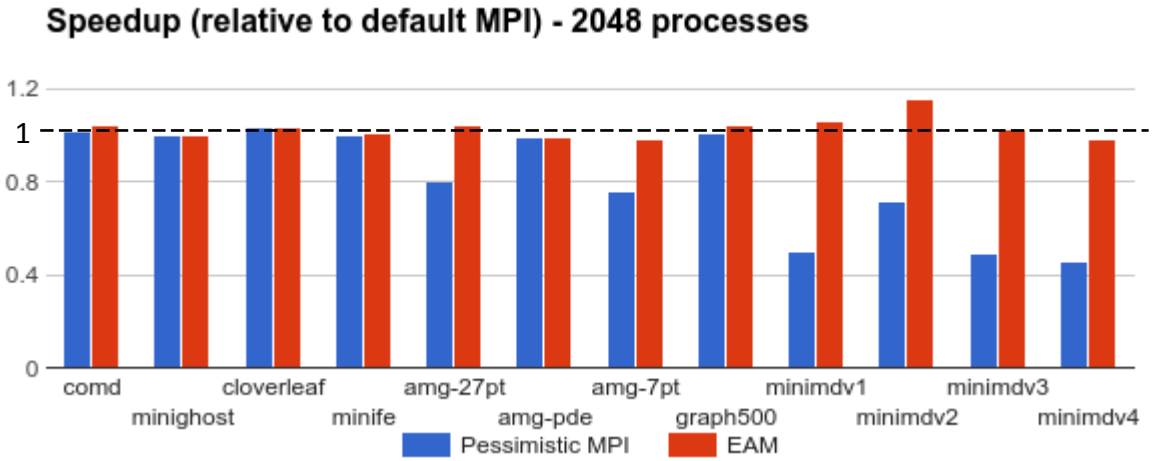
Energy-Aware MVAPICH2 Library and OSU Energy Management Tool (OEMT)

- MVAPICH2-EA (Energy-Aware) MPI Library
  - Production-ready Energy-Aware MPI Library
  - New Energy-Efficient communication protocols for pt-pt and collective operations
  - Intelligently apply the appropriate energy saving techniques
  - Application oblivious energy saving
  - Released 08/28/15
- OEMT
  - A library utility to measure energy consumption for MPI applications
  - Works with all MPI runtimes
  - PRELOAD option for precompiled applications
  - Does not require ROOT permission:
    - A safe kernel module to read only a subset of MSRs
- Available from: http://mvapich.cse.ohio-state.edu

MVAPICH2-EA: Application Oblivious Energy-Aware-MPI (EAM)

- An energy efficient runtime that provides energy savings without application knowledge
- A **white-box** approach
- **Automatically and transparently** use the best energy lever
- Provides **guarantees on maximum degradation** with 5-41% savings at <= 5% degradation
- Pessimistic MPI applies energy reduction lever to each MPI call



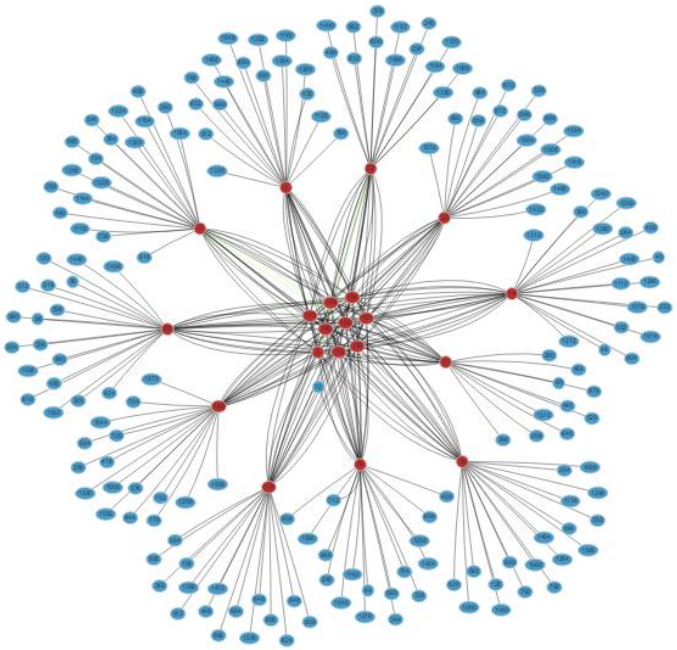


A Case for Application-Oblivious Energy-Efficient MPI Runtime A. Venkatesh , A. Vishnu , K. Hamidouche , N. Tallent , D. K. Panda , D. Kerbyson , and A. Hoise - Supercomputing '15, Nov 2015 , *Best Student Paper Finalist, presented in the Technical Papers Program, Tuesday 3:30-4:00pm (Room 18CD)*

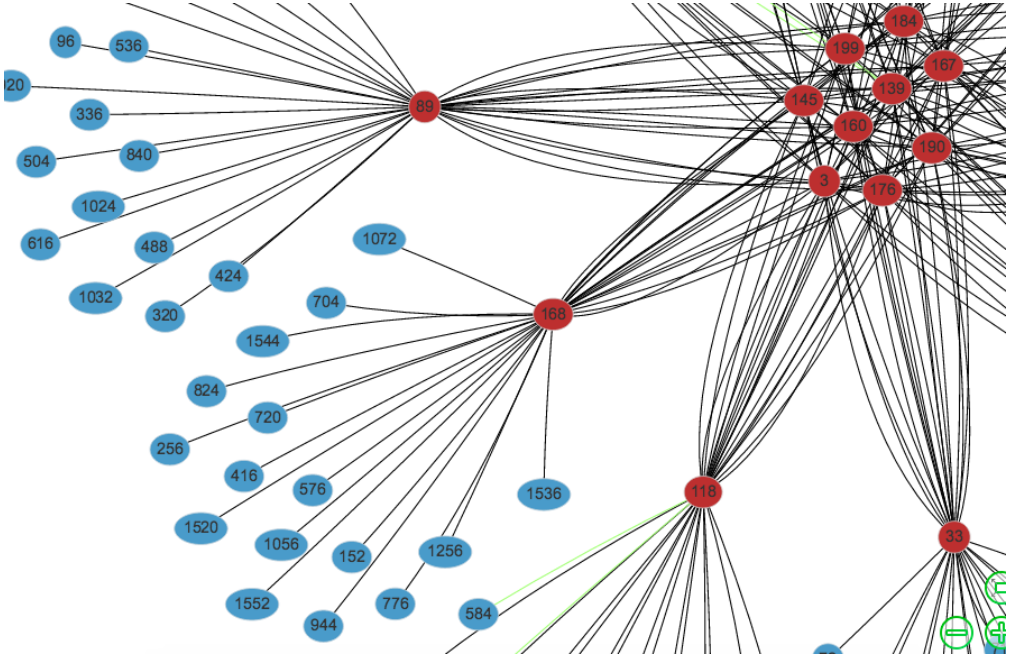
MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPIC2-MIC
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

OSU InfiniBand Network Analysis Monitoring Tool (INAM) – Network Level View



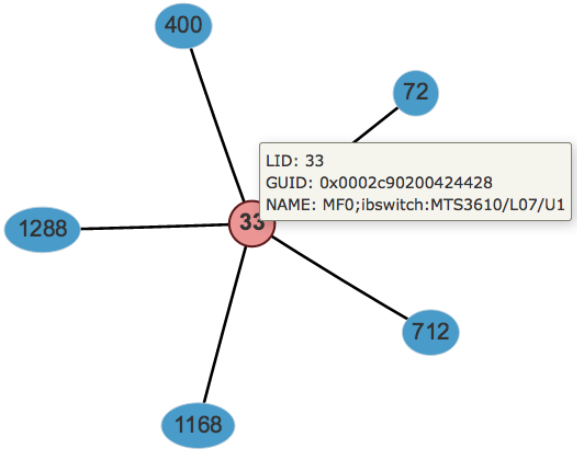
Full Network (152 nodes)



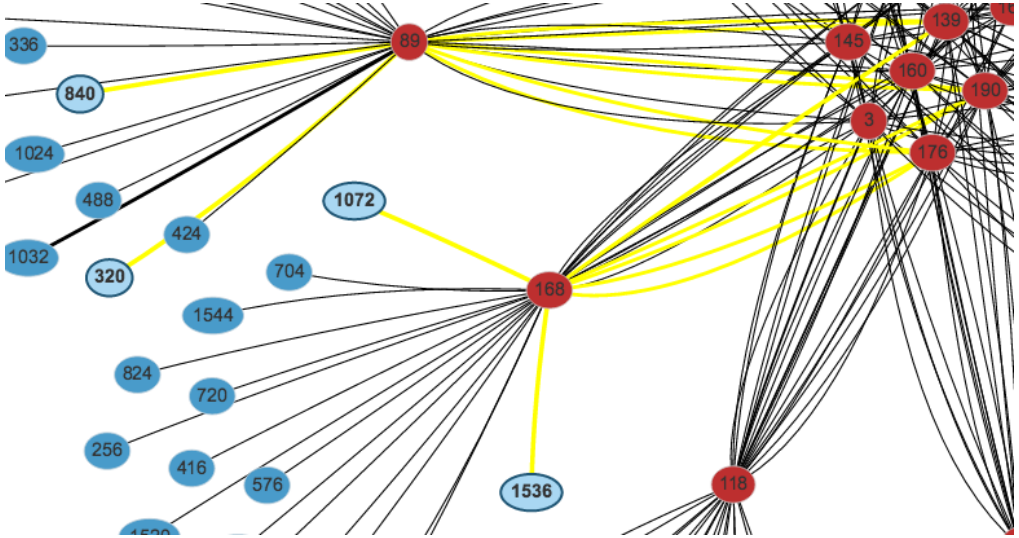
Zoomed-in View of the Network

- Show network topology of large clusters
- Visualize traffic pattern on different links
- Quickly identify congested links/links in error state
- See the history unfold – play back historical state of the network

OSU INAM Tool – Job and Node Level Views



Visualizing a Job (5 Nodes)



Finding Routes Between Nodes

- Job level view
  - Show different network metrics (load, error, etc.) for any live job
  - Play back historical data for completed jobs to identify bottlenecks
- Node level view provides details per process or per node
  - CPU utilization for each rank/node
  - Bytes sent/received for MPI operations (pt-to-pt, collective, RMA)
  - Network metrics (e.g. XmitDiscard, RcvError) per rank/node

MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 500K-1M cores
  - Dynamically Connected Transport (DCT) service with Connect-IB
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - Support for UPC++
- Enhanced Optimization for GPU Support and Accelerators
- Taking advantage of advanced features
  - User Mode Memory Registration (UMR)
  - On-demand Paging
- Enhanced Inter-node and Intra-node communication schemes for upcoming OmniPath enabled Knights Landing architectures
- Extended RMA support (as in MPI 3.0)
- Extended topology-aware collectives
- Energy-aware point-to-point (one-sided and two-sided) and collectives
- Extended Support for MPI Tools Interface (as in MPI 3.0)
- Extended Checkpoint-Restart and migration support with SCR

Two Additional Talks

- **Today, Wednesday (2:30-3:00pm)**
  - **High Performance Big Data (HiBD): Accelerating Hadoop, Spark and Memcached on Modern Clusters**
- **Tomorrow, Thursday (11:00-11:30am)**
  - **Exploiting Full Potential of GPU Clusters with InfiniBand using MVAPICH2-GDR**

Funding Acknowledgments

Funding Support by

Equipment Support by

Personnel Acknowledgments

Current Students

- A. Augustine (M.S.)
- A. Awan (Ph.D.)
- S. Chakraborthy (Ph.D.)
- C.-H. Chu (Ph.D.)
- N. Islam (Ph.D.)
- M. Li (Ph.D.)
- K. Kulkarni (M.S.)
- M. Rahman (Ph.D.)
- D. Shankar (Ph.D.)
- A. Venkatesh (Ph.D.)
- J. Zhang (Ph.D.)

Current Research Scientists

- H. Subramoni
- X. Lu

Current Senior Research Associate

- K. Hamidouche

Current Post-Doc

- J. Lin
- D. Banerjee

Current Programmer

- J. Perkins

Current Research Specialist

- M. Arnold

Past Students

- P. Balaji (Ph.D.)
- 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.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (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.)

Past Post-Docs

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

Past Research Scientist

- S. Sur

Past Programmers

- D. Bureddy

Web Pointers

NOWLAB Web Page

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

MVAPICH Web Page

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