# MVAPICH2 and MVAPICH2-MIC: Latest Status

Presentation at IXPUG Meeting, July 2014

by

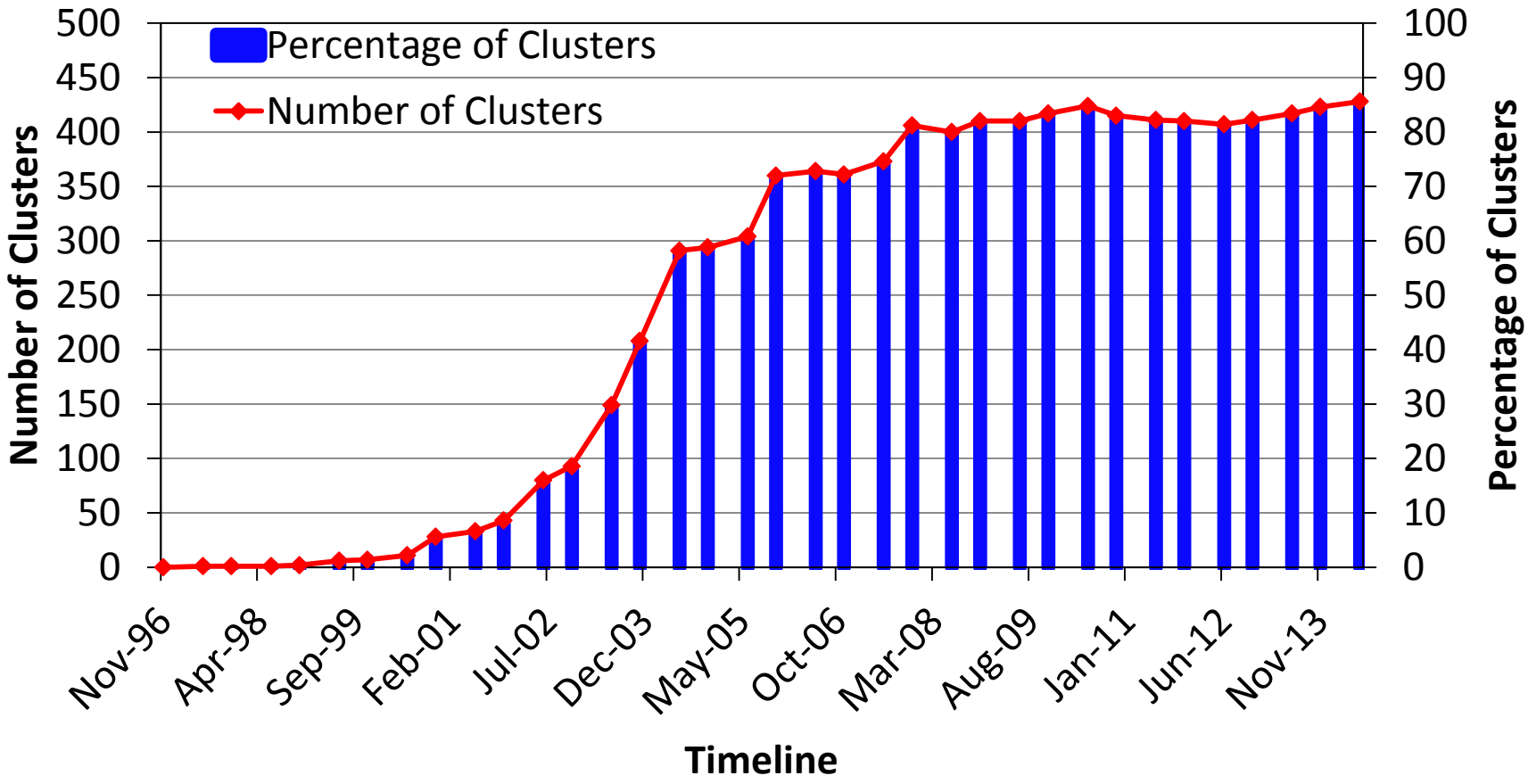
**Dhabaleswar K. (DK) Panda and Khaled Hamidouche**

The Ohio State University

E-mail: {panda, hamidouc}@cse.ohio-state.edu

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

# Trends for Commodity Computing Clusters in the Top 500 List (http://www.top500.org)



# Large-scale InfiniBand Installations

- 223 IB Clusters (44.3%) in the June 2014 Top500 list (http://www.top500.org)
- Installations in the Top 50 (25 systems):

| | |
|---|---|
| **519,640 cores (Stampede) at TACC (7th)** | 120, 640 cores (Nebulae) at China/NSCS (28th) |
| 62,640 cores (HPC2) in Italy (11th) | 72,288 cores (Yellowstone) at NCAR (29th) |
| 147, 456 cores (Super MUC) in Germany (12th) | 70,560 cores (Helios) at Japan/IFERC (30th) |
| 76,032 cores (Tsubame 2.5) at Japan/GSIC (13th) | 138,368 cores (Tera-100) at France/CEA (35th) |
| 194,616 cores (Cascade) at PNNL (15th) | 222,072 cores (QUARTETTO) in Japan (37th) |
| 110,400 cores (Pangea) at France/Total (16th) | 53,504 cores (PRIMERGY) in Australia (38th) |
| 96,192 cores (Pleiades) at NASA/Ames (21st) | 77,520 cores (Conte) at Purdue University (39th) |
| 73,584 cores (Spirit) at USA/Air Force (24th) | 44,520 cores (Spruce A) at AWE in UK (40th) |
| 77,184 cores (Curie thin nodes) at France/CEA (26h) | 48,896 cores (MareNostrum) at Spain/BSC (41st) |
| 65,320-cores, iDataPlex DX360M4 at Germany/Max-Planck (27th) | **and many more!** |

# MVAPICH2/MVAPICH2-X Software

- High Performance open-source MPI Library for InfiniBand, 10Gig/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 2012
  - Support for GPGPUs and MIC
  - **Used by more than 2,150 organizations in 72 countries**
  - **More than 218,000 downloads from OSU site directly**
  - Empowering many TOP500 clusters
    - 7th ranked 519,640-core cluster (Stampede) at TACC
    - 13th ranked 74,358-core cluster (Tsubame 2.5) at Tokyo Institute of Technology
    - 23rd ranked 96,192-core cluster (Pleiades) at NASA
  - Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- Partner in the U.S. NSF-TACC Stampede System

# MVAPICH2 2.0 GA and MVAPICH2-X 2.0 GA

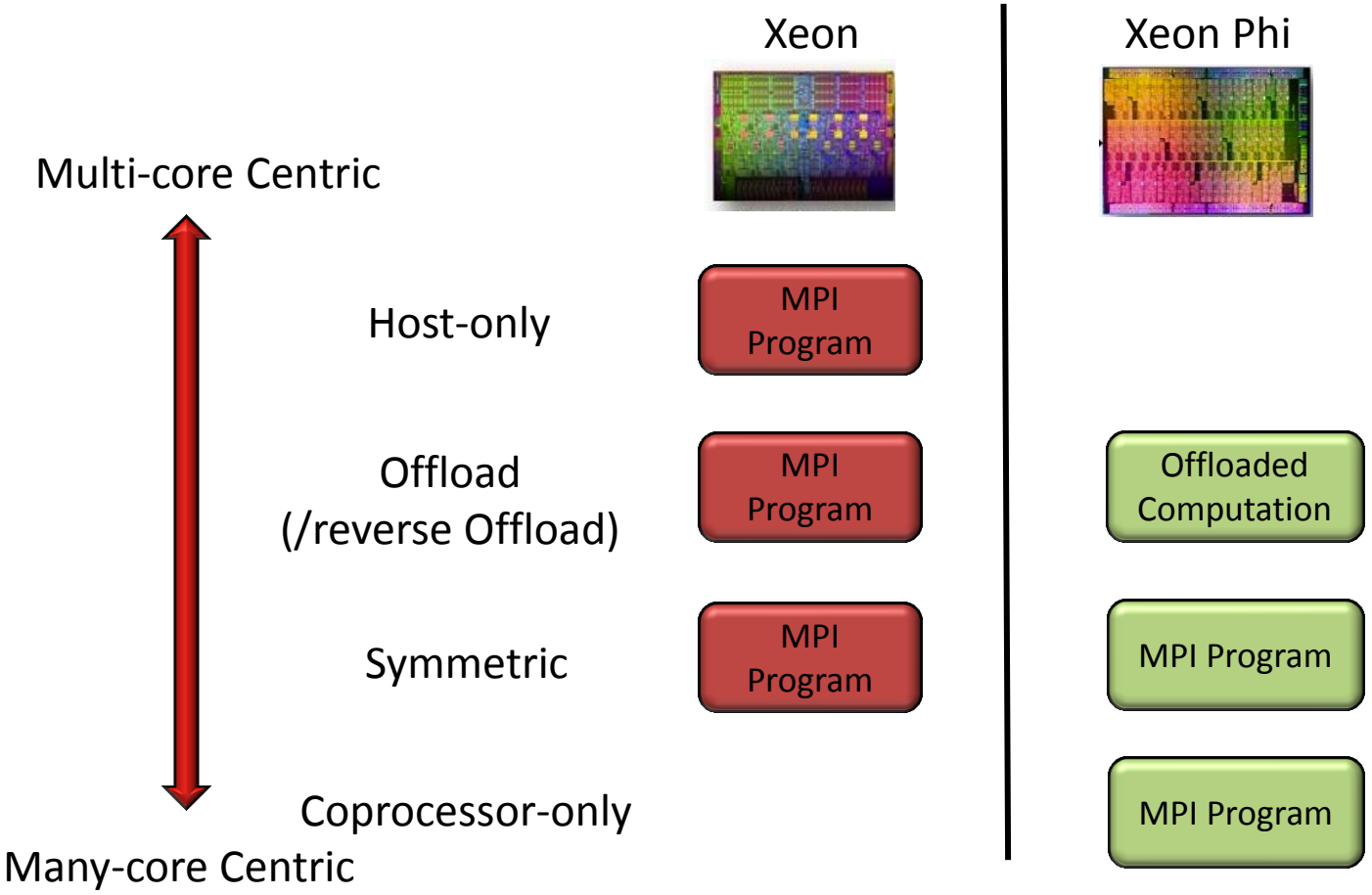
- Released on 06/20/14
- Major Features and Enhancements
  - Based on MPICH-3.1
  - MPI-3 RMA Support
  - Support for Non-blocking collectives
  - MPI-T support
  - CMA support is default for intra-node communication
  - Optimization of collectives with CMA support
  - Large message transfer support
  - Reduced memory footprint
  - Improved Job start-up time
  - Optimization of collectives and tuning for multiple platforms
  - Updated hwloc to version 1.9
- MVAPICH2-X 2.0 GA supports hybrid MPI + PGAS (UPC and OpenSHMEM) programming models
  - Based on MVAPICH2 2.0 GA including MPI-3 features
  - Compliant with UPC 2.18.0 and OpenSHMEM v1.0f

# MVAPICH2-MIC: Optimized MPI Library for Xeon Phi Clusters

- MPI libraries run out of box (or with minor changes) on the Xeon Phi
- Critical to optimize the runtimes for better performance
  - Tune existing designs
  - Designs using lower level features offered by MPSS
  - Designs to address system-level limitations
- Initial version of MVAPICH2-MIC (based on MVAPICH2 2.0a release) is available on Stampede since Oct '13
  - Supports all modes of usage – host-only, offload, coprocessor-only and symmetric
  - Improved shared memory communication channel
  - SCIF-based designs for improved communication within MIC and between MICs and Hosts
  - Proxy-based design to work around bandwidth limitations on Sandy Bridge platform
- Enhanced version based on MVAPICH2 2.0GA release is being worked out. Will be coming out soon!!

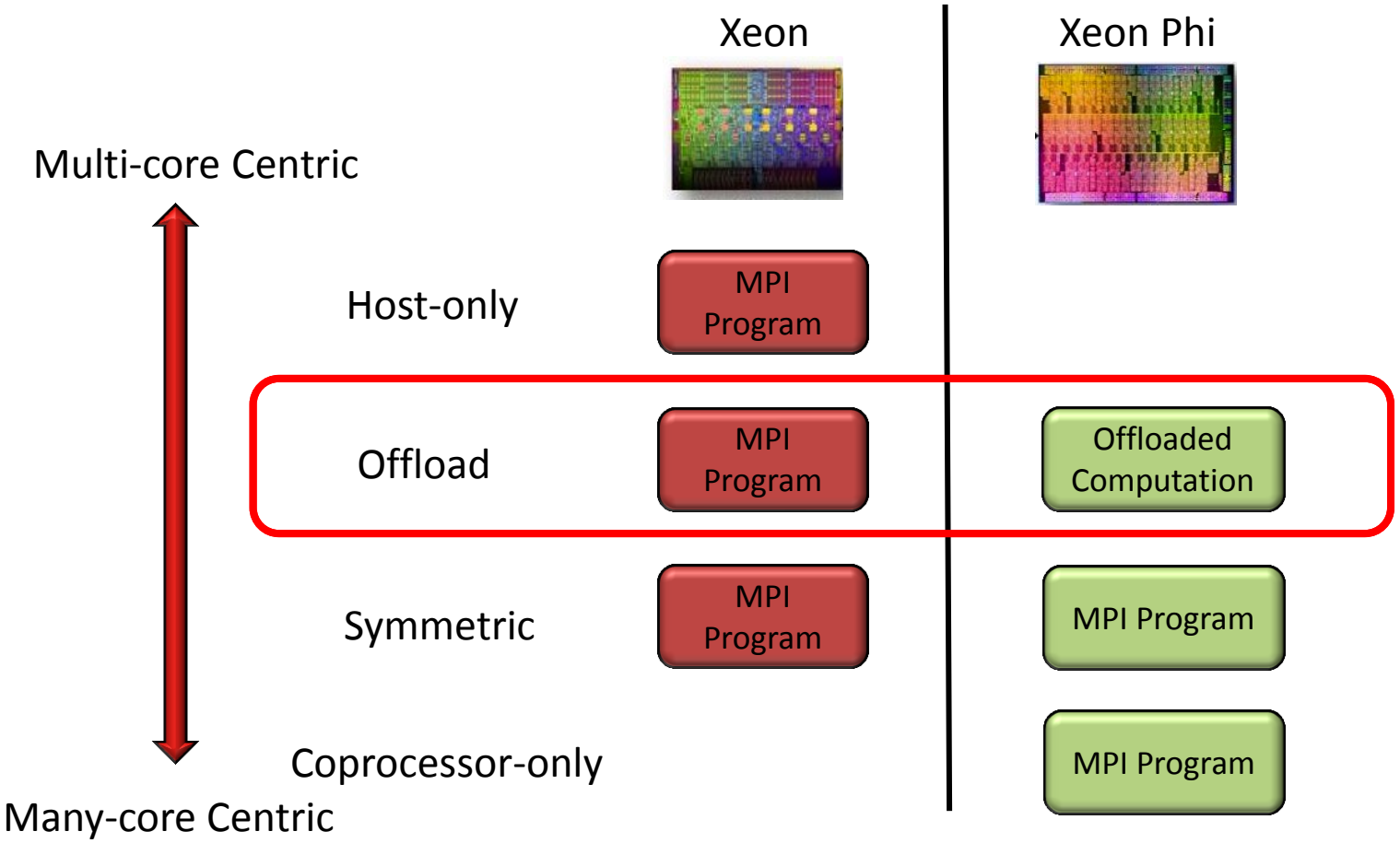
# MPI Applications on MIC Clusters

- MPI (+X) continues to be the predominant programming model in HPC
- Flexibility in launching MPI jobs on clusters with Xeon Phi



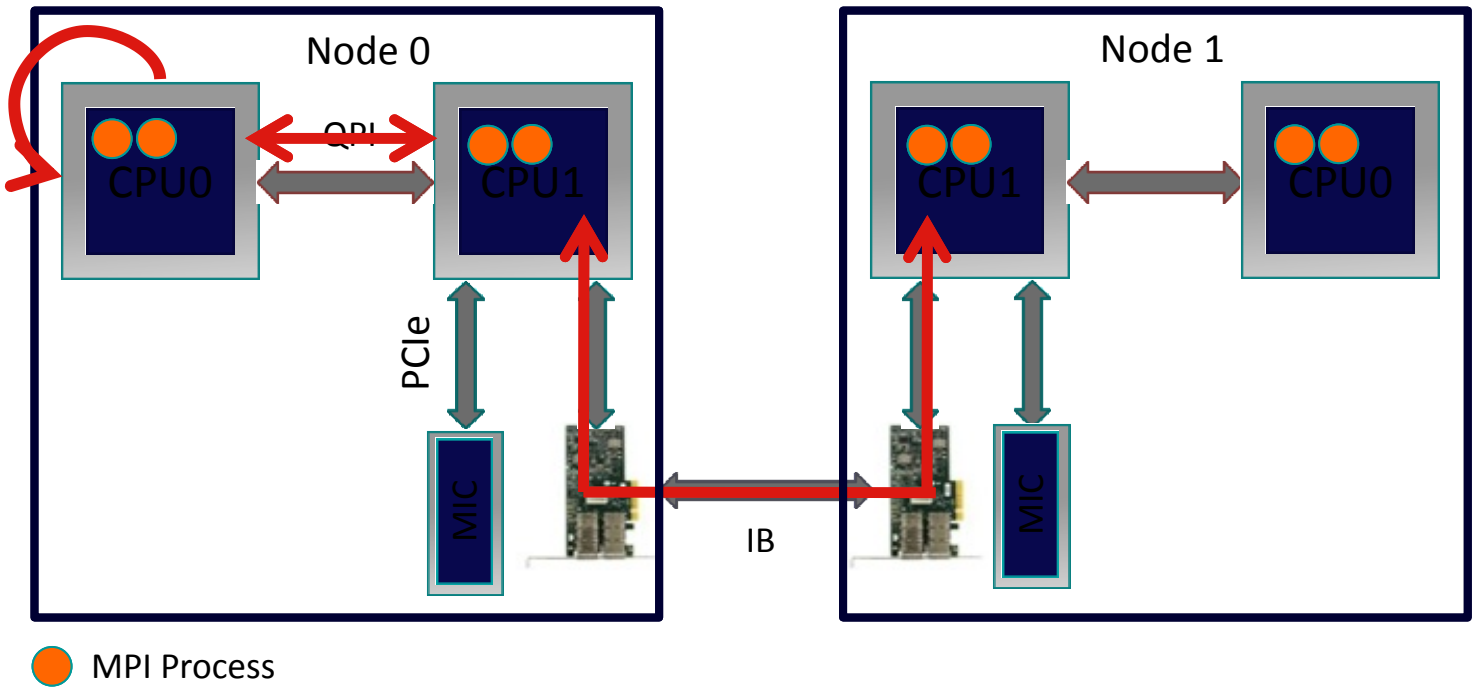
# MPI Applications on MIC Clusters

- Offload mode: lesser overhead way to extract performance from Xeon Phi



# MPI Data Movement in Offload Mode

- MPI communication only from the host

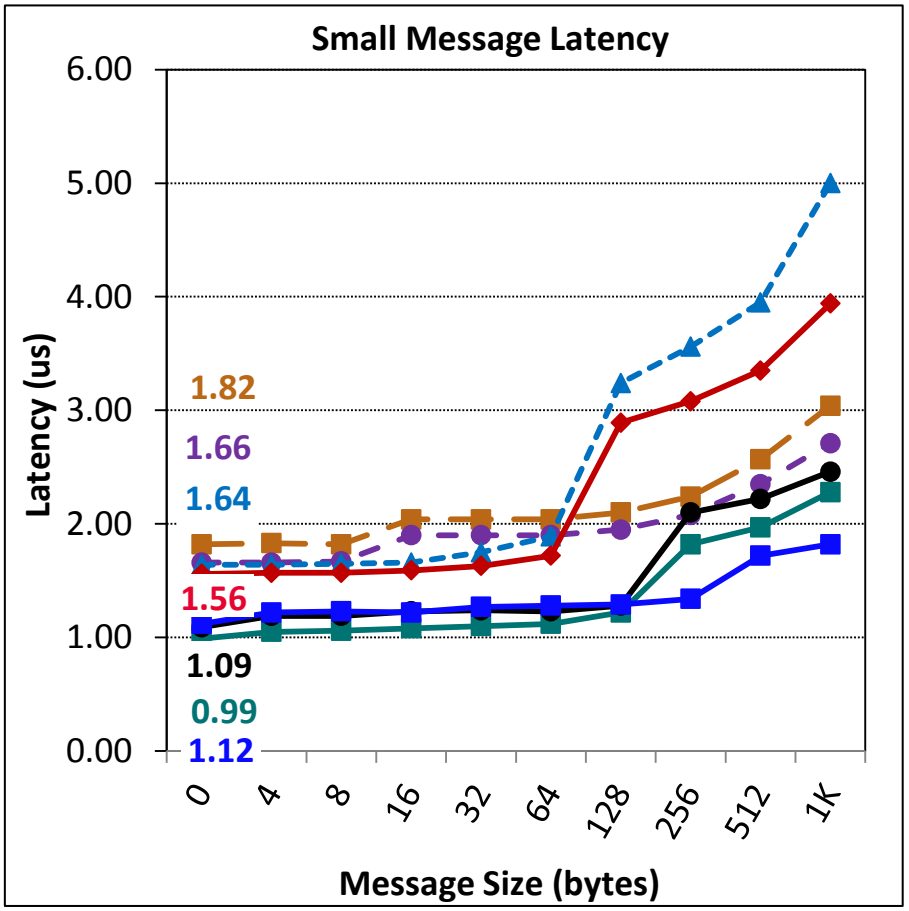


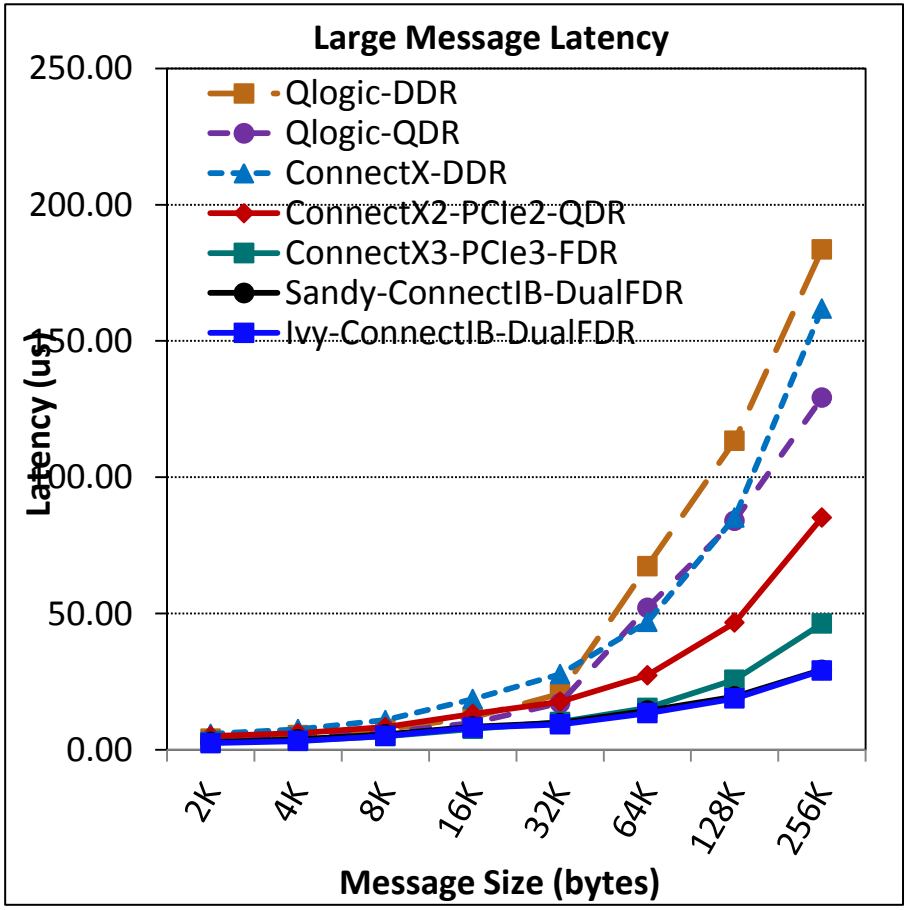

1. Intra-Socket
2. Inter-Socket
3. Inter-Node

***Supported by the Standard MVAPICH2 Library***

***(Latest Release - MVAPICH2 2.0 GA)***

# One-way Latency: MPI over IB with MVAPICH



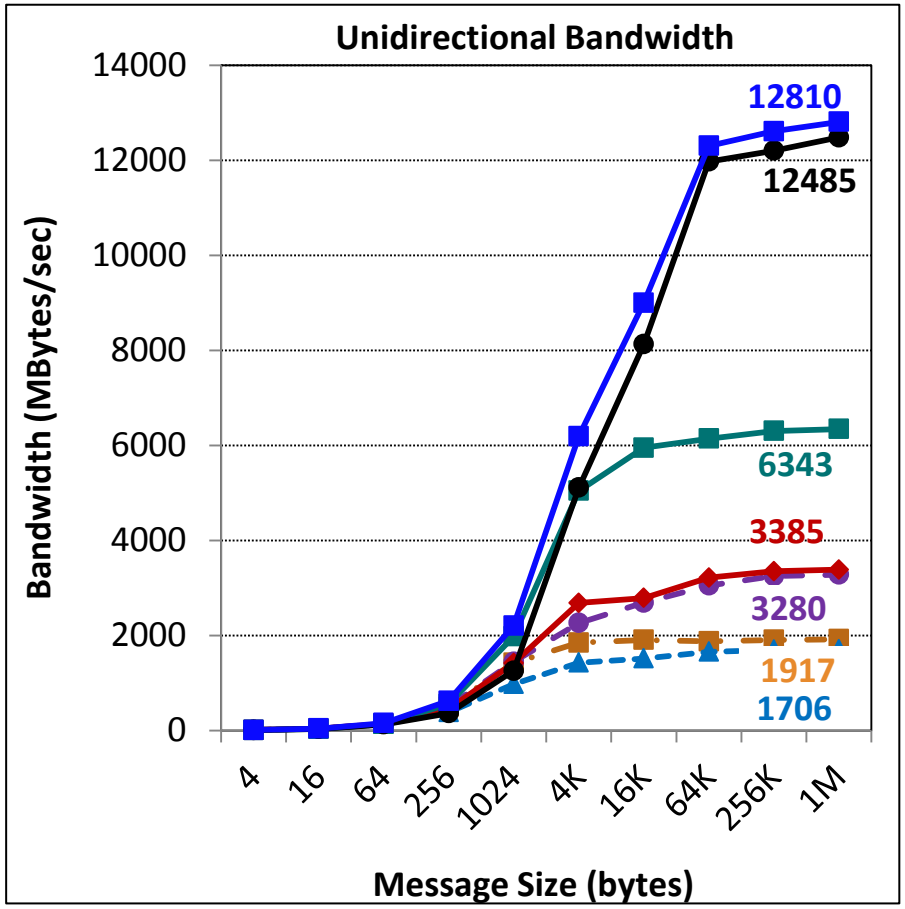

DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

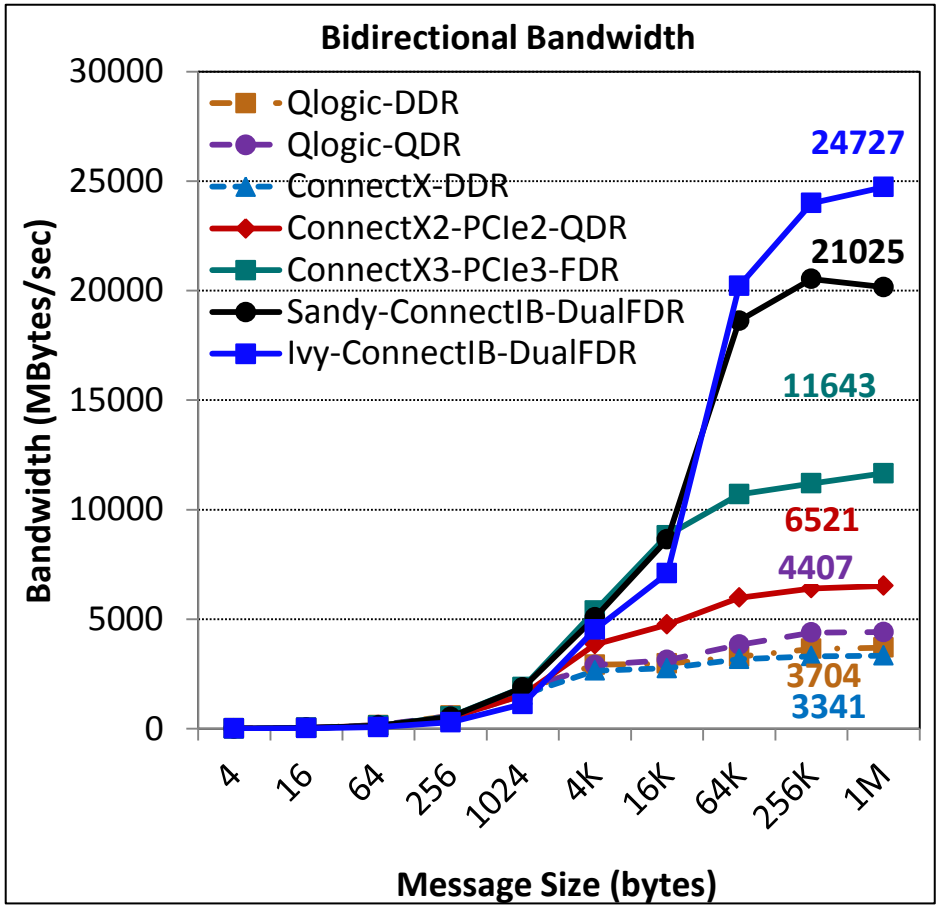
FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

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

# Bandwidth: MPI over IB with MVAPICH



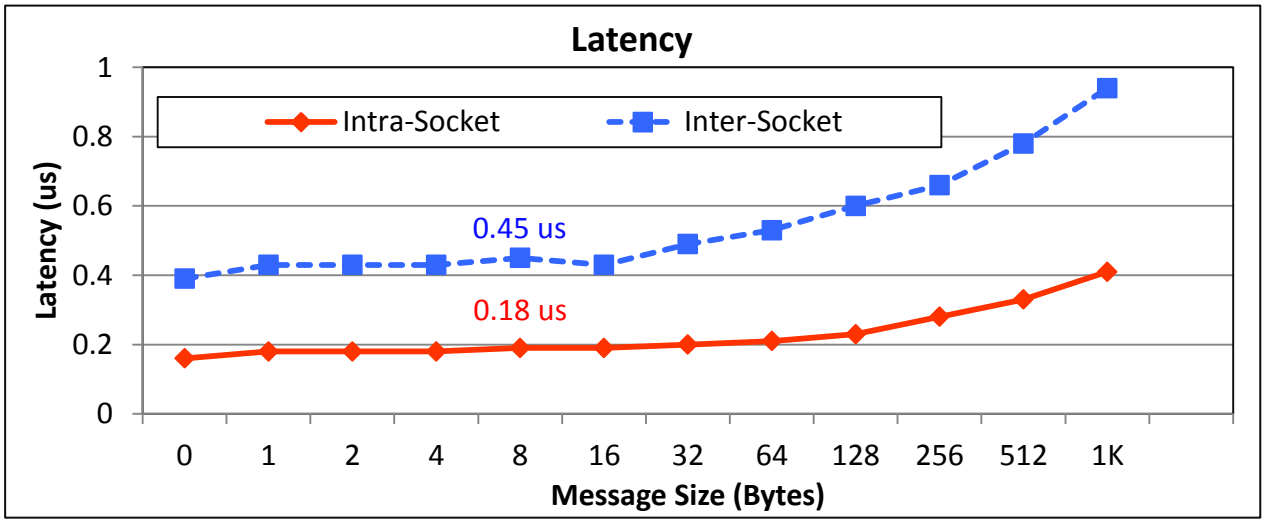

DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

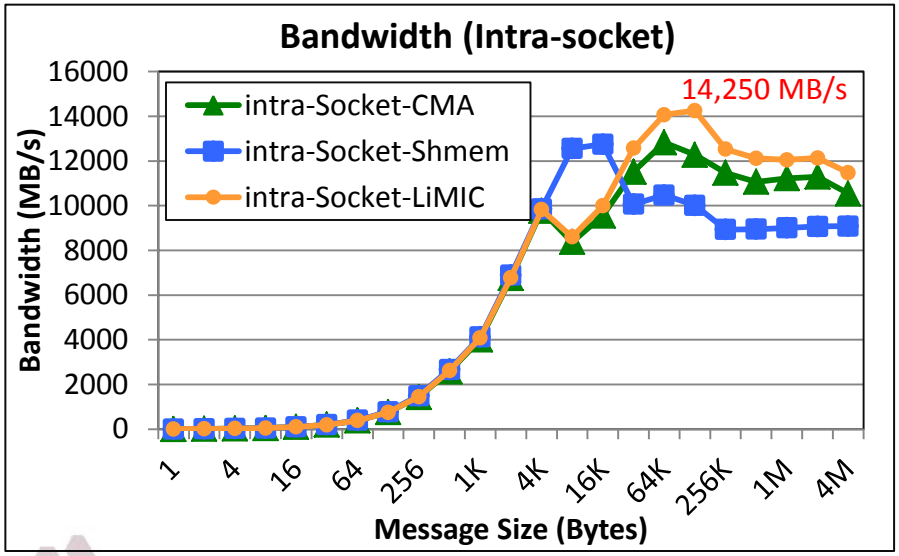
ConnectIB-Dual FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

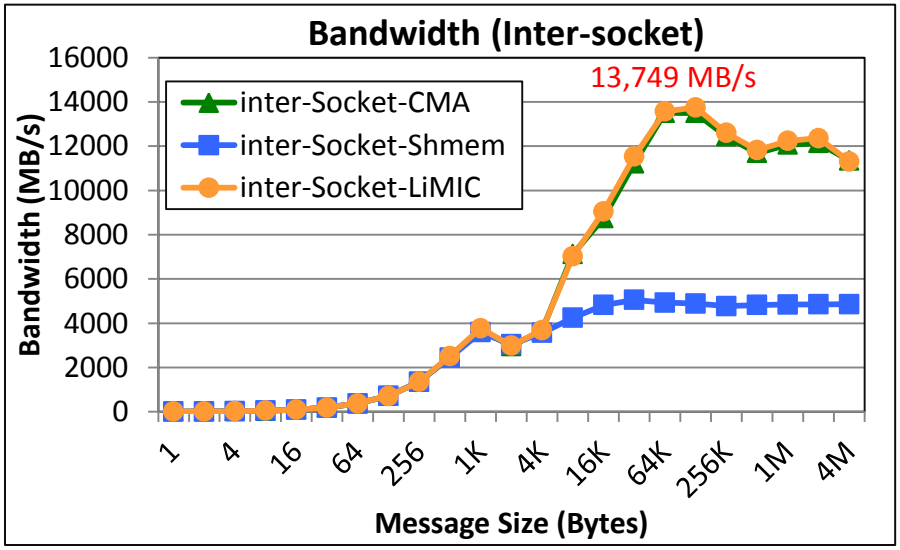
# MVAPICH2 Two-Sided Intra-Node Performance
## (Shared memory and Kernel-based Zero-copy Support (LiMIC and CMA))



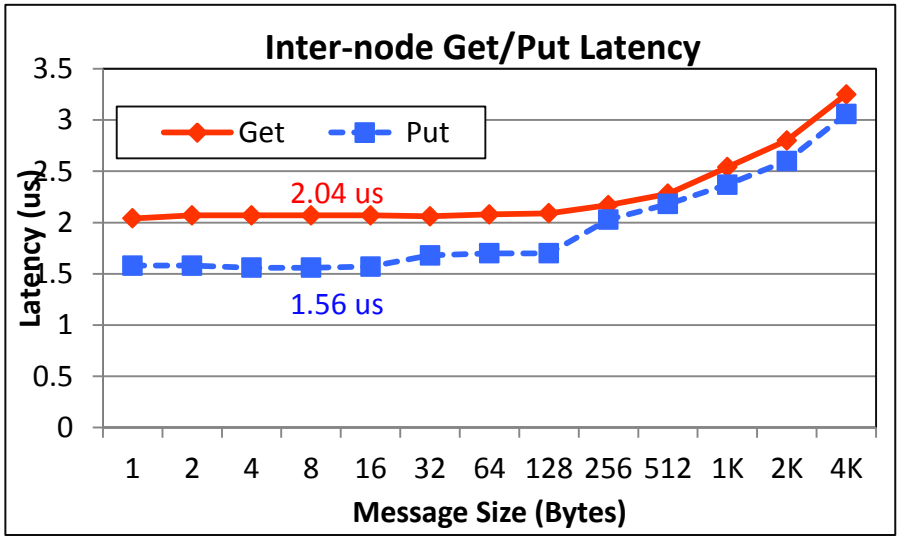

Latest MVAPICH2 2.0

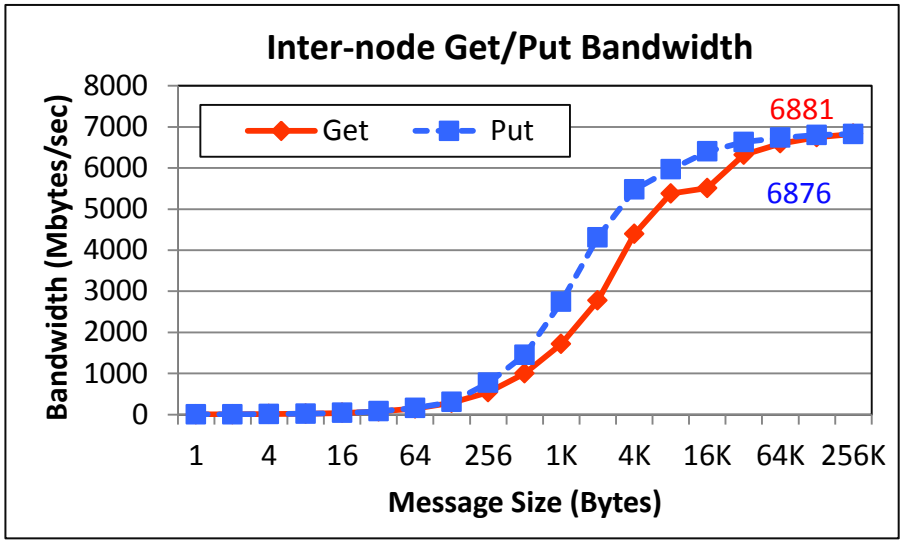
Intel Ivy-bridge

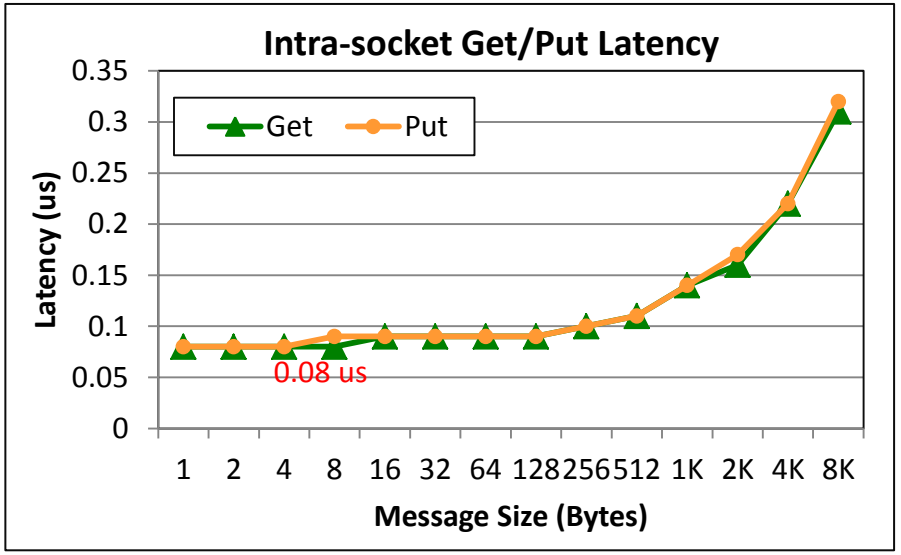


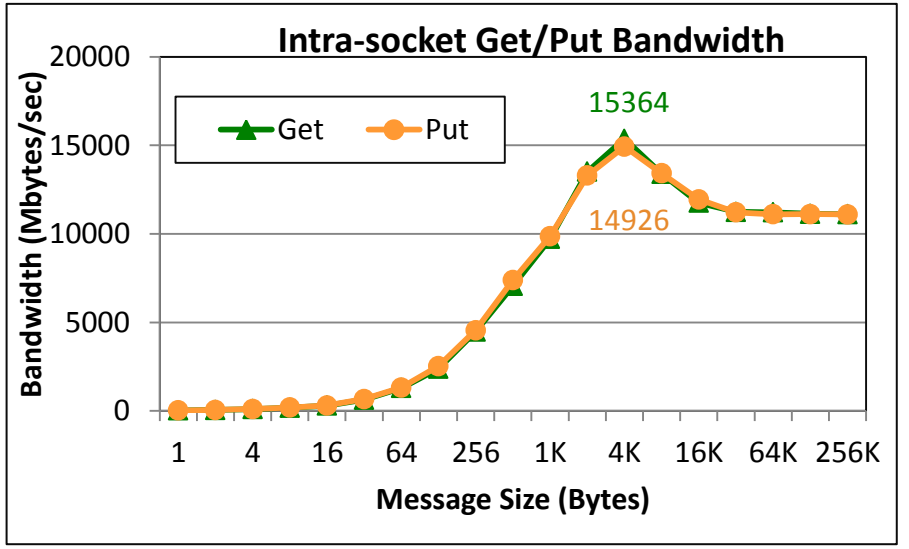



# MPI-3 RMA Get/Put with Flush Performance



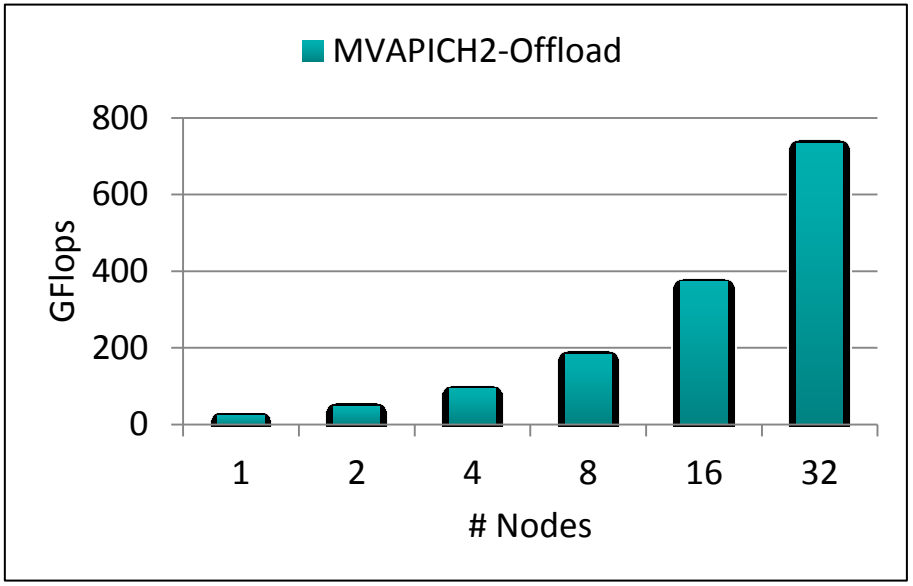



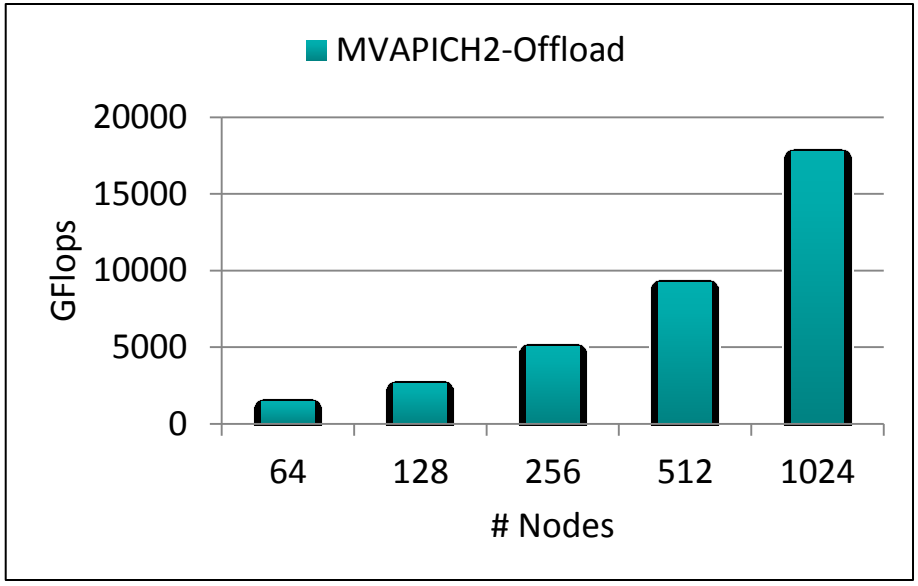




Latest MVAPICH2 2.0, Intel Sandy-bridge with Connect-IB (single-port)

# HPCG Benchmark with Offload Mode

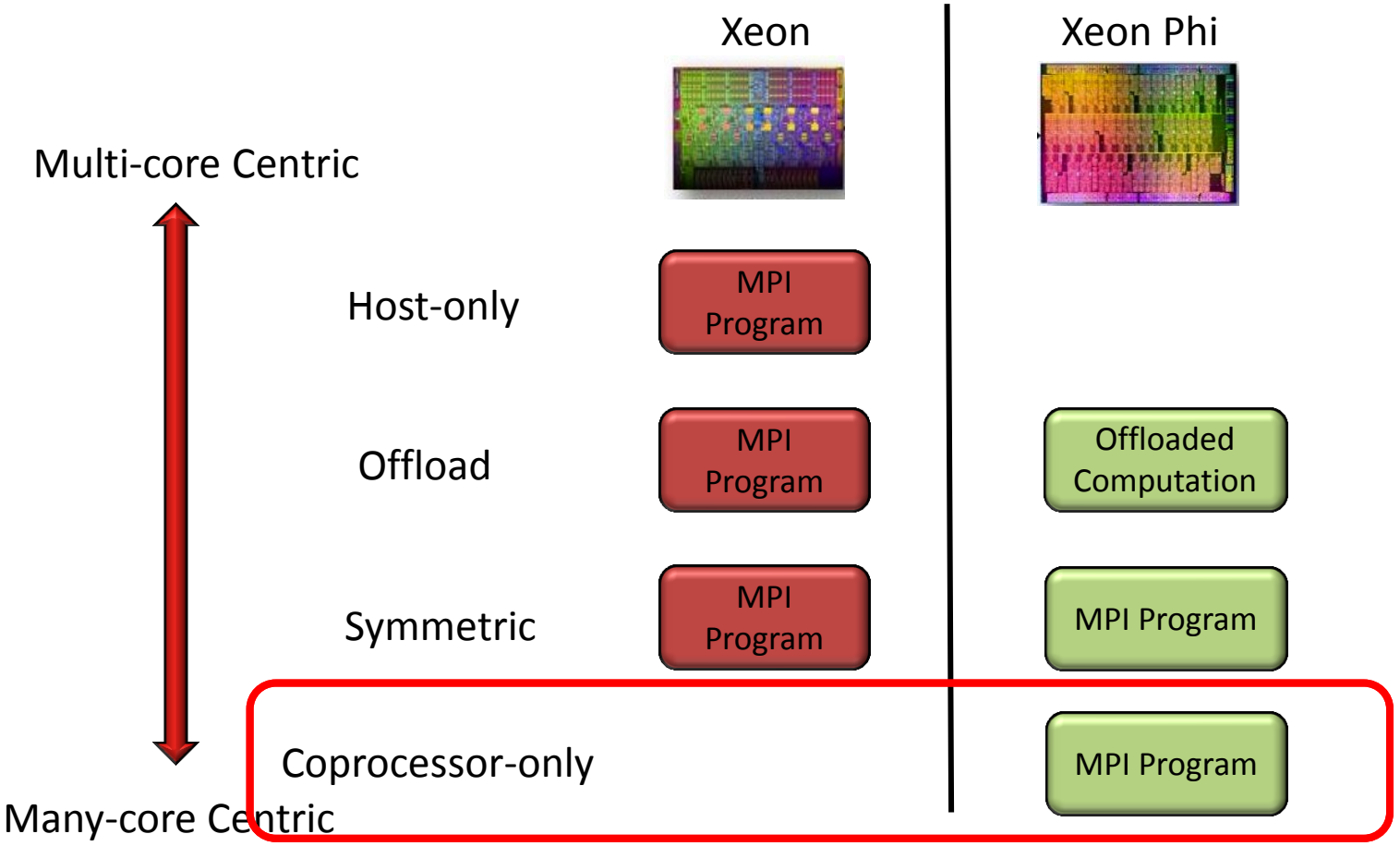


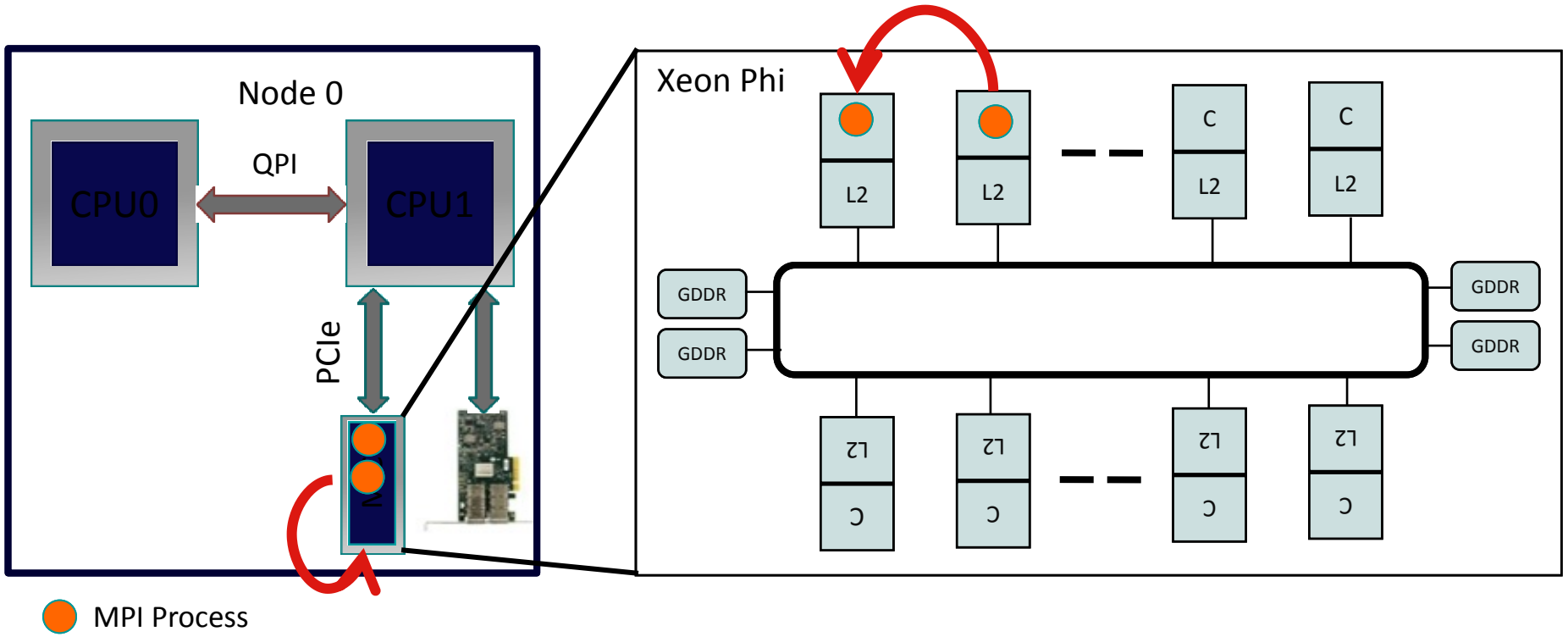

- Full subscription of a node:
  - 1 MPI process + 10 OpenMP threads on the CPU
  - 3 MPI processes + 6 OpenMP threads on CPU + 240 OpenMP threads offload on MIC
  - Input data size = 128^3
- Binding using MV2_CPU_MAPPING
- 1 node MVAPICH2 achieves 24 GFlops
- 1024 nodes MVAPICH2 achieves 18.2 TFlops

# MPI Applications on Xeon Phi Clusters - IntraNode

- Xeon Phi as a many-core node



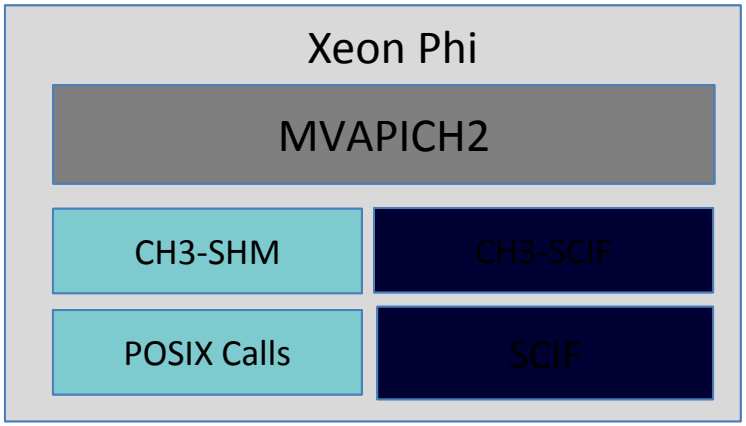
# MPI Data Movement in Coprocessor-Only Mode




**Intra-MIC Communication**

# MVAPICH2-MIC Channels for Intra-MIC Communication

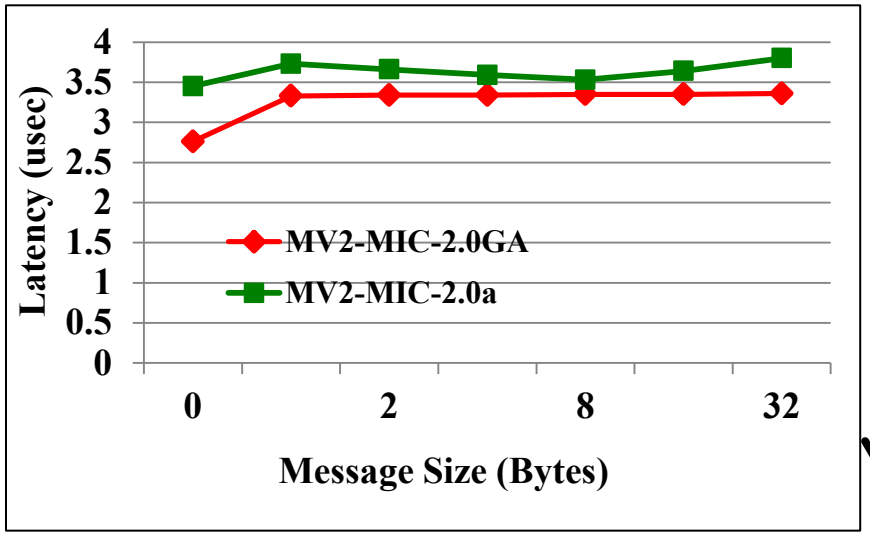
- MVAPICH2 provides a hybrid of two channels
  - CH3-SHM
    - Derives the design for shared memory communication on host
    - Tuned for the Xeon Phi architecture
  - CH3-SCIF
    - SCIF is a lower level API provided by MPSS
    - Provides user control of the DMA
    - Takes advantage of SCIF to improve performance



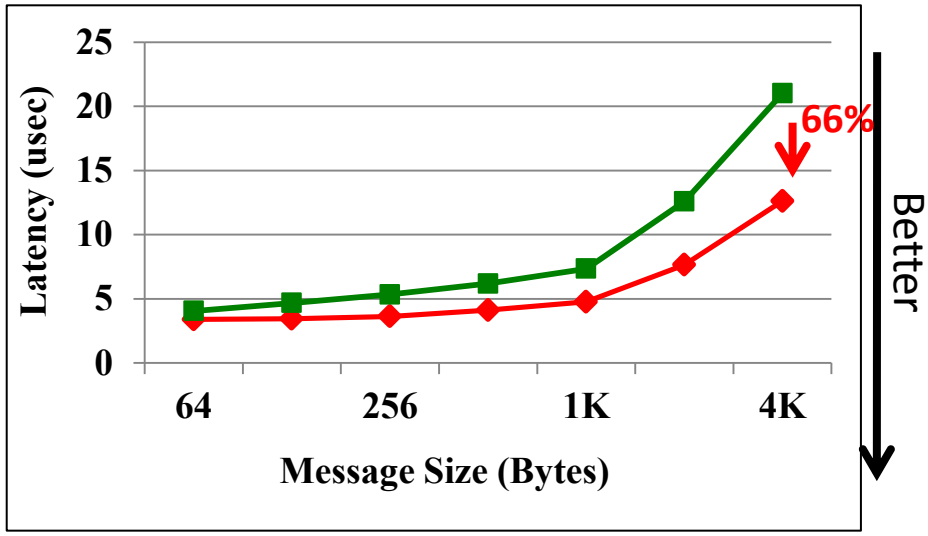

S. Potluri, A. Venkatesh, D. Bureddy, K. Kandalla, and D. K. Panda - Efficient Intra-node Communication on Intel-MIC Clusters - International Symposium on Cluster, Cloud and Grid Computing, May 2013 (CCGrid' 13).

S. Potluri, K. Tomko, D. Bureddy and D. K. Panda - Intra-MIC MPI Communication using MVAPICH2: Early Experience - TACC-Intel Highly-Parallel Computing Symposium (TI-HPCS), April 2012 - Best Student Paper

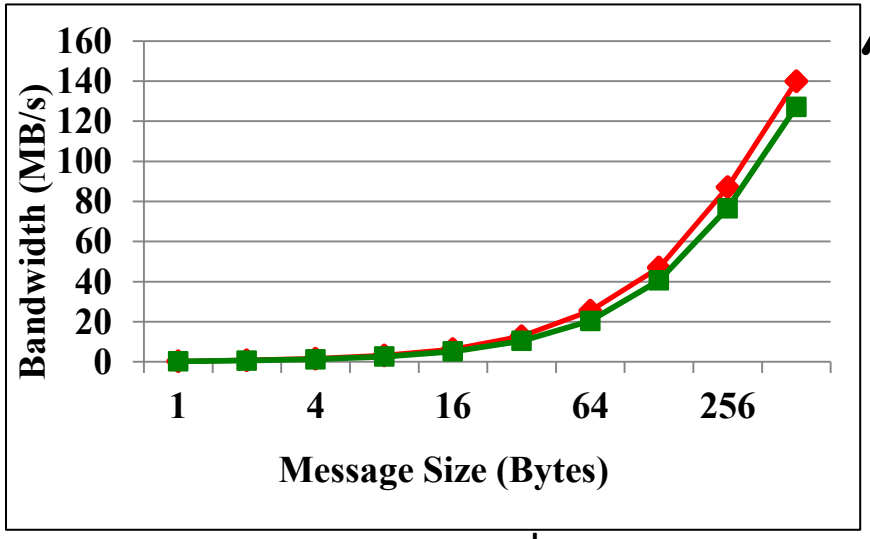
# Intra-MIC - Point-to-Point Communication



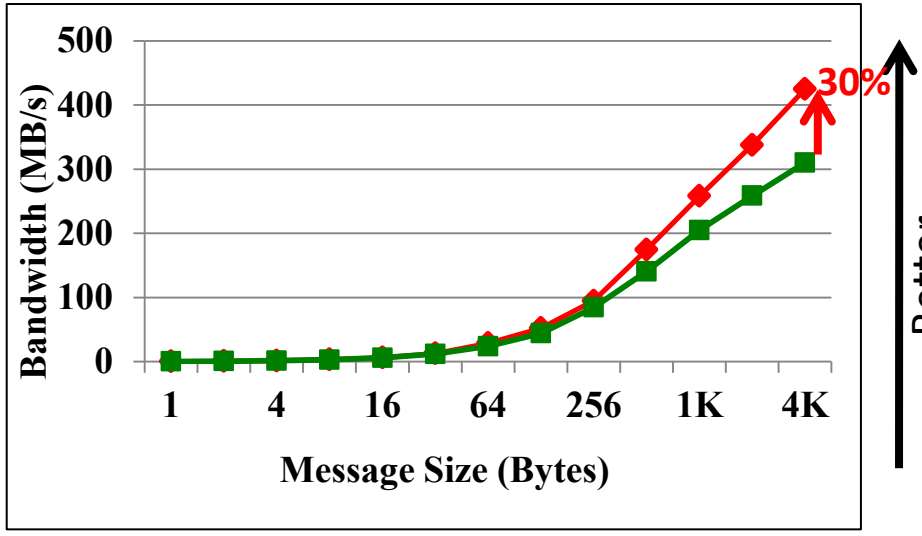

osu_latency (small)




osu_latency (medium)

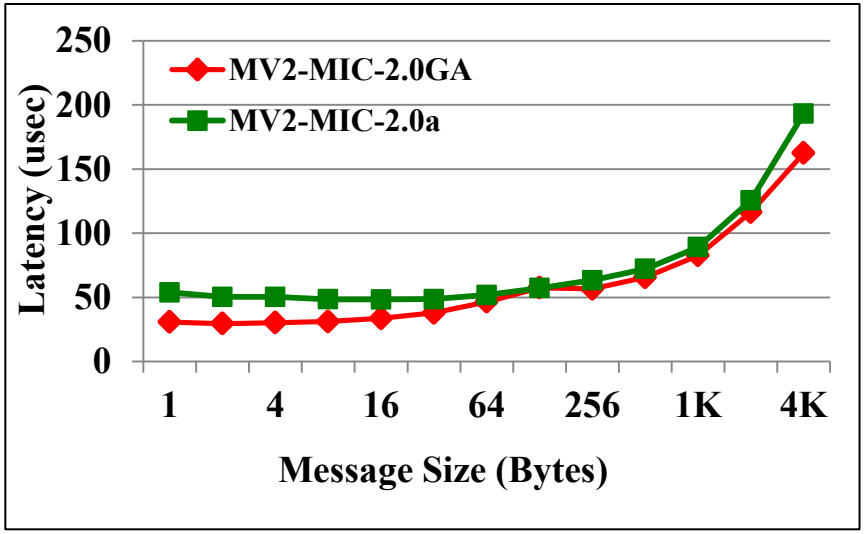



osu_bw

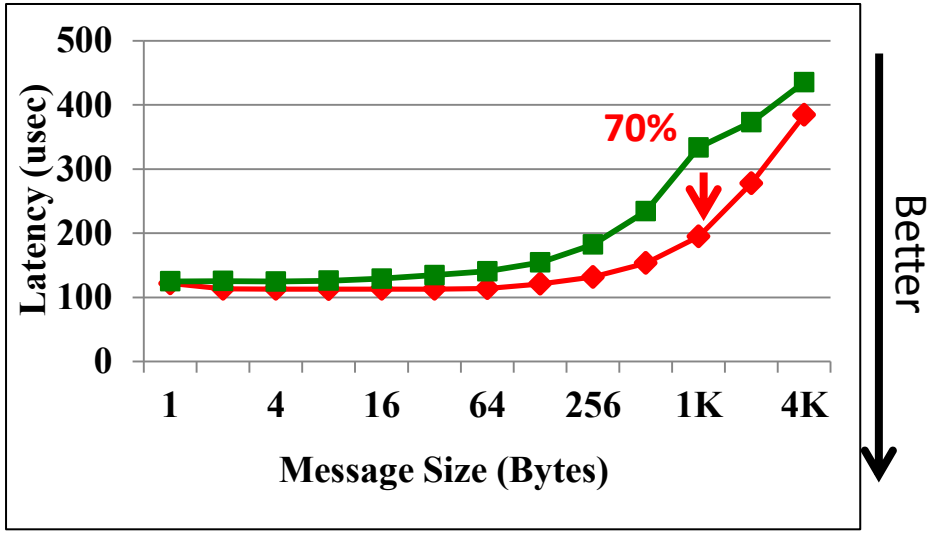



osu_bibw

# Intra-MIC - Collective Communication – AlltoAll



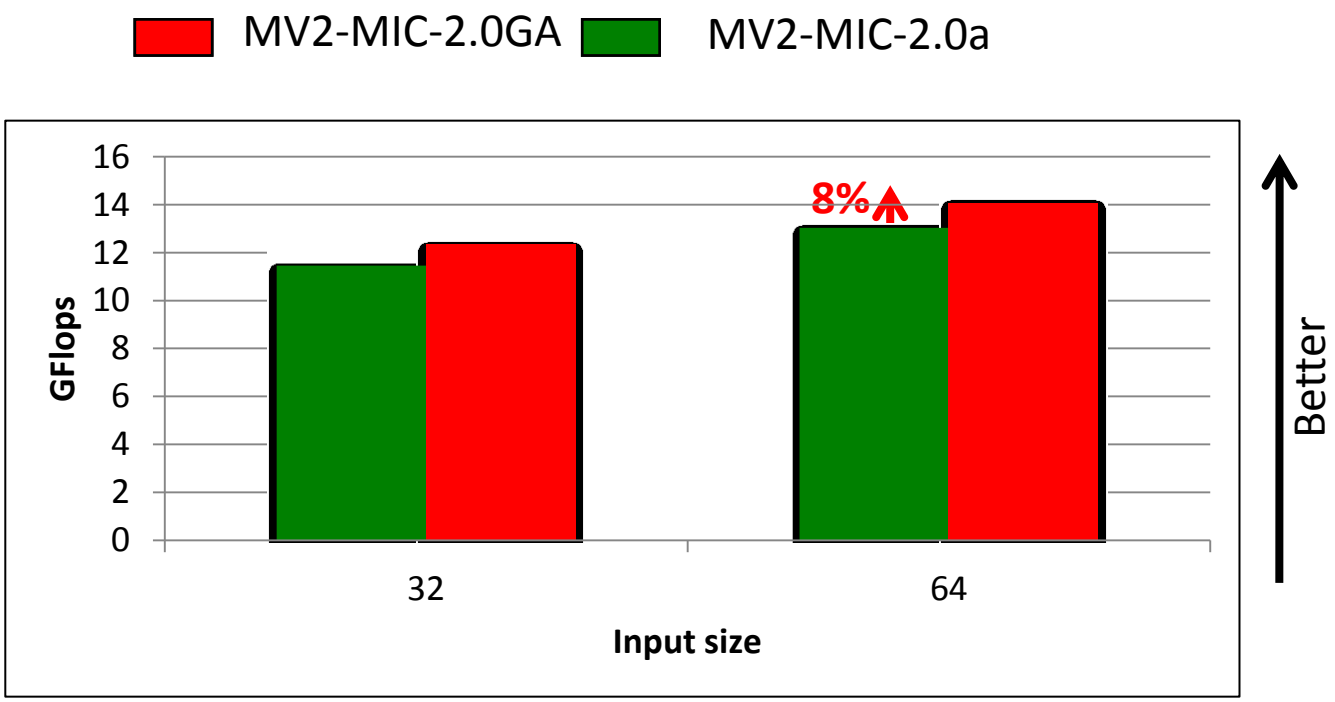

8 processes – osu_alltoall (small)




16 processes – osu_alltoall (small)

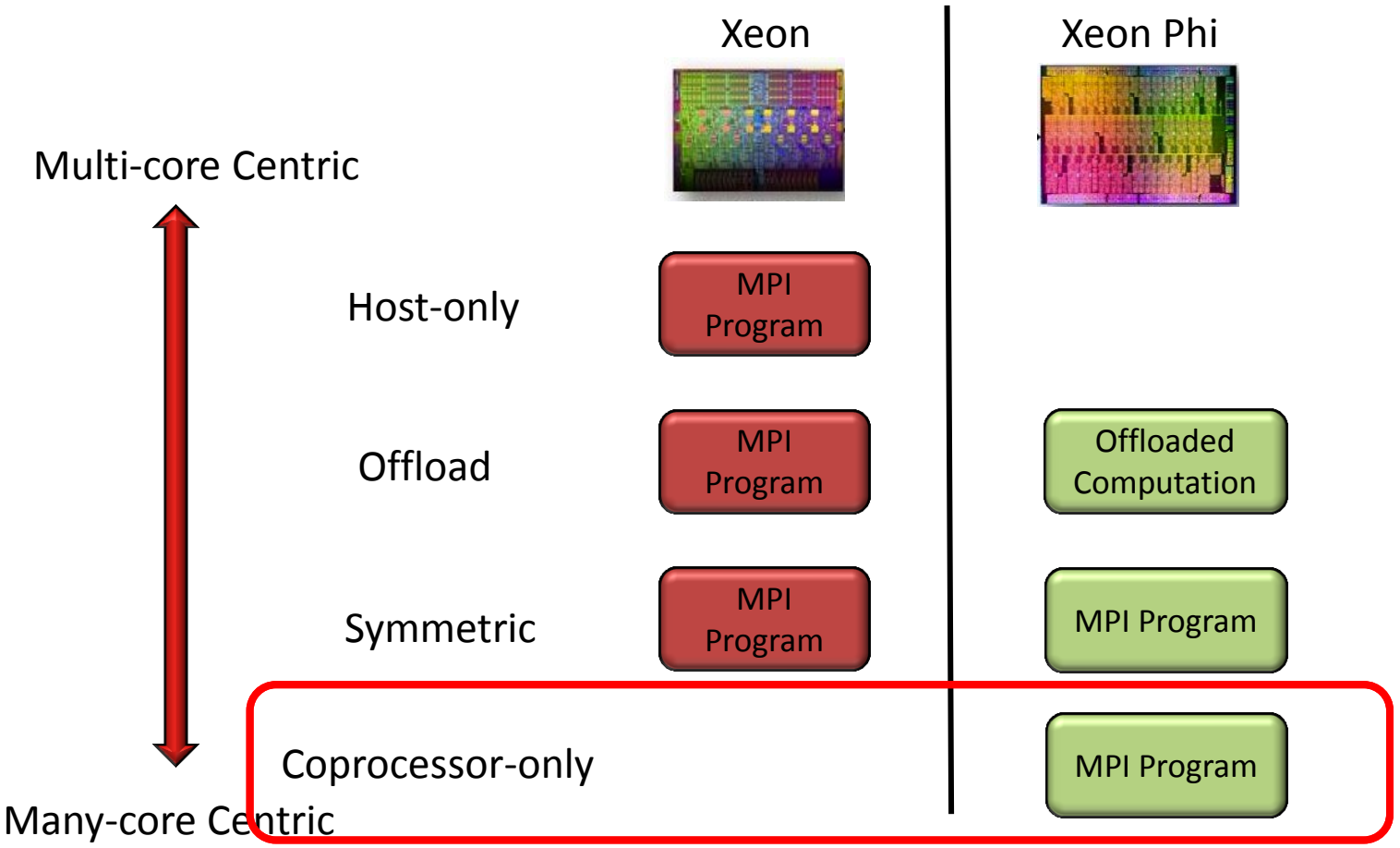
- 8 Processes: 40% and 19% improvement for 4 and 4K bytes messages
- 16 Processes: 38% and 71% improvement for 64 and 1K bytes messages

# Intra-MIC – HPCG Benchmark

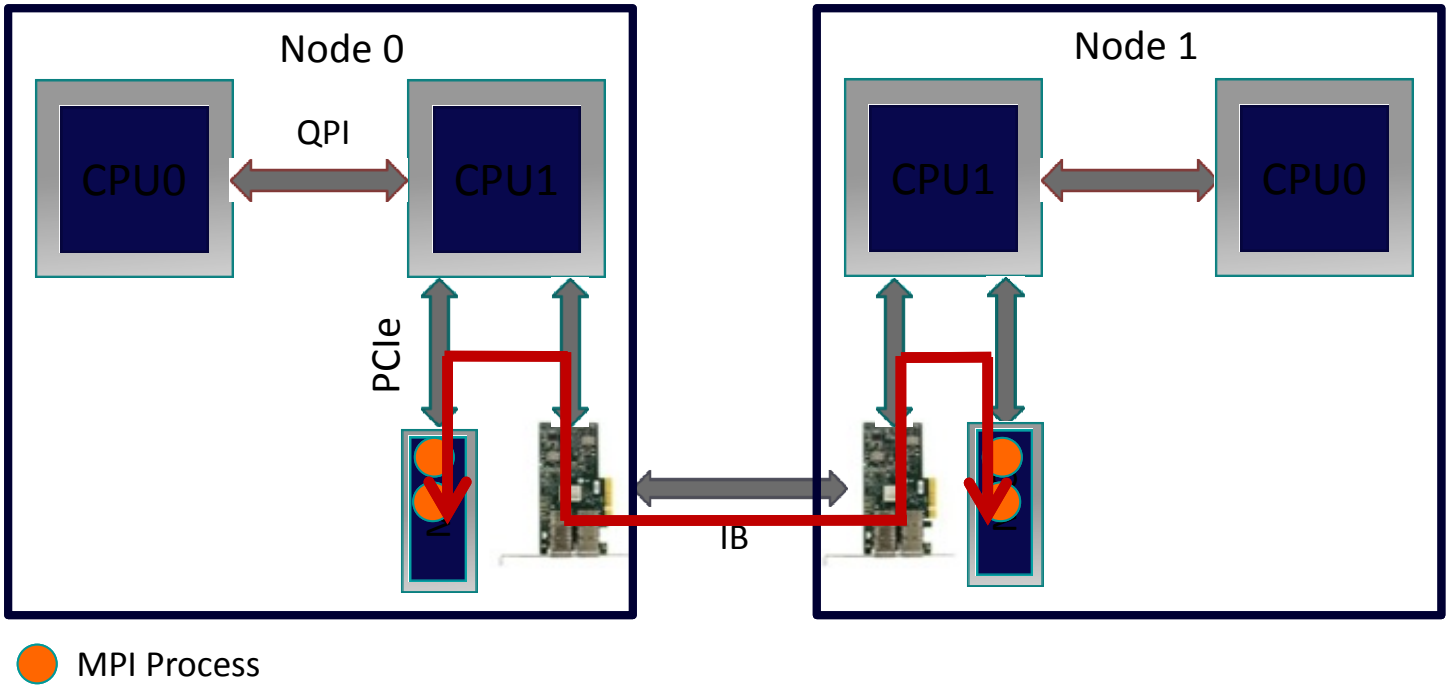



- 4 MPI processes with 60 OpenMP threads per process
- Binding using MV2_MIC_MAPPING environment variable

# MPI Applications on MIC Clusters - InterNode

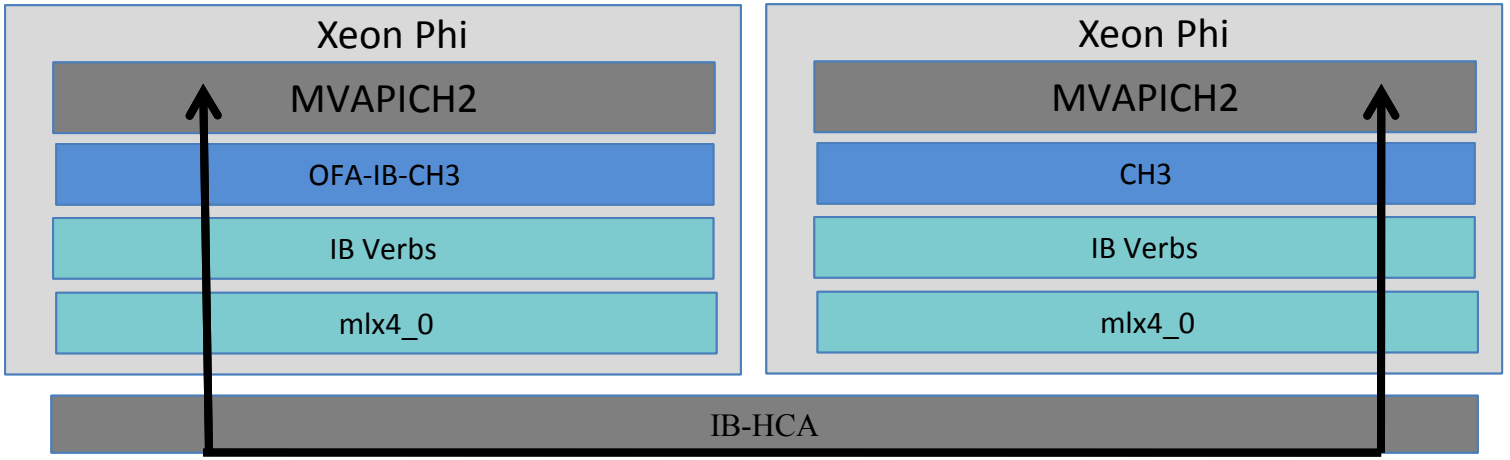


# MPI Data Movement in Coprocessor-Only Mode



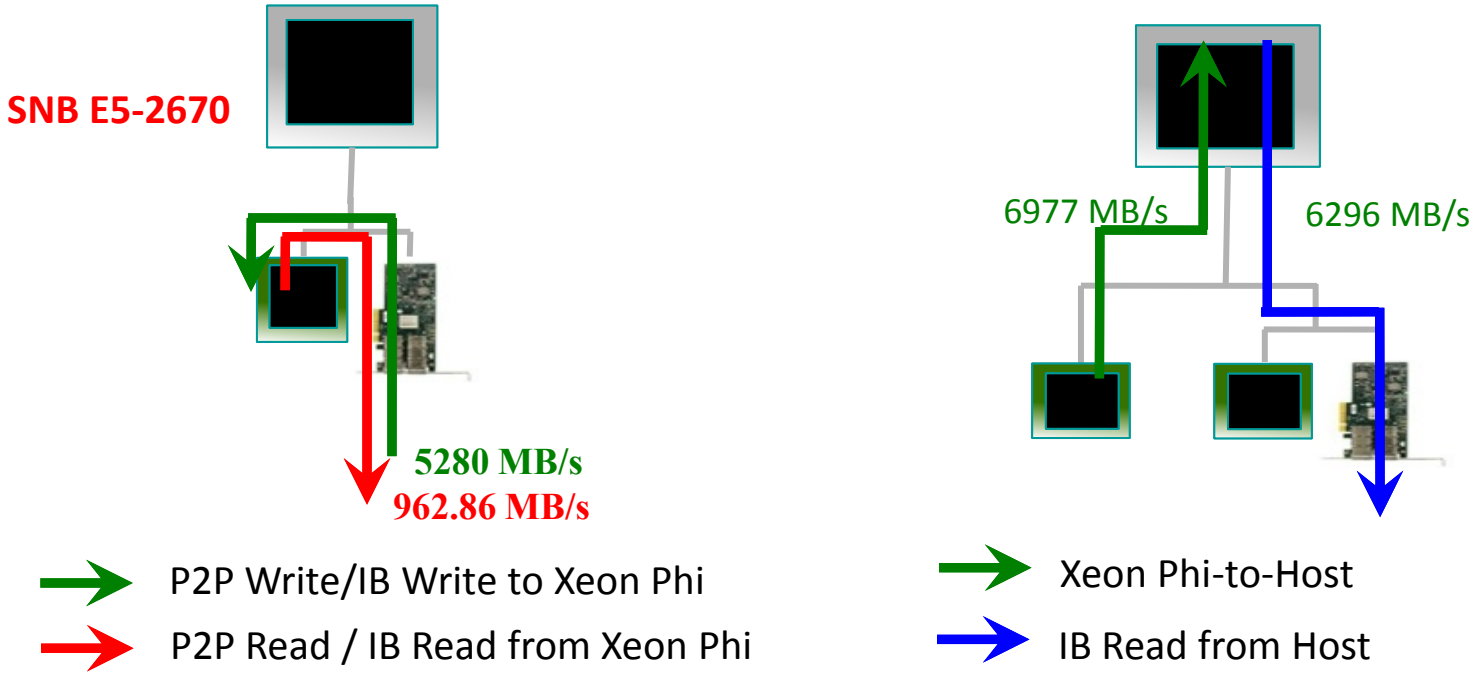

**1. MIC-RemoteMIC**

# MVAPICH2 Channels for MIC-RemoteMIC Communication




- CH3-IB channel
  - High-performance InfiniBand channel in MVAPICH2
  - Implemented using IB Verbs - DirectIB
  - Currently does not support advanced features like hardware multicast, etc.
  - Limited by P2P Bandwidth offered on Sandy Bridge platform

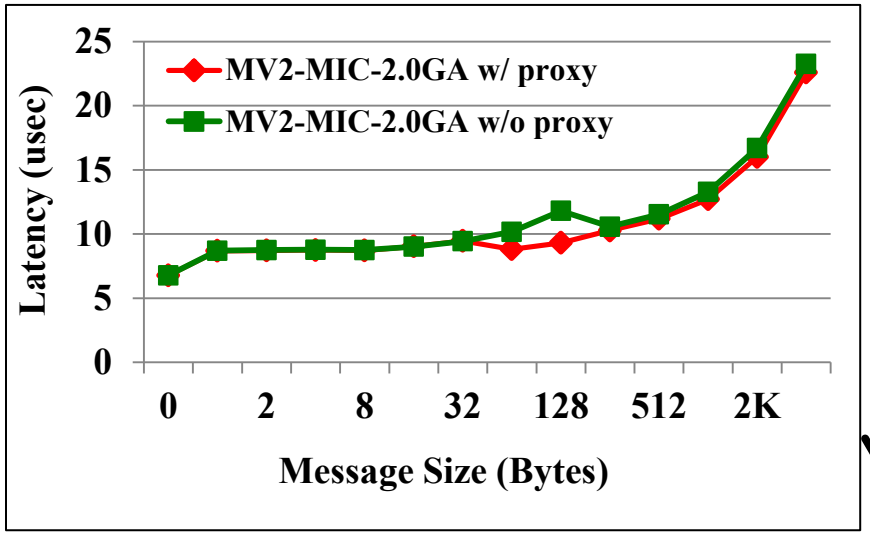
# Host Proxy-based Designs in MVAPICH2-MIC



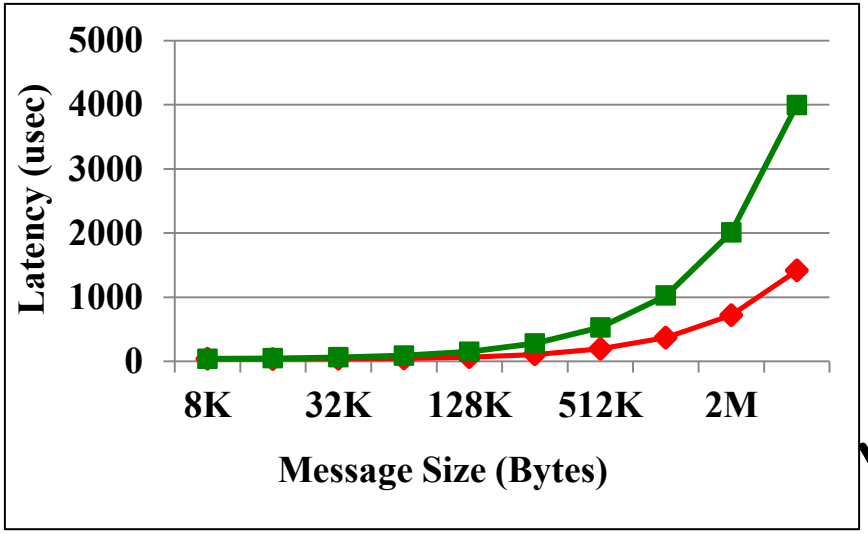

- Direct IB channels is limited by P2P read bandwidth
- MVAPICH2-MIC uses a hybrid DirectIB + host proxy-based approach to work around this

S. Potluri, D. Bureddy, K. Hamidouche, A. Venkatesh, K. Kandalla, H. Subramoni and D. K. Panda, MVAPICH-PRISM: A Proxy-based Communication Framework using InfiniBand and SCIF for Intel MIC Clusters Int'l Conference on Supercomputing (SC '13), November 2013.

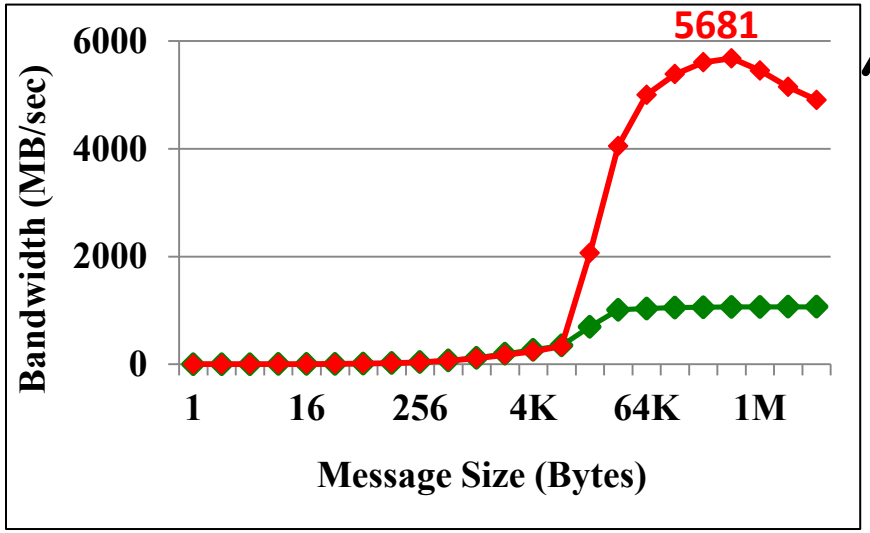
# MIC-RemoteMIC Point-to-Point Communication (Active Proxy)



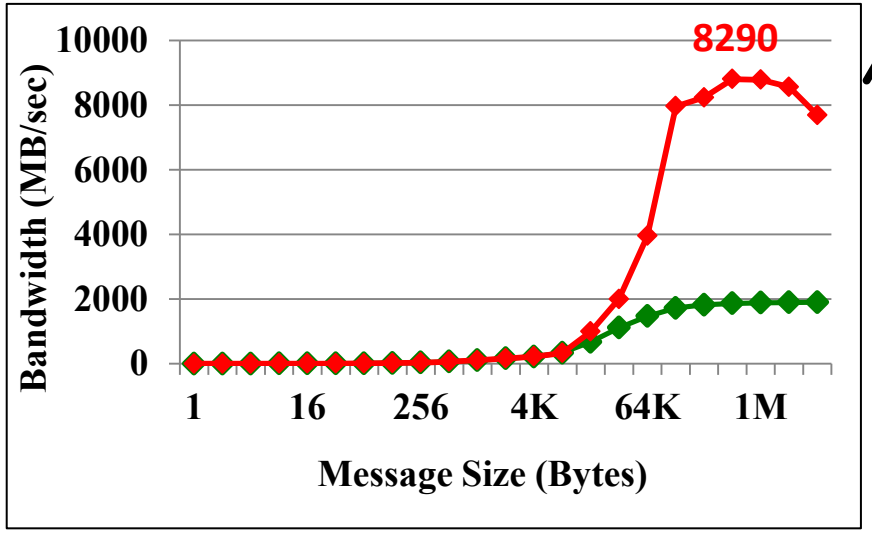

osu_latency (small)




osu_latency (large)

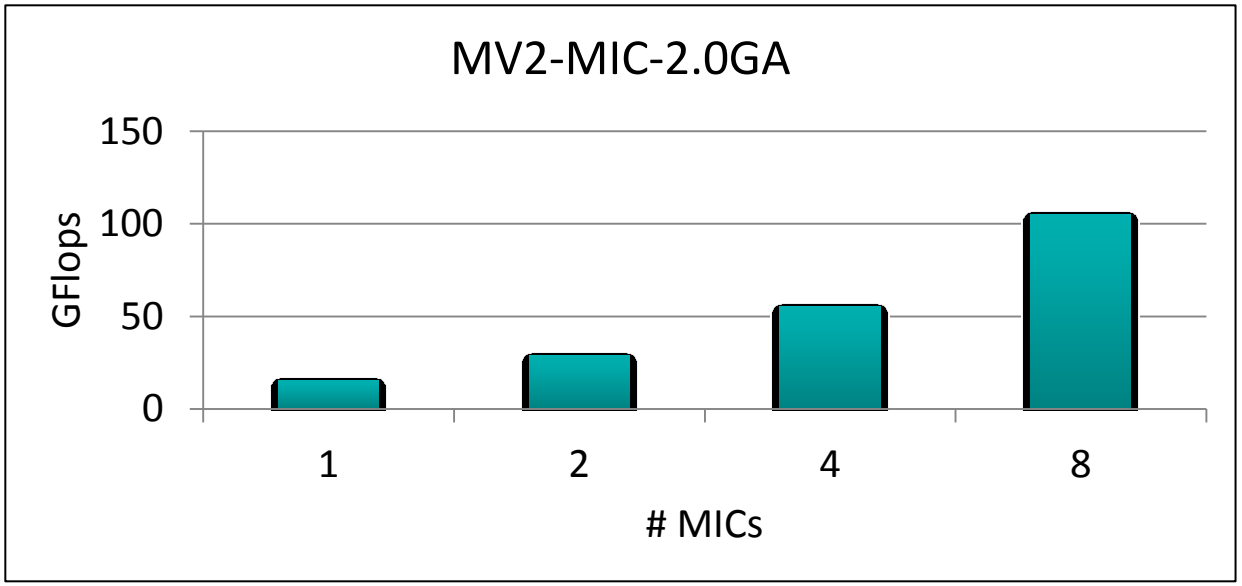



osu_bw



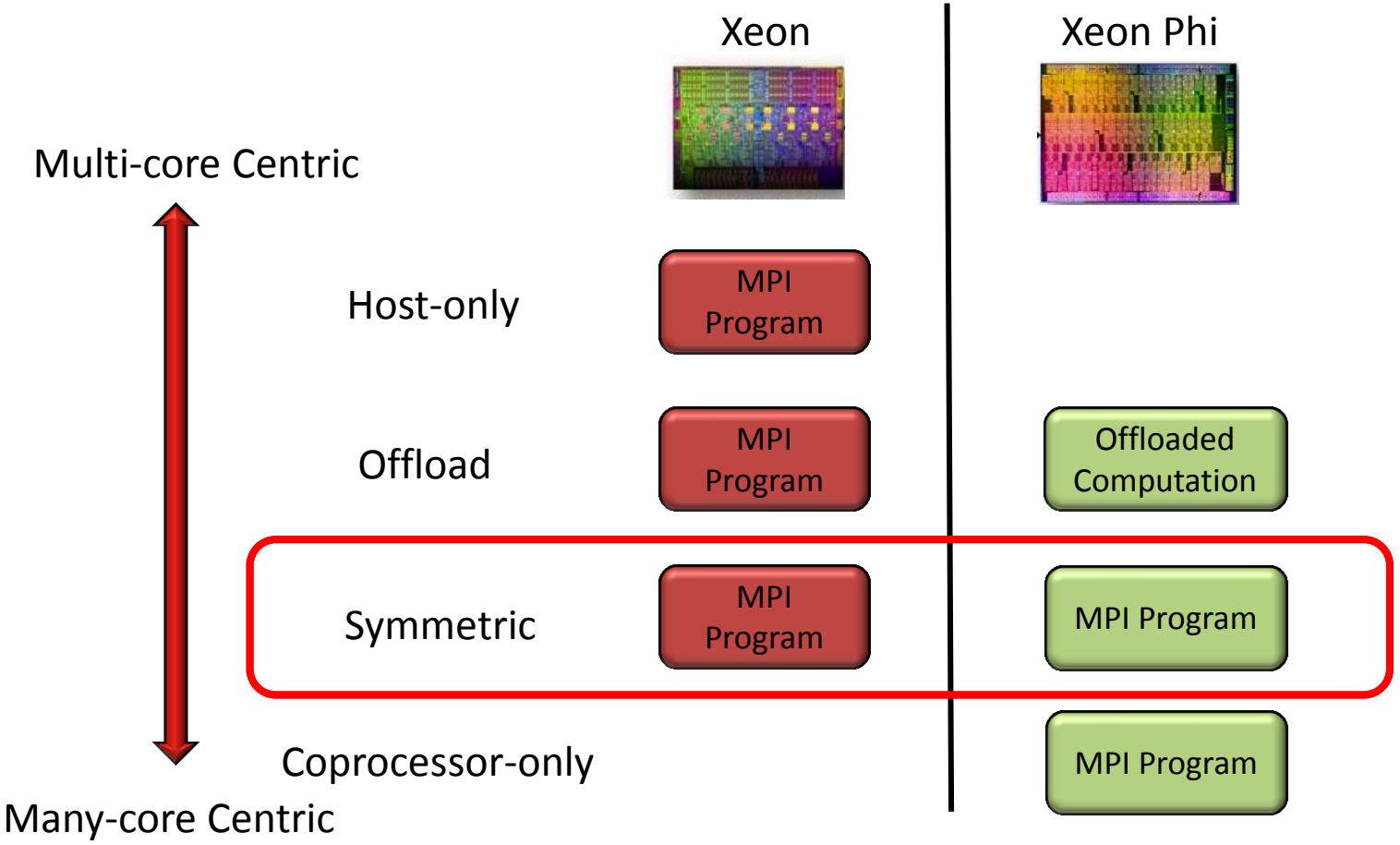

osu_bibw

# InterNode - Coprocessor-only Mode - HPCG

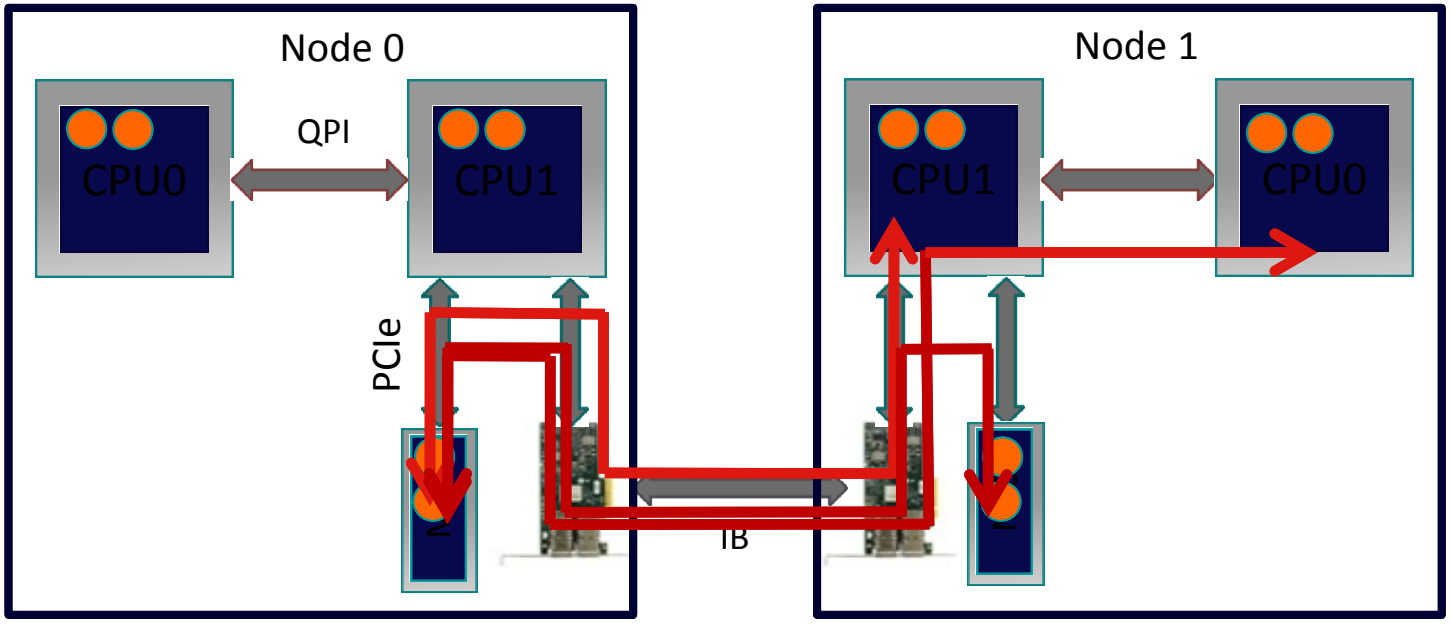



- Full subscription of a MIC: (Host is not involved)
  - 3 MPI processes + 240 OpenMP threads on MIC
  - Input size = 96^3
- Binding using MV2_MIC_MAPPING
- 1 node MVAPICH2-MIC achieves 15 Gflops
- 8 nodes MVAPICH2-MIC achieves 105 GFlops

# MPI Applications on MIC Clusters - InterNode



# MPI Data Movement in Symmetric Mode - InterNode

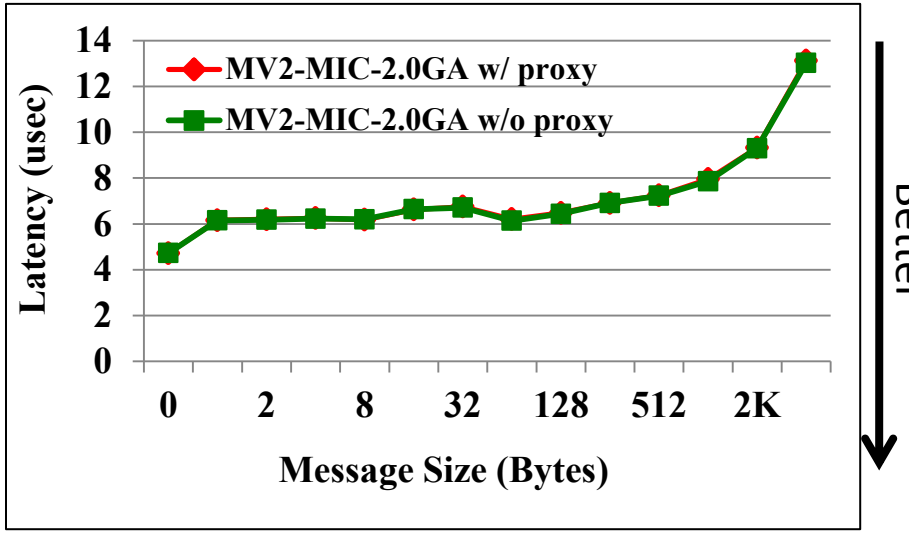



1. MIC- RemoteMIC **2. MIC- RemoteHost (Closer to HCA)**

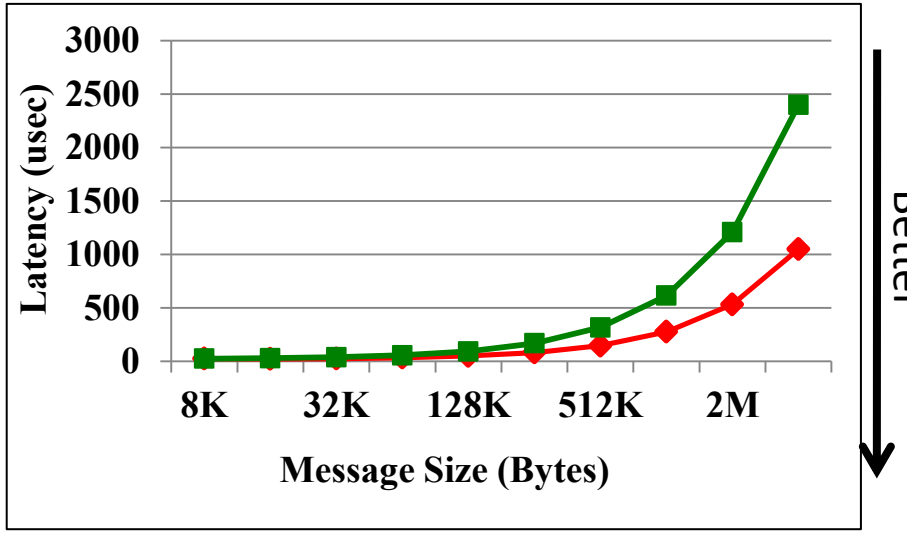
**3. MIC- RemoteHost (Farther from HCA)**

- Uses OFA-IB-CH3 channel backed by host-based proxy

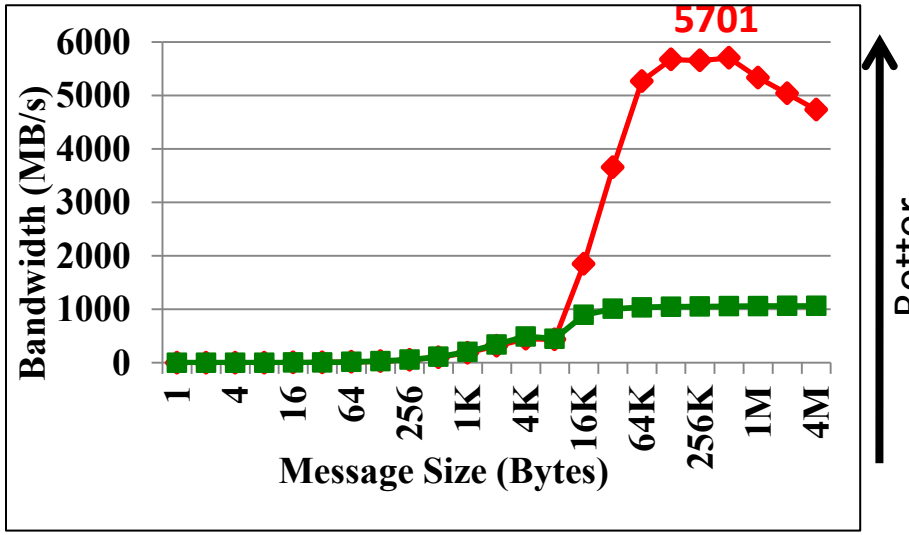
# MIC-RemoteHost Point-to-Point Communication (Active Proxy)



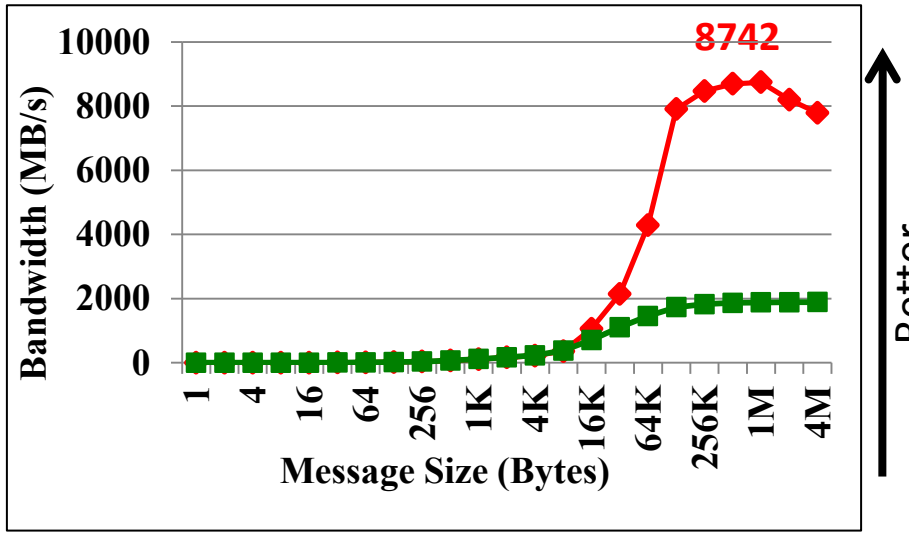

osu_latency (small)




osu_latency (large)



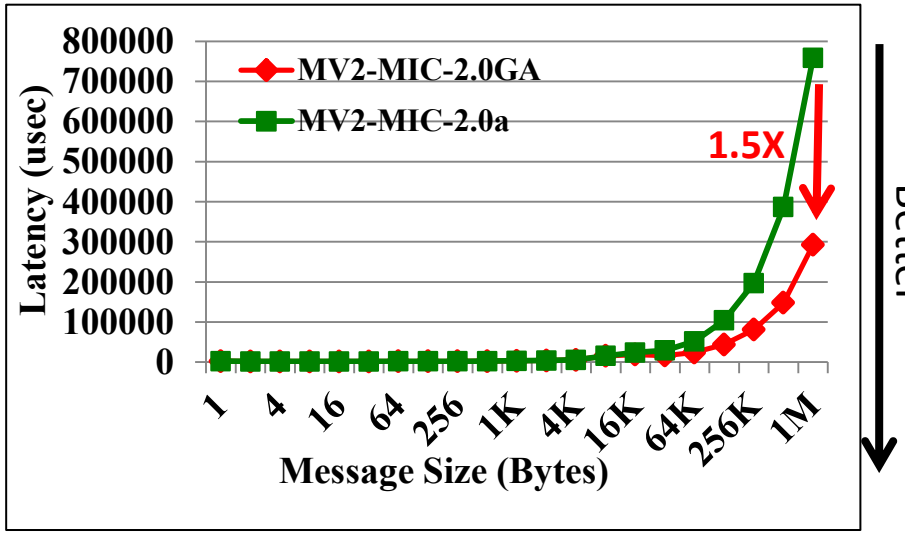

osu_bw



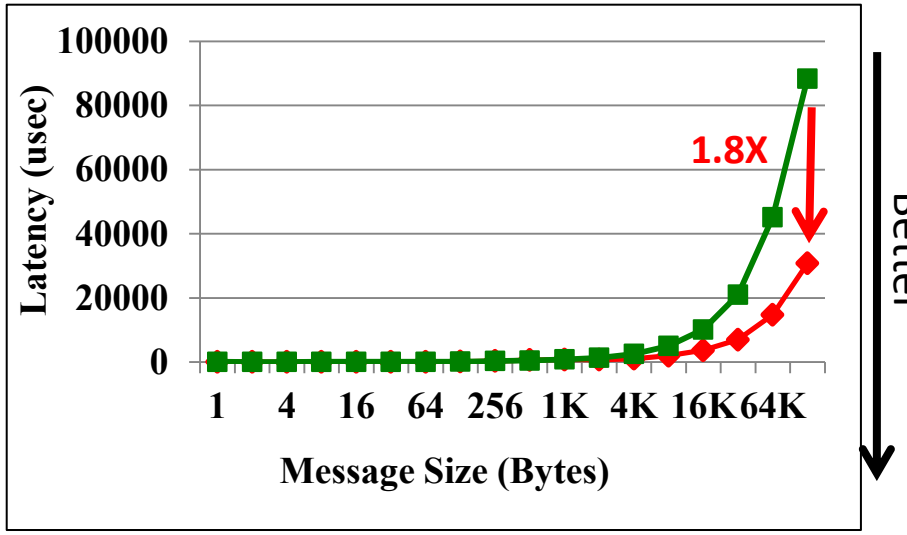

osu_bibw

# Inter-Node Symmetric Mode – Passive Proxy

64 MPI processes on 2 nodes (16H+16M per node)



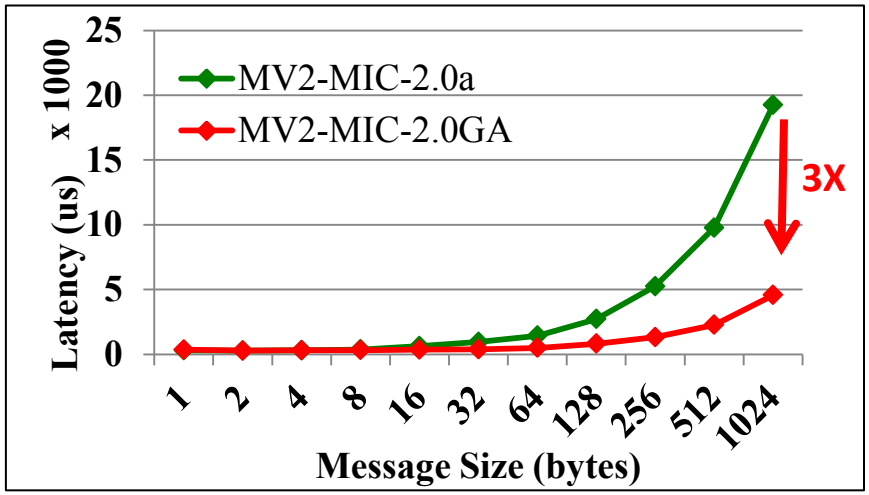

osu_alltoall



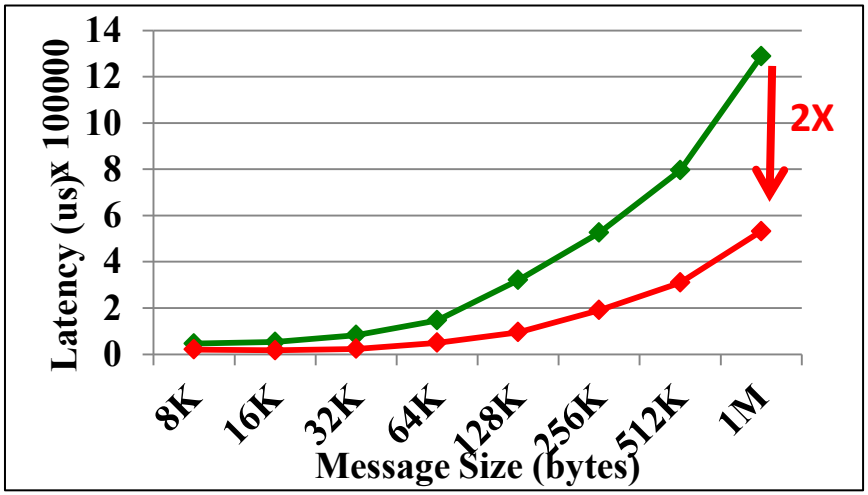

osu_allgather

- Fully-subscribed mode (16 processes on the Host)
- Passive proxy design outperforms the active proxy (default in MV2-MIC-2.0a)

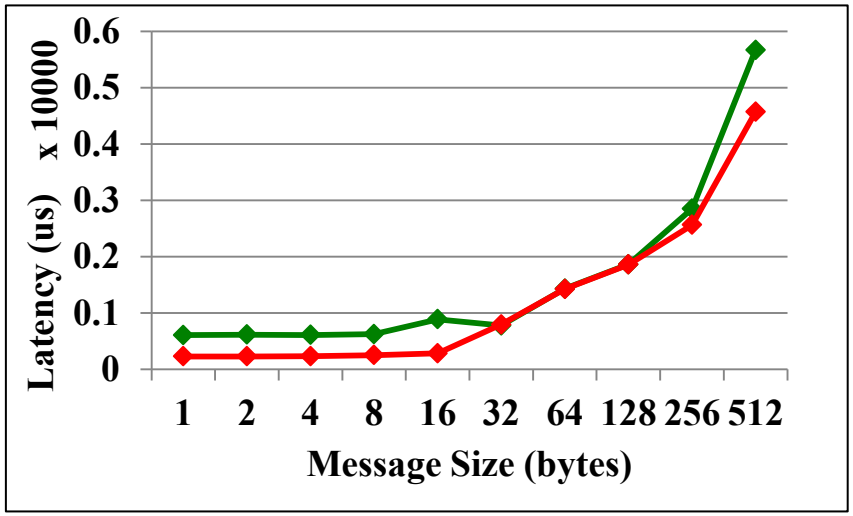
# Inter-Node Symmetric Mode – Redesigning Collectives



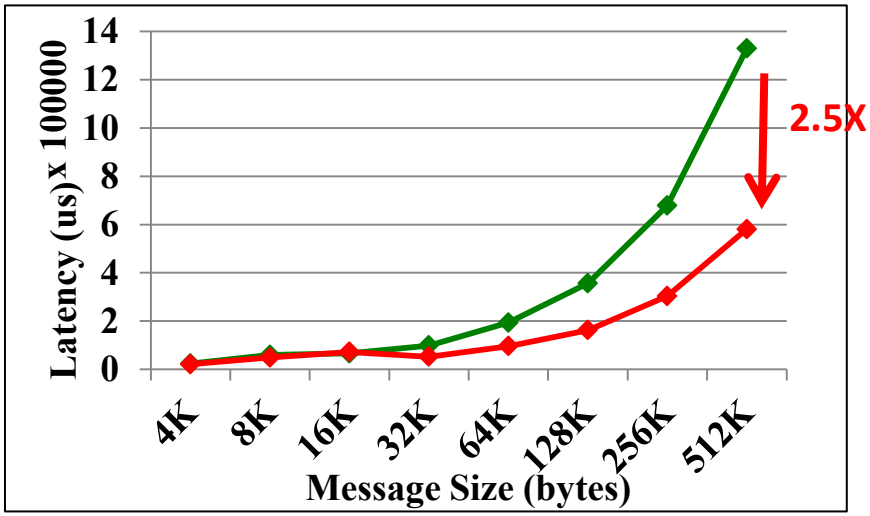

Allgather: 32 Nodes (16H+16M)




Allgather: 32 Nodes (8H+8M)



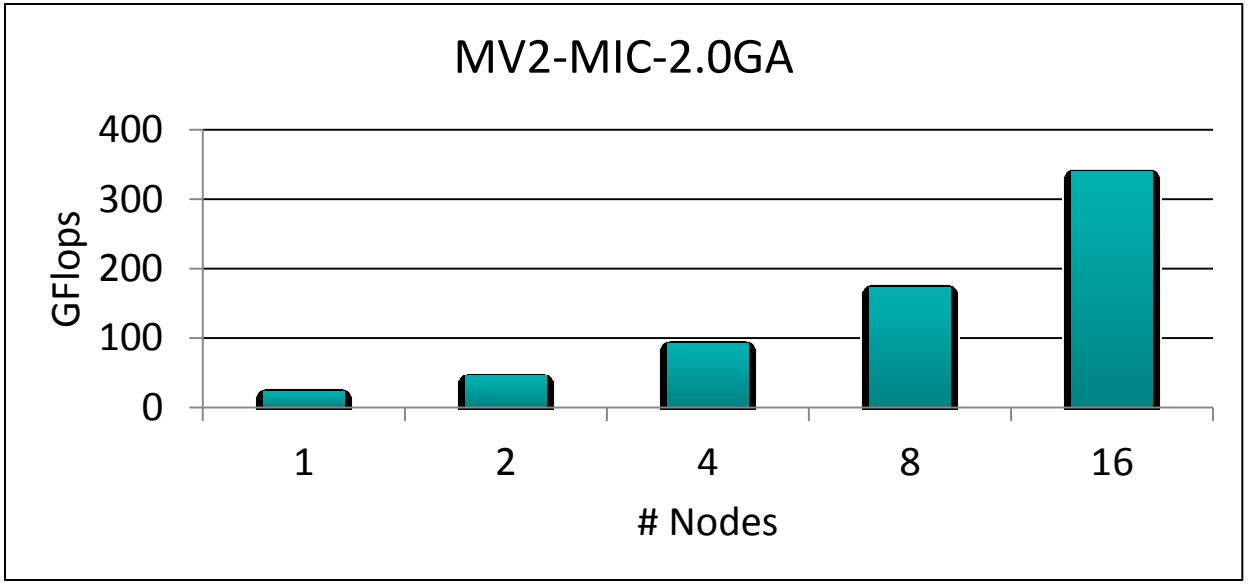

Alltoall: 16 Nodes (8H+8M)




Alltoall: 16 Nodes (8H+8M)

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

# InterNode – Symmetric Mode - HPCG




- Full subscription of a node:
  - 2 MPI processes + 16 OpenMP threads on CPU
  - 3 MPI processes + 240 OpenMP threads on MIC
  - Input size = 96^3
- Binding using both MV2_CPU_MAPPING and MV2_MIC_MAPPING
- To explore different threading levels for host and Xeon Phi
- 1 node MVAPICH2-MIC achieves 23.5 GFlops
- 16 nodes MVAPICH2-MIC achieves 340 GFlops

# On-Going and Future Work

- Redesigning of other collective operations
- Optimizing MPI3-RMA operations for MIC
- Automatic proxy selection using architecture topology detection
- Evaluating application performance and scaling
- Extending designs to KNL-based self-hosted nodes

# Conclusion

- MVAPICH2-MIC provides a high-performance and scalable MPI library for InfiniBand clusters using Xeon Phi
- Initial version takes advantage of SCIF to improve Intra-MIC and Intranode MIC-Host communication
- Host-Proxy based designs help work around P2P bandwidth bottlenecks
- Provides initial support for heterogeneity-aware collectives
- Enhanced version of MVAPICH2-MIC (based on 2.0 GA) will be available soon

# Upcoming 2nd Annual MVAPICH User Group (MUG) Meeting

- August 25-27, 2014; Columbus, Ohio, USA
- Keynote Talks, Invited Talks, Contributed Presentations
- Tutorial on MVAPICH2 and MVAPICH2-X optimization and tuning
- Speakers (Confirmed so far):
  - Keynote: Dan Stanzione (TACC)
  - Keynote: Darren Kerbyson (PNNL)
  - Sadaf Alam (CSCS, Switzerland)
  - Onur Celebioglu (Dell)
  - Jens Glaser (Univ. of Michigan)
  - Jeff Hammond (Intel)
  - Adam Moody (LLNL)
  - David Race (Cray)
  - Davide Rossetti (NVIDIA)
  - Gilad Shainer (Mellanox)
  - Sayantan Sur (Intel)
  - Mahidhar Tatineni (SDSC)
  - Jerome Vienne (TACC)
- More details at: http://mug.mvapich.cse.ohio-state.edu

# Web Pointers

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

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

MVAPICH Web Page

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


panda@cse.ohio-state.edu