The MVAPICH2 Project: Pushing the Frontier of InfiniBand and RDMA Networking Technologies

Talk at OSC/OH-TECH Booth (SC '15)

by

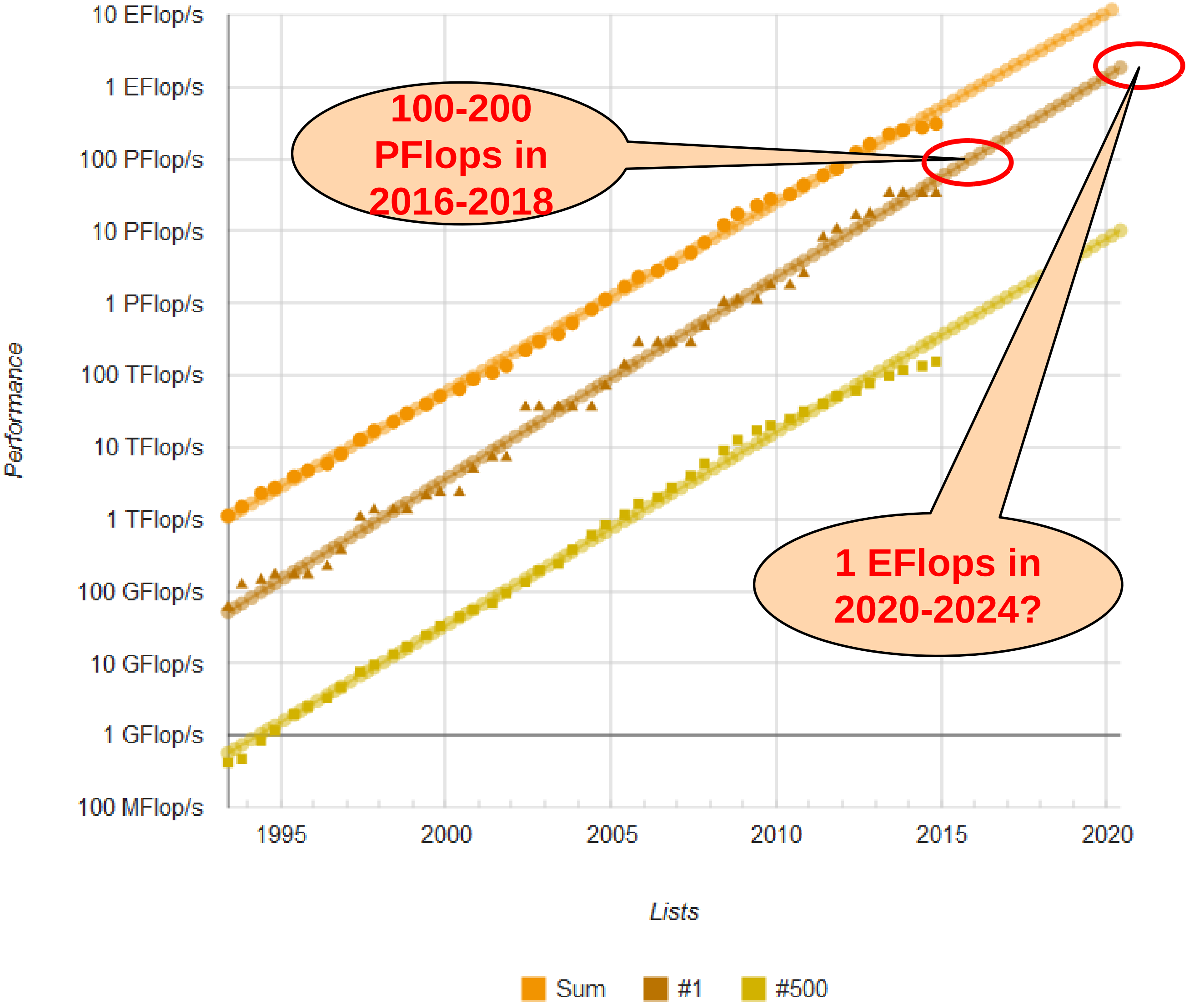
Dhabaleswar K. (DK) Panda

The Ohio State University

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

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

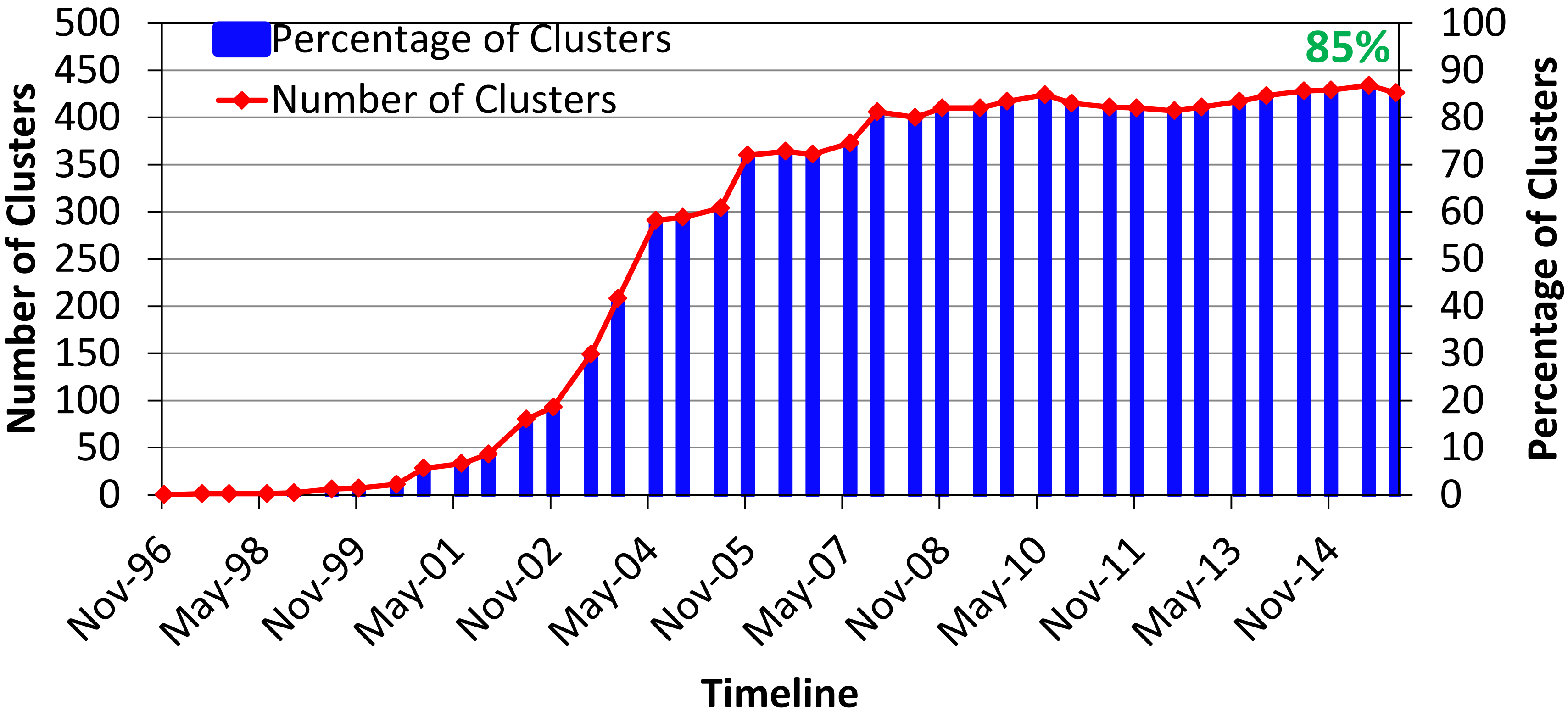
High-End Computing (HEC): PetaFlop to ExaFlop



100-200 PFlops in 2016-2018

1 EFlops in 2020-2024?

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



Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand
<1usec latency, 100Gbps 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)

Tianhe – 2

Titan

Stampede

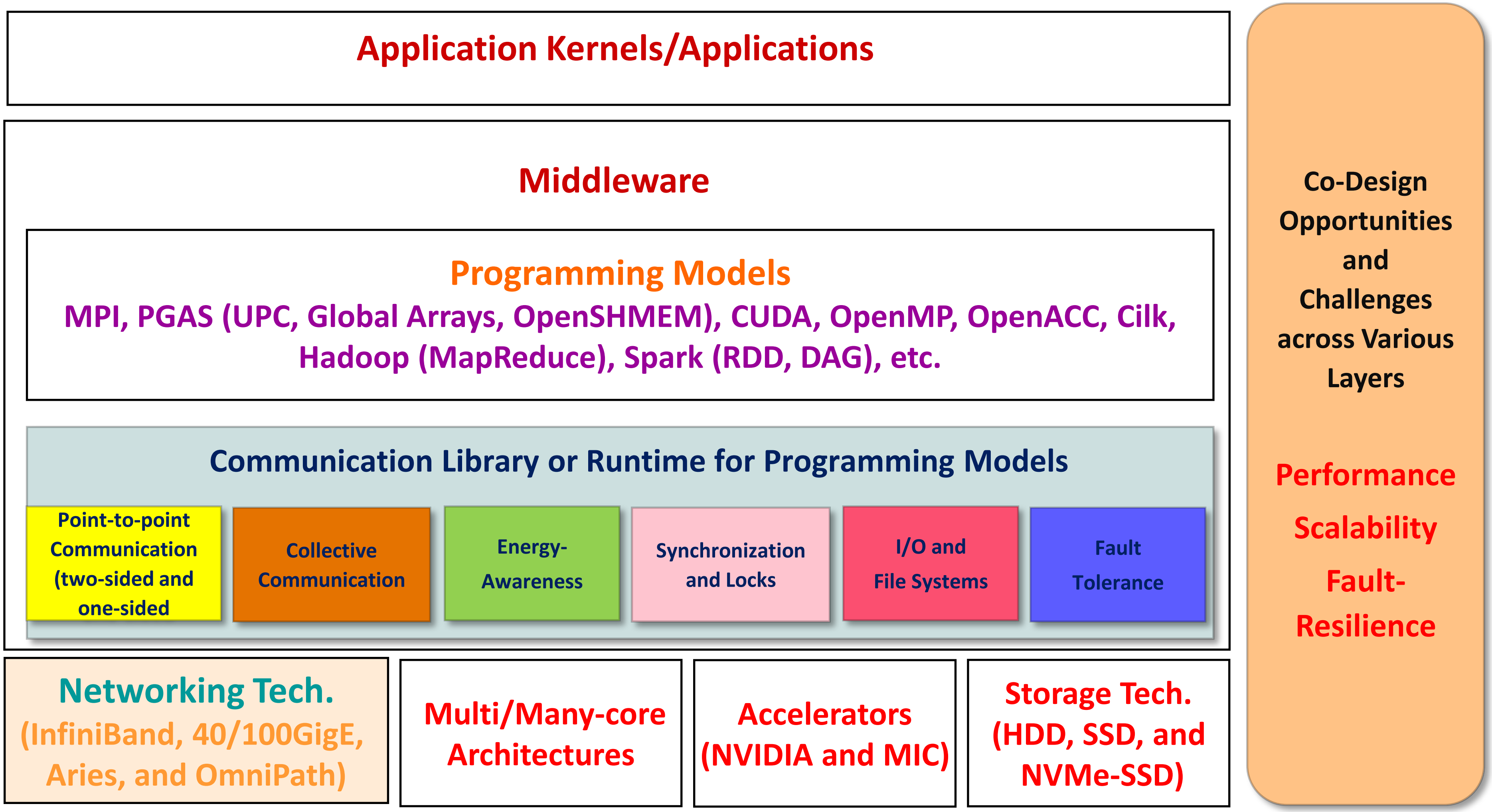
Tianhe – 1A

Large-scale InfiniBand Installations

- 235 IB Clusters (47%) in the Nov' 2015 Top500 list (http://www.top500.org)
- Installations in the Top 50 (21 systems):

462,462 cores (Stampede) at TACC (10th)	76,032 cores (Tsubame 2.5) at Japan/GSIC (25th)
185,344 cores (Pleiades) at NASA/Ames (13th)	194,616 cores (Cascade) at PNNL (27th)
72,800 cores Cray CS-Storm in US (15th)	76,032 cores (Makman-2) at Saudi Aramco (32nd)
72,800 cores Cray CS-Storm in US (16th)	110,400 cores (Pangea) in France (33rd)
265,440 cores SGI ICE at Tulip Trading Australia (17th)	37,120 cores (Lomonosov-2) at Russia/MSU (35th)
124,200 cores (Topaz) SGI ICE at ERDC DSRC in US (18th)	57,600 cores (SwiftLucy) in US (37th)
72,000 cores (HPC2) in Italy (19th)	55,728 cores (Prometheus) at Poland/Cyfronet (38th)
152,692 cores (Thunder) at AFRL/USA (21st)	50,544 cores (Occigen) at France/GENCI-CINES (43rd)
147,456 cores (SuperMUC) in Germany (22nd)	76,896 cores (Salomon) SGI ICE in Czech Republic (47th)
86,016 cores (SuperMUC Phase 2) in Germany (24th)	**and many more!**

Designing High-Performance Middleware for HPC: Challenges



Application Kernels/Applications

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 (two-sided and one-sided

Collective Communication

Energy-Awareness

Synchronization and Locks

I/O and File Systems

Fault Tolerance

Networking Tech. (InfiniBand, 40/100GigE, Aries, and OmniPath)

Multi/Many-core Architectures

Accelerators (NVIDIA and MIC)

Storage Tech. (HDD, SSD, and NVMe-SSD)

Co-Design Opportunities and Challenges across Various Layers

Performance

Scalability

Fault-Resilience

Broad Challenges in Designing Communication Libraries at Exascale

- Scalable Job Startup
- 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
- Support for GPGPUs Support for MICs
- QoS support for communication and I/O

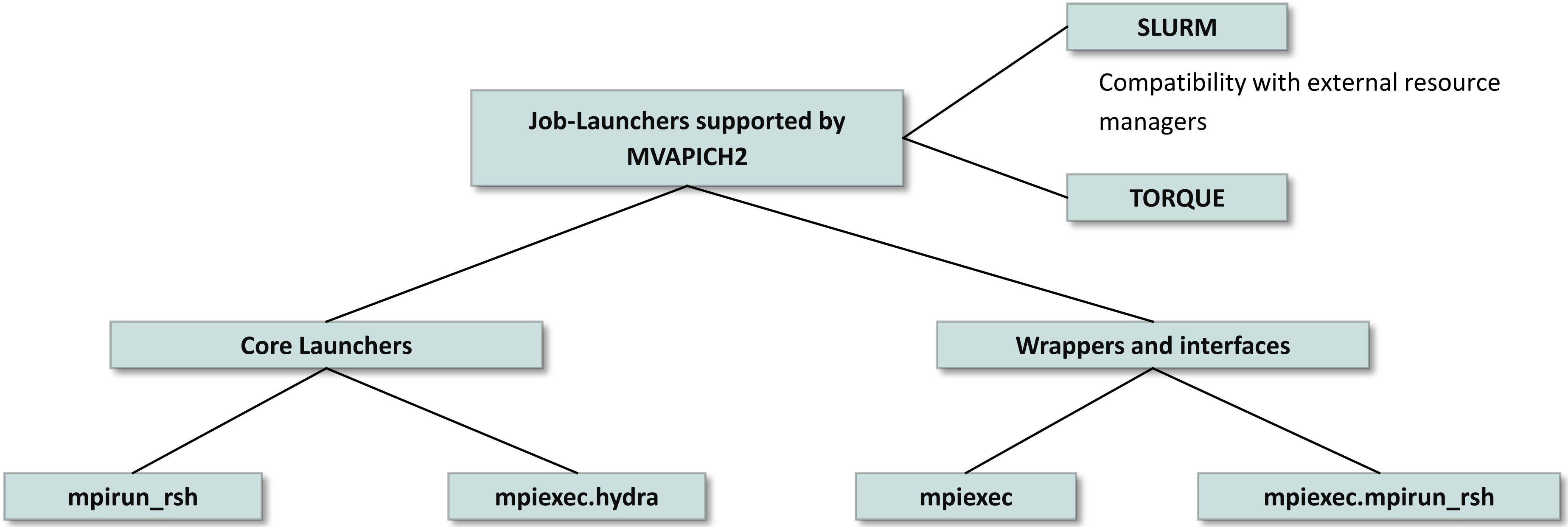
MVAPICH2 Software

- 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)

Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- Scalable Job Startup
- Scalability for million to billion processors
  - Support for highly-efficient Inter-node communication
  - Support for highly-efficient Intra-node communication
  - Support for highly-efficient One-sided / RMA communication
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Support for GPGPUs
- Support for MICs
- QoS support for communication and I/O

Job-Launchers supported by MVAPICH2



Job-Launchers supported by MVAPICH2

SLURM

Compatibility with external resource managers

TORQUE

Core Launchers

Wrappers and interfaces

mpirun_rsh

mpiexec.hydra

mpiexec

mpiexec.mpirun_rsh

mpirun_rsh

- Default and preferred launcher for all interfaces

- Supports scalable hierarchical job-startup mechanism

mpiexec.hydra

Developed at ANL by the MPICH team

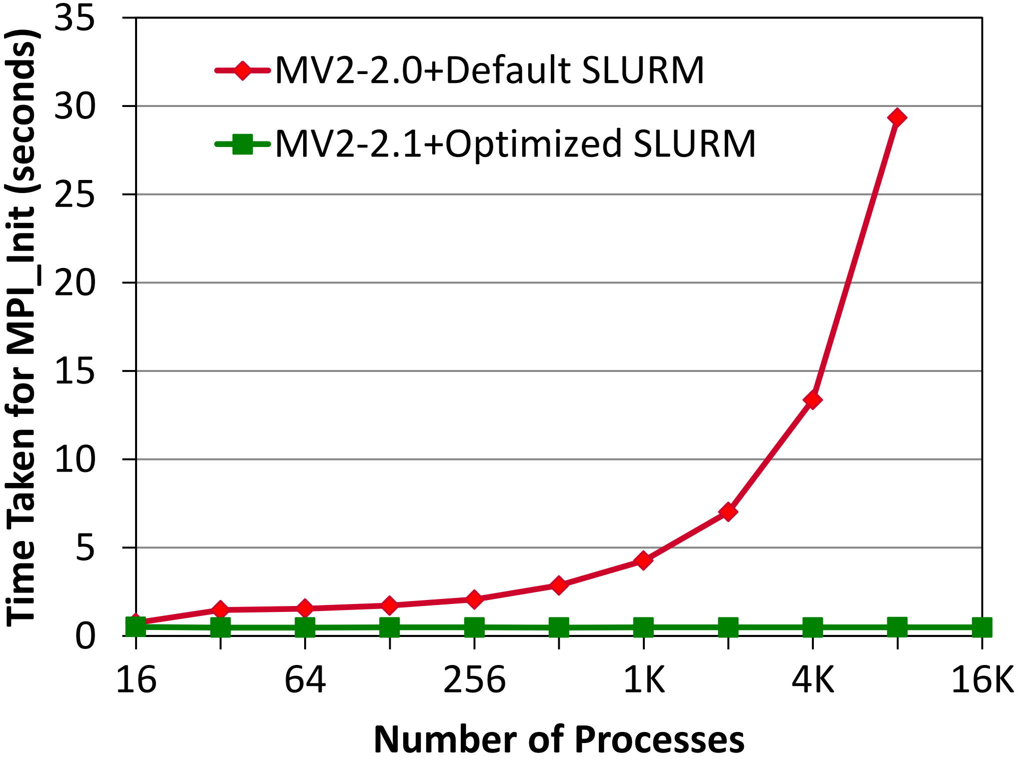
mpiexec

To support interfaces prescribed by the MPI standard

mpiexec.mpirun_rsh

Hydra-like interface to mpirun_rsh

MPI_Init Performance on TACC Stampede



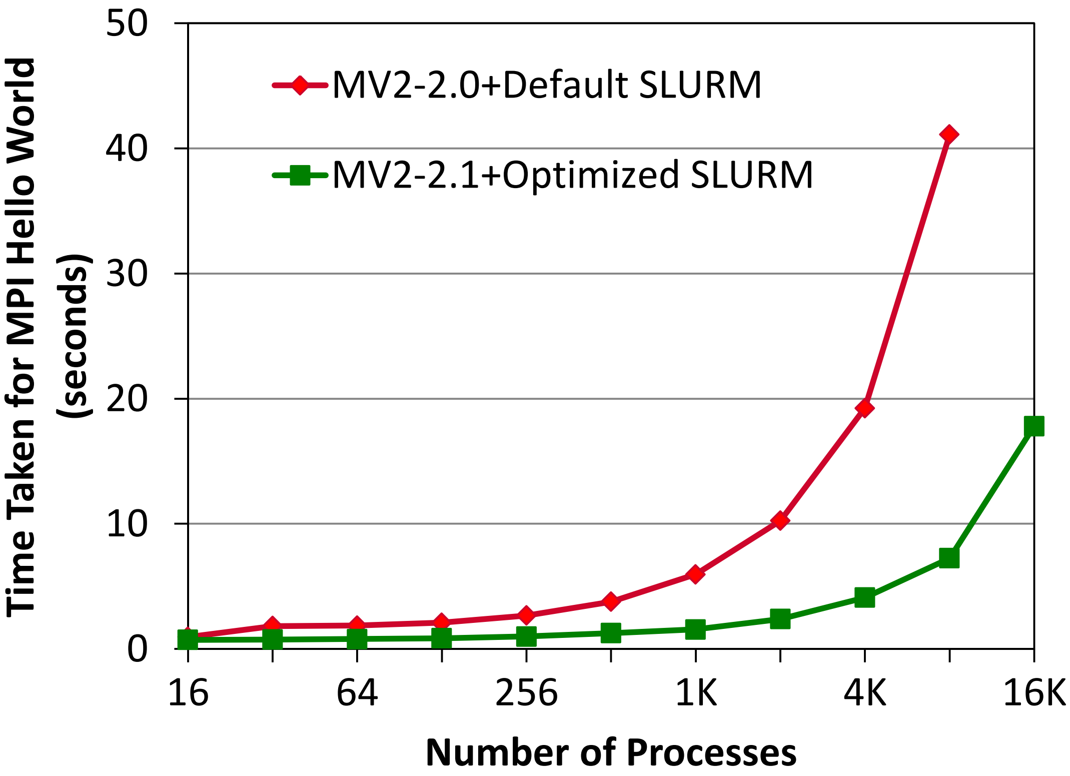
- Near-constant MPI_Init performance
- 59 times improvement at 8,192 processes (512 nodes)
- New designs show good scaling with 16K processes and above

"Non-blocking PMI Extensions for Fast MPI Startup"

S. Chakraborty, H. Subramoni, A. Moody, A. Venkatesh, J. Perkins and D. K. Panda

Int'l Symposium on Cluster, Cloud, and Grid Computing (CCGrid '15)

MPI Hello World Performance on TACC Stampede



- PMI Exchange costs overlapped with application computation
- 5.7 times improvement at 8,192 processes (512 nodes)
- *New designs to be available as part of upcoming releases*

"Non-blocking PMI Extensions for Fast MPI Startup"

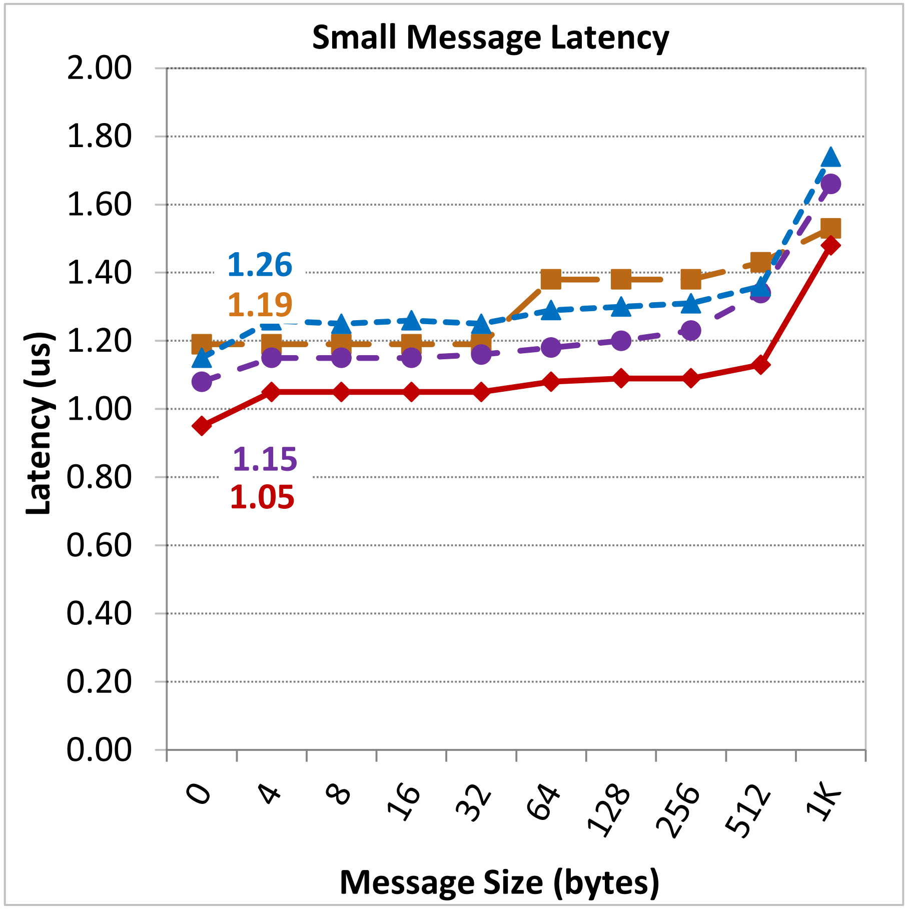
S. Chakraborty, H. Subramoni, A. Moody, A. Venkatesh, J. Perkins and D. K. Panda

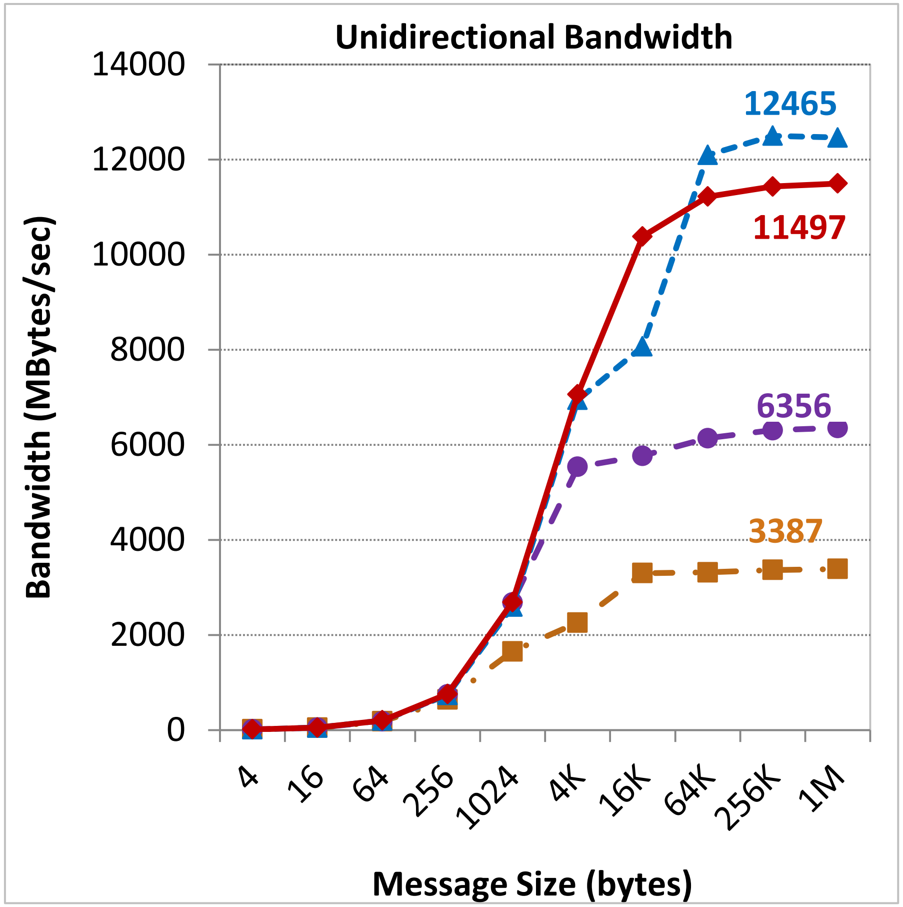
Int'l Symposium on Cluster, Cloud, and Grid Computing (CCGrid '15)

Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- Scalable Job Startup
- Scalability for million to billion processors
  - Support for highly-efficient Inter-node communication
  - Support for highly-efficient Intra-node communication
  - Support for highly-efficient One-sided / RMA communication
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Support for GPGPUs
- Support for MICs
- QoS support for communication and I/O

Latency & Bandwidth: MPI over IB with MVAPICH2





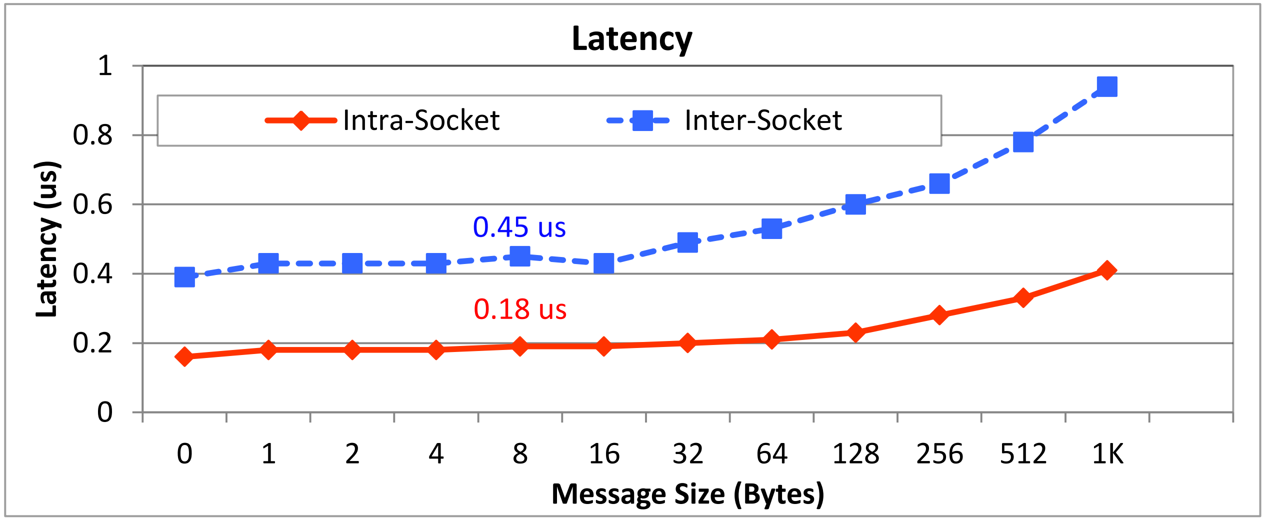
TrueScale-QDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

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

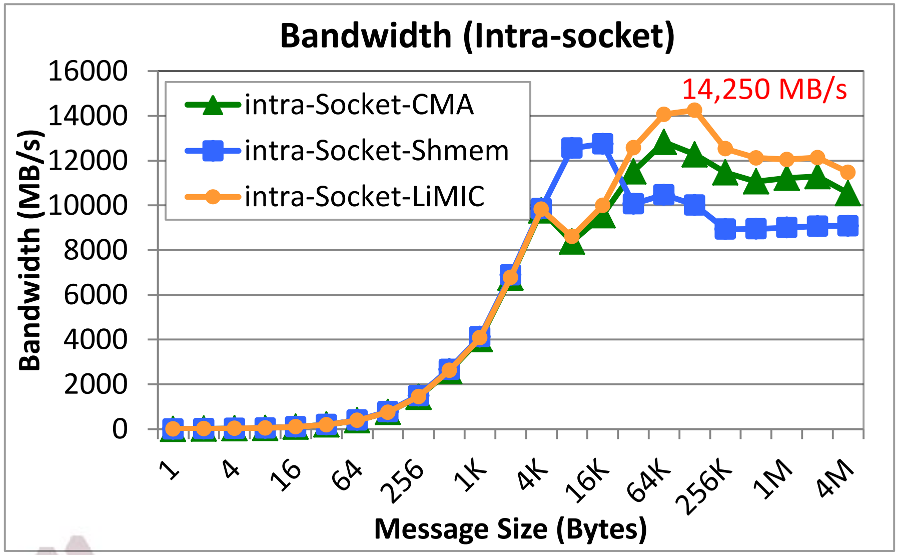
ConnectX-4-EDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 Back-to-back

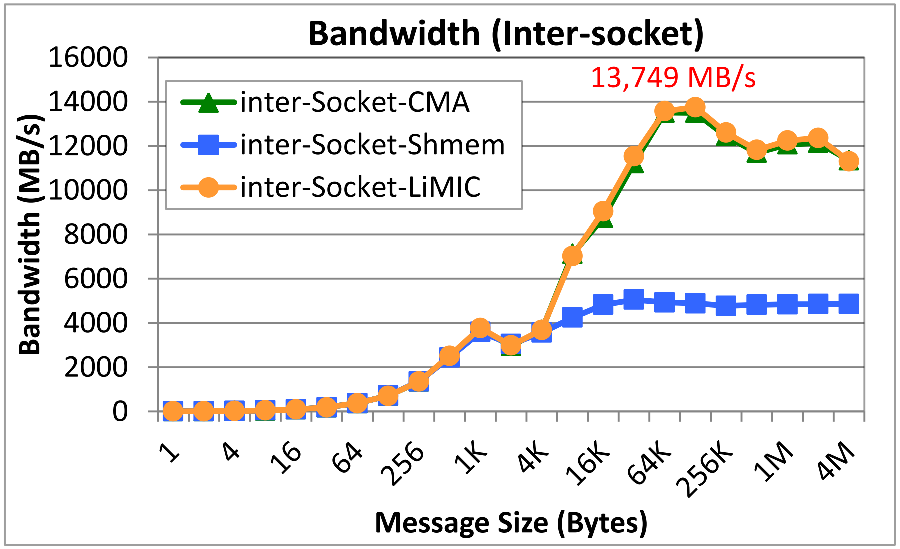
MVAPICH2 Two-Sided Intra-Node Performance
(Shared memory and Kernel-based Zero-copy Support (LiMIC and CMA))



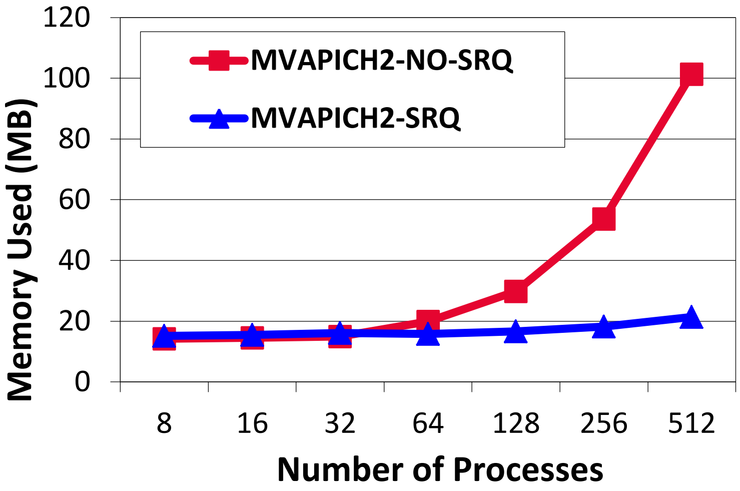
Latest MVAPICH2 2.2a

Intel Ivy-bridge

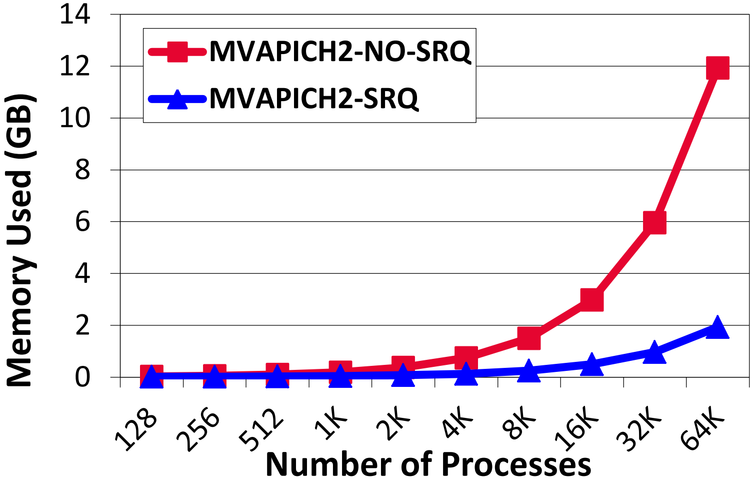




Memory Utilization using Shared Receive Queues, UD



MPI_Init memory utilization

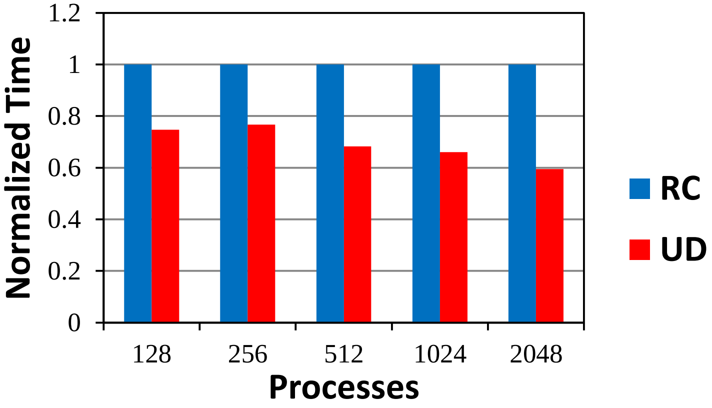


Analytical model

- SRQ reduces the memory used by 1/6th at 64,000 processes

S. Sur, L. Chai, H. –W. Jin and D. K. Panda, "Shared Receive Queue Based Scalable MPI Design for InfiniBand Clusters", IPDPS 2006

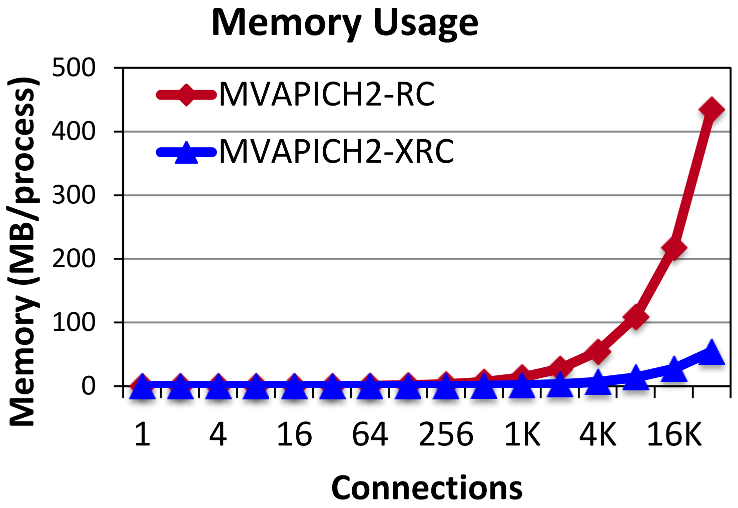
	RC (MVAPICH2 1.8)				UD (MVAPICH2 1.8)		
Number of Processes	Conn.	Buffers	Struct	Total	Buffers	Struct	Total
512	22.9	24	0.3	47.2	24	0.2	24.2
1024	29.5	24	0.6	54.1	24	0.4	24.4
2048	42.4	24	1.2	67.6	24	0.9	24.9

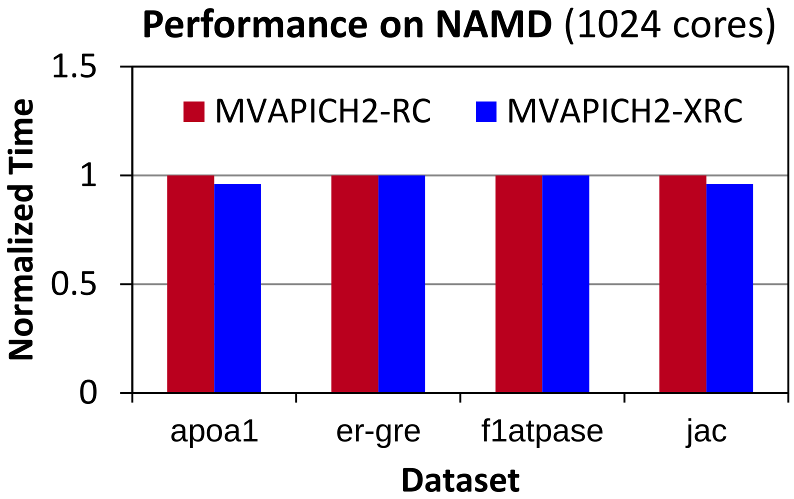


- UD reduces HCA QP cache trashing

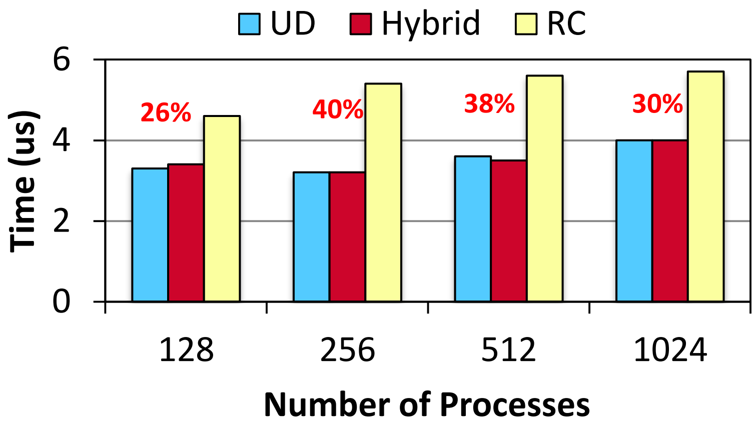
M. Koop, S. Sur, Q. Gao and D. K. Panda, "High Performance MPI Design using Unreliable Datagram for Ultra-Scale InfiniBand Clusters," ICS '07

eXtended Reliable Connection (XRC) and Hybrid Mode





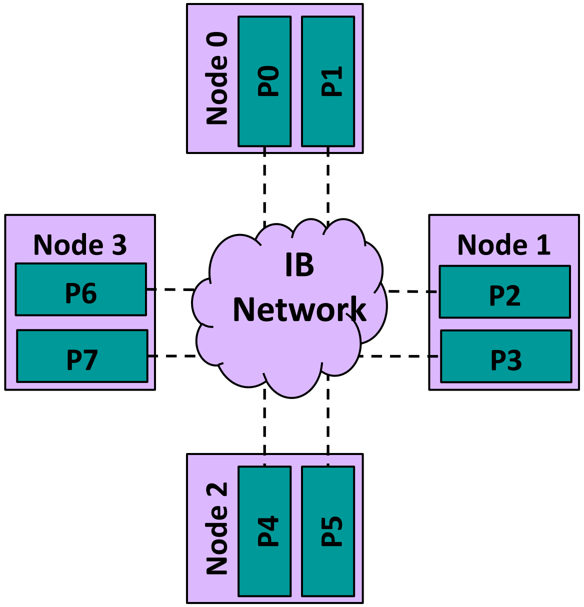
- Memory usage for 32K processes with 8-cores per node can be 54 MB/process (for connections)
- NAMD performance improves when there is frequent communication to many peers



- Both UD and RC/XRC have benefits
  - Hybrid for the best of both
- Available since MVAPICH2 1.7 as integrated interface
- Runtime Parameters: RC - default;
  - **UD - MV2_USE_ONLY_UD=1**
  - **Hybrid - MV2_HYBRID_ENABLE_THRESHOLD=1**

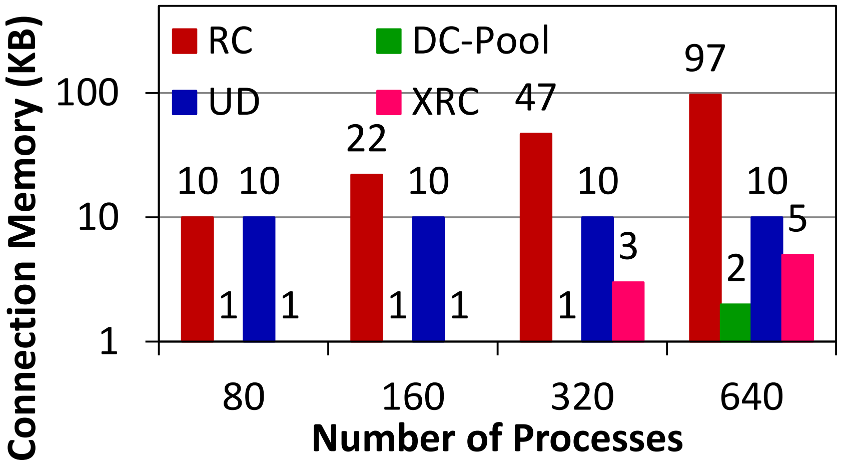
M. Koop, J. Sridhar and D. K. Panda, "Scalable MPI Design over InfiniBand using eXtended Reliable Connection," Cluster '08

Dynamic Connected (DC) Transport in MVAPICH2

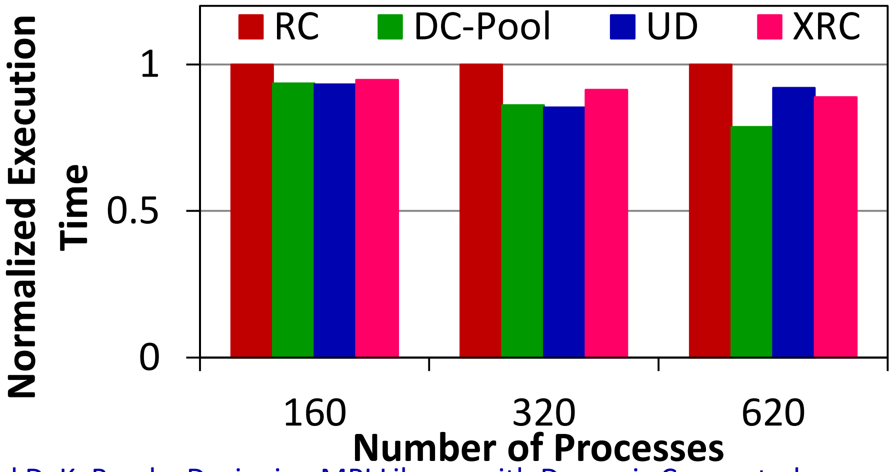


- Constant connection cost (*One QP for any peer*)
- Full Feature Set (RDMA, Atomics etc)
- Separate objects for send (DC Initiator) and receive (DC Target)
  - DC Target identified by "DCT Number"
  - Messages routed with (DCT Number, LID)
  - Requires same "DC Key" to enable communication
- Initial study done in MVAPICH2
- DCT support available in Mellanox OFED – 2.2.0.1

Memory Footprint for Alltoall



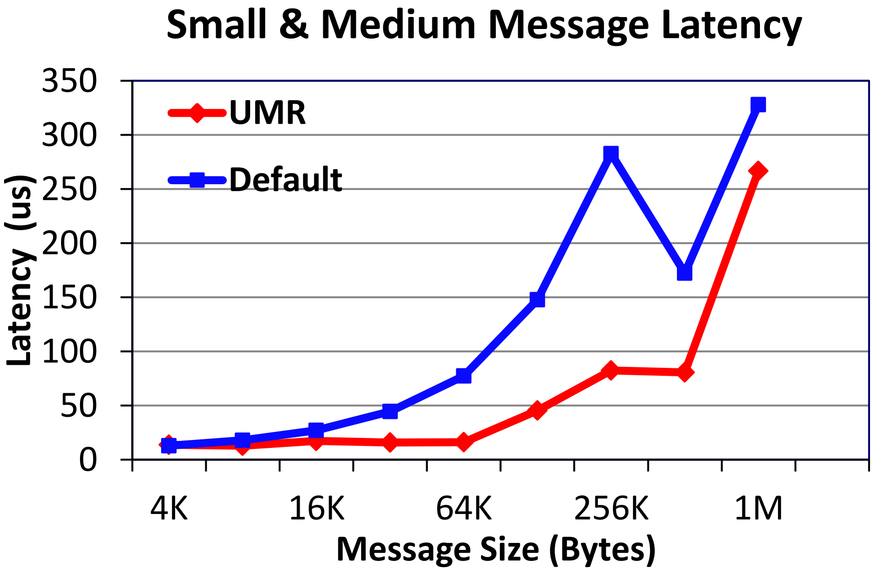
NAMD - Apoa1: Large data set

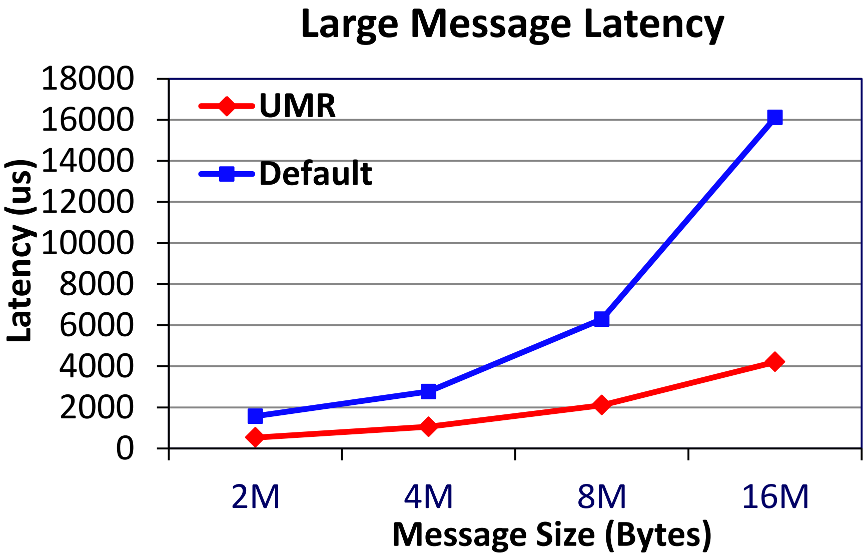


H. Subramoni, K. Hamidouche, A. Venkatesh, S. Chakraborty and D. K. Panda, Designing MPI Library with Dynamic Connected Transport (DCT) of InfiniBand : Early Experiences. IEEE International Supercomputing Conference (ISC '14).

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



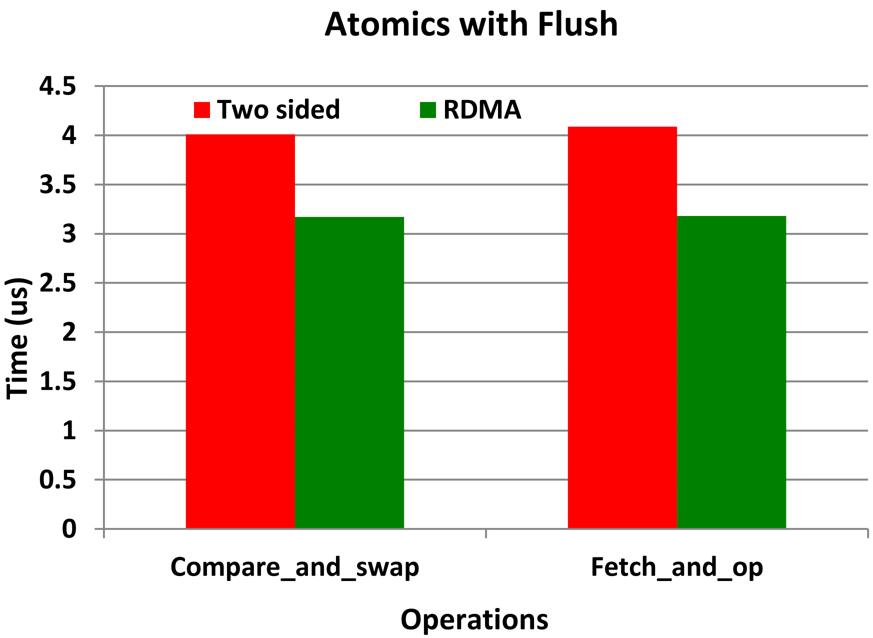


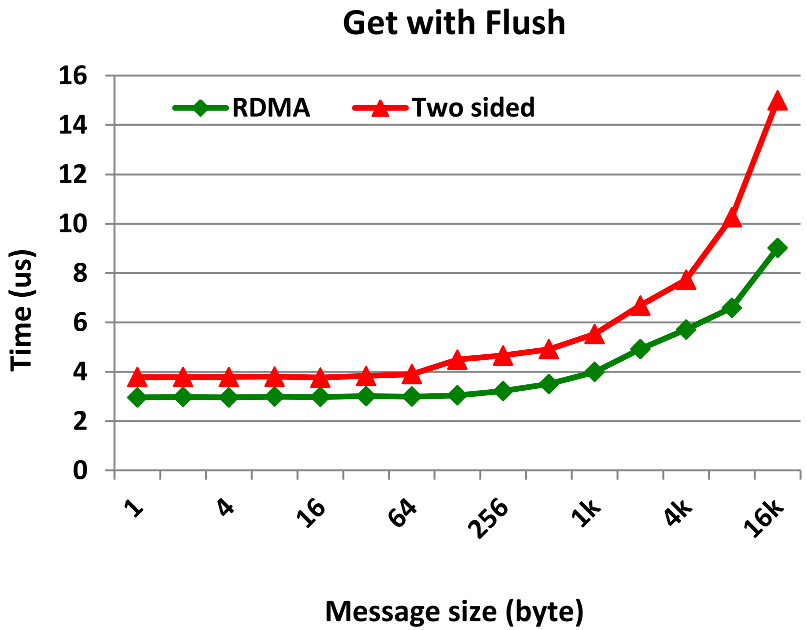
Connect-IB (54 Gbps): 2.8 GHz Dual Ten-core (IvyBridge) Intel PCI Gen3 with Mellanox IB FDR switch

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

MPI-3 RMA Model: Performance

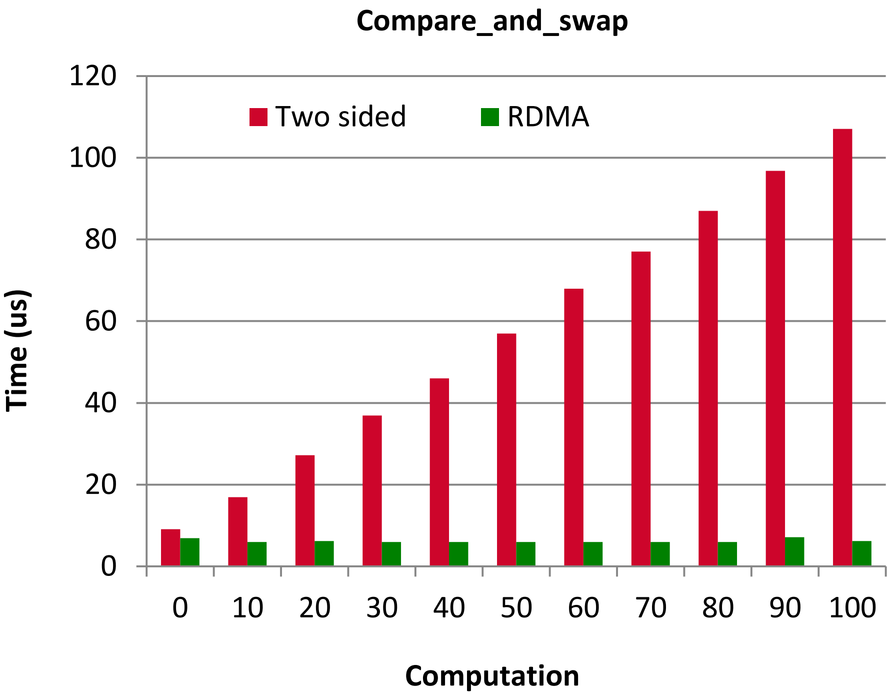
- RDMA-based and truly one-sided implementation of MPI-3 RMA in progress

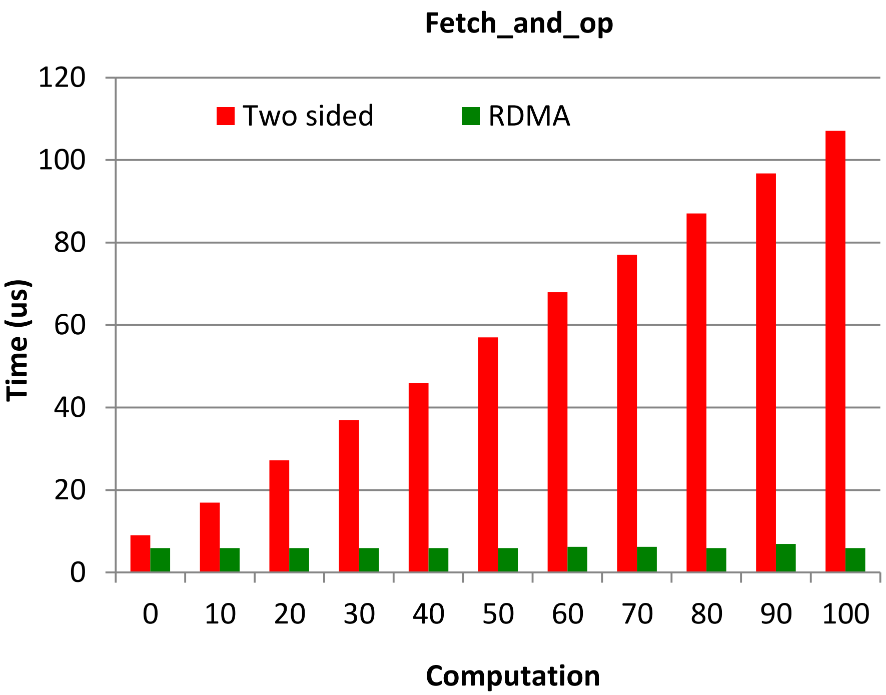




- MVAPICH2-2.1 and OSU micro-benchmarks (OMB v4.1)
- Better performance for MPI_Compare_and_swap and MPI_Fetch_and_op and MPI_Get performance with RDMA-based design

MPI-3 RMA Model: Overlap





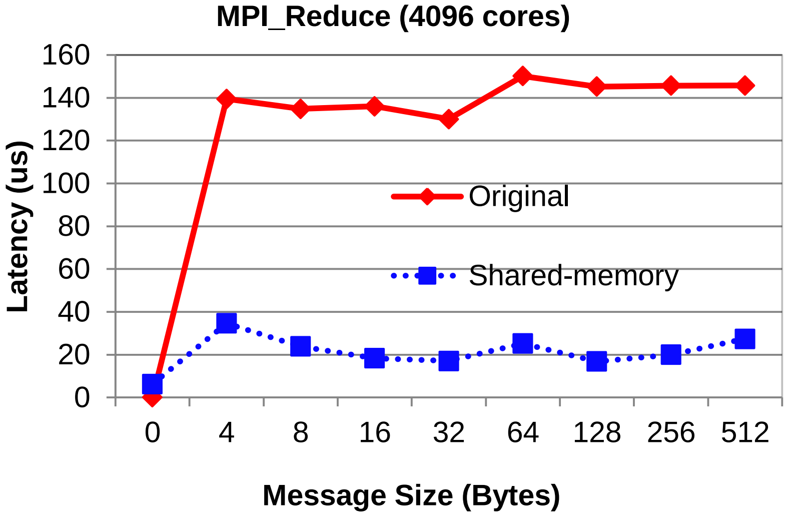
- Process 0 is busy in computation, Process 1 performance atomic operations at P0
- These benchmarks show the latency of atomic operations. For RDMA based design, the atomic latency at P1 remains consistent even as the busy time at P0 increases

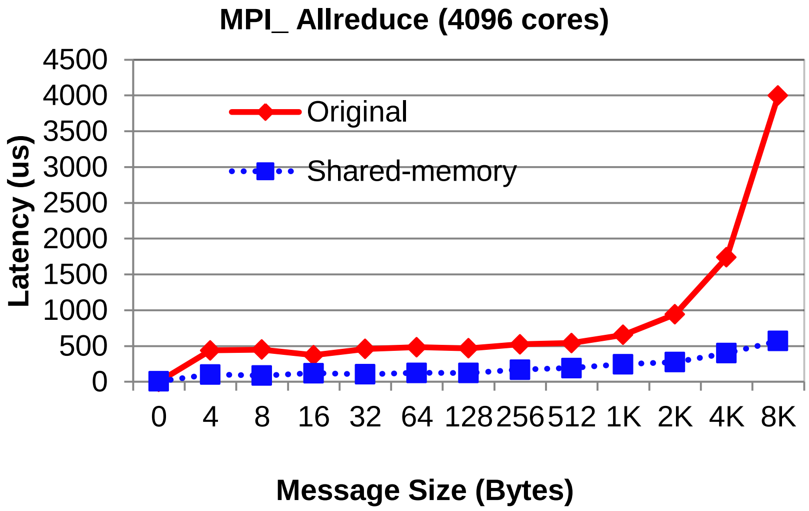
Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- Scalable Job Startup
- Scalability for million to billion processors
  - Support for highly-efficient Inter-node communication
  - Support for highly-efficient Intra-node communication
  - Support for highly-efficient One-sided / RMA communication
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Support for GPGPUs
- Support for MICs
- QoS support for communication and I/O

Shared-memory Aware Collectives

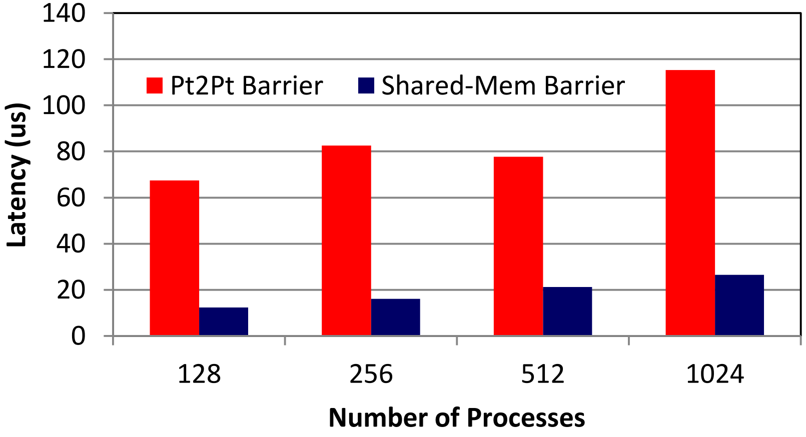
- MVAPICH2 Reduce/Allreduce with 4K cores on TACC Ranger (AMD Barcelona, SDR IB)





MV2_USE_SHMEM_REDUCE=0/1

MV2_USE_SHMEM_ALLREDUCE=0/1

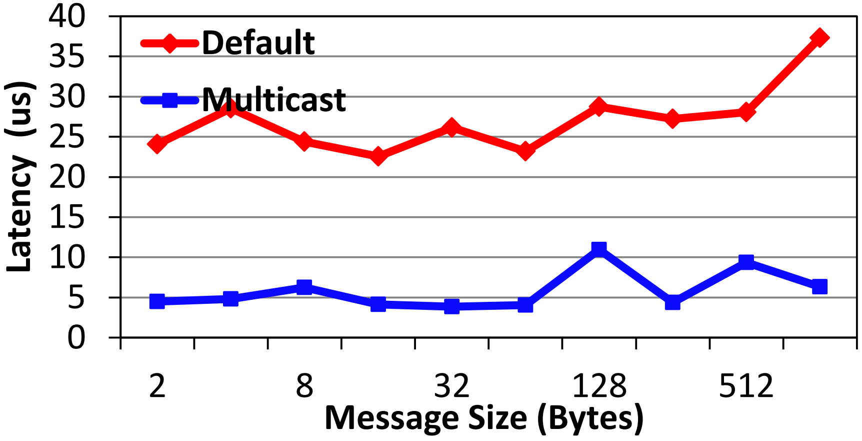


- MVAPICH2 Barrier with 1K Intel Westmere cores , QDR IB

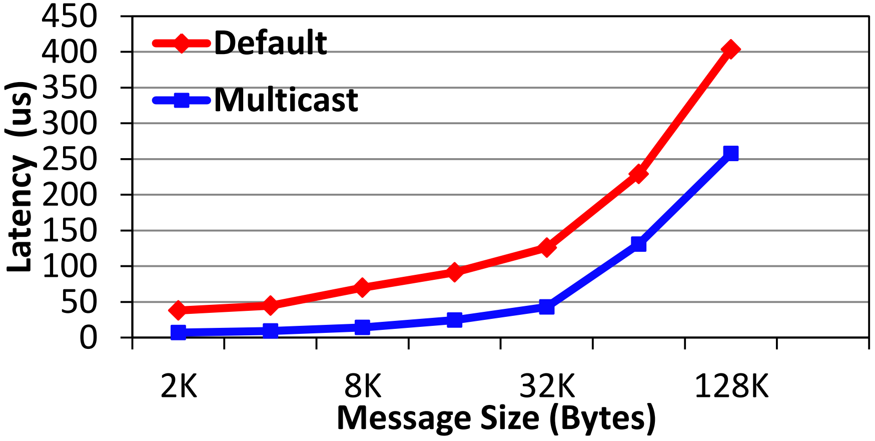
MV2_USE_SHMEM_BARRIER=0/1

Hardware Multicast-aware MPI_Bcast on Stampede

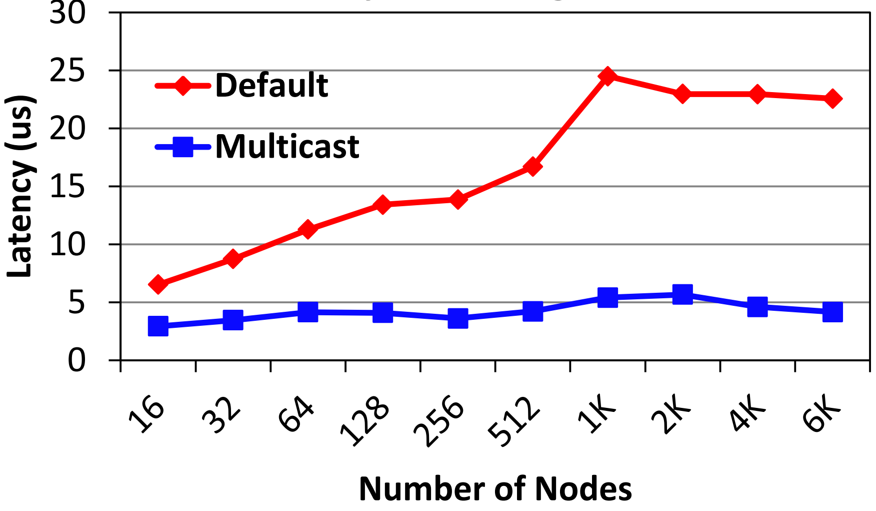
Small Messages (102,400 Cores)



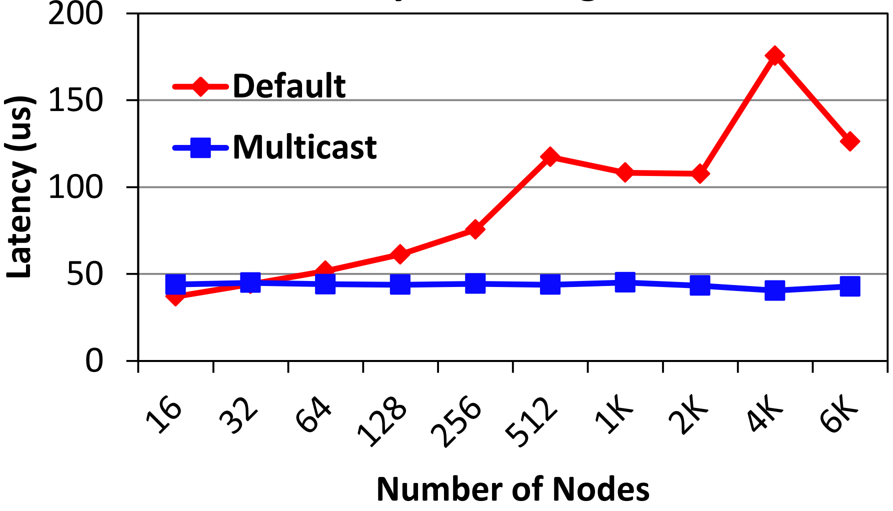
Large Messages (102,400 Cores)



16 Byte Message

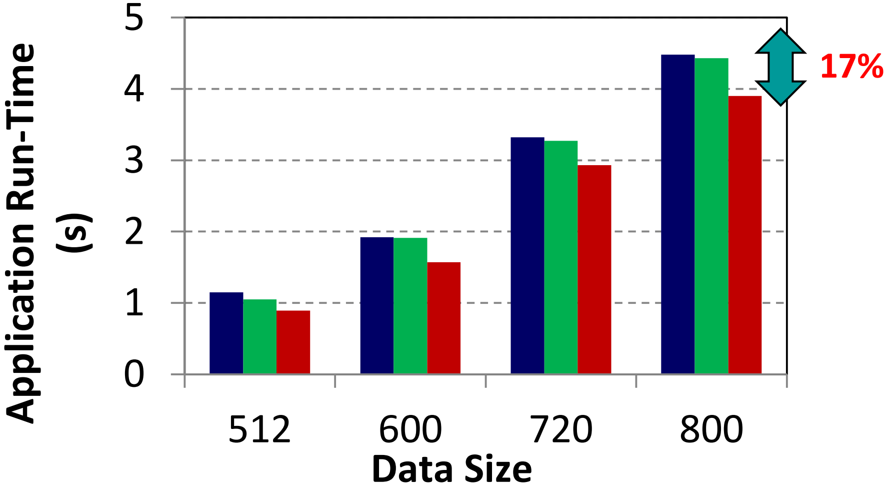


32 KByte Message

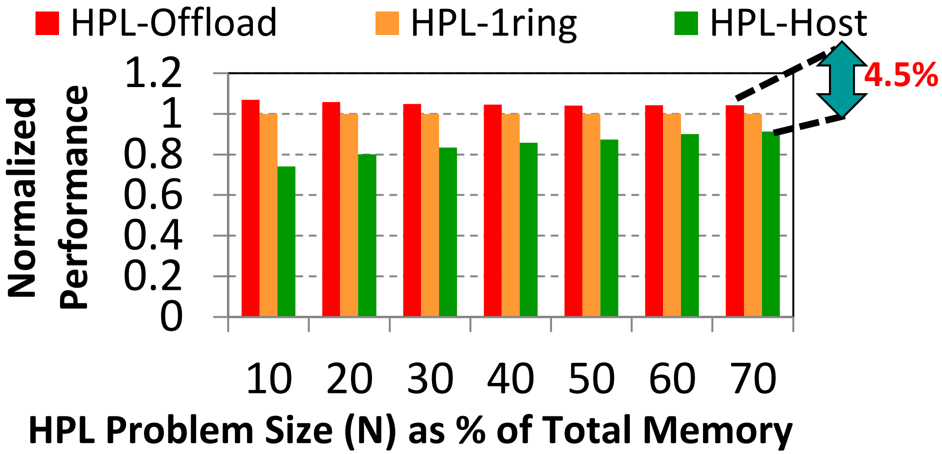


ConnectX-3-FDR (54 Gbps): 2.7 GHz Dual Octa-core (SandyBridge) Intel PCI Gen3 with Mellanox IB FDR switch

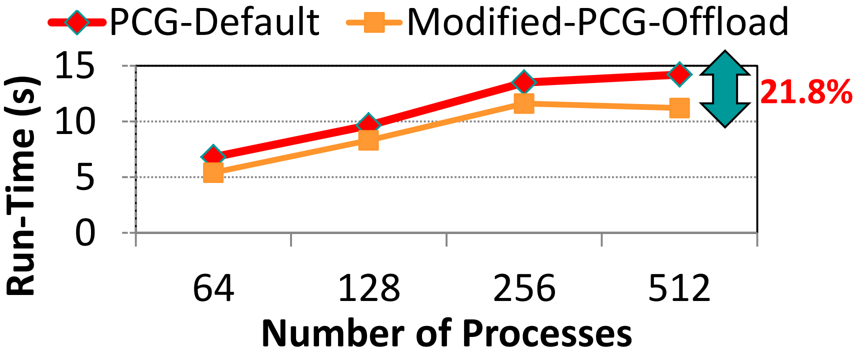
Application benefits with Non-Blocking Collectives based on CX-2 Collective Offload



Modified P3DFFT with Offload-Alltoall does up to 17% better than default version (128 Processes)



Modified HPL with Offload-Bcast does up to 4.5% better than default version (512 Processes)



Modified Pre-Conjugate Gradient Solver with Offload-Allreduce does up to 21.8% better than default version

K. Kandalla, et. al.. High-Performance and Scalable Non-Blocking All-to-All with Collective Offload on InfiniBand Clusters: A Study with Parallel 3D FFT. ISC 2011

K. Kandalla, et. al, Designing Non-blocking Broadcast with Collective Offload on InfiniBand Clusters: A Case Study with HPL, HotI 2011

K. Kandalla, et. al., Designing Non-blocking Allreduce with Collective Offload on InfiniBand Clusters: A Case Study with Conjugate Gradient Solvers, IPDPS '12

Can Network-Offload based Non-Blocking Neighborhood MPI Collectives Improve Communication Overheads of Irregular Graph Algorithms? K. Kandalla, A. Buluc, H. Subramoni, K. Tomko, J. Vienne, L. Oliker, and D. K. Panda, IWPAPS' 12

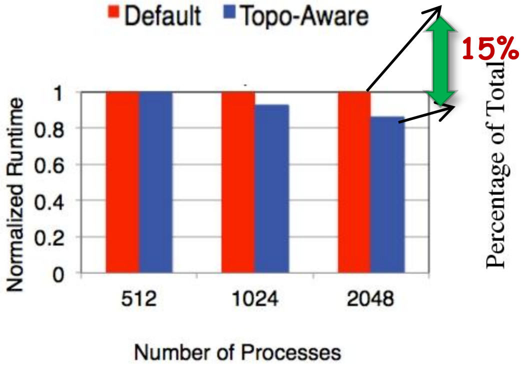
Network-Topology-Aware Placement of Processes

Can we design a highly scalable network topology detection service for IB?

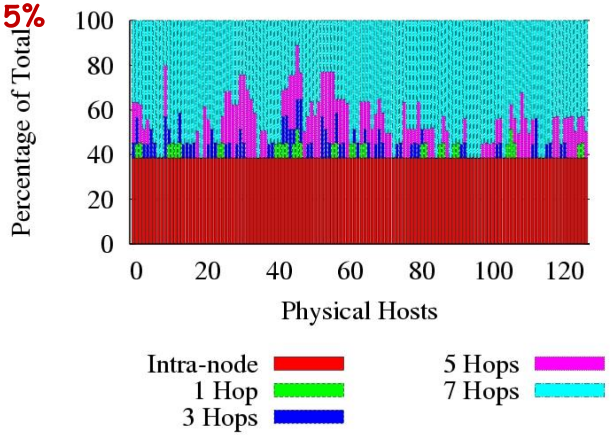
How do we design the MPI communication library in a network-topology-aware manner to efficiently leverage the topology information generated by our service?

What are the potential benefits of using a network-topology-aware MPI library on the performance of parallel scientific applications?

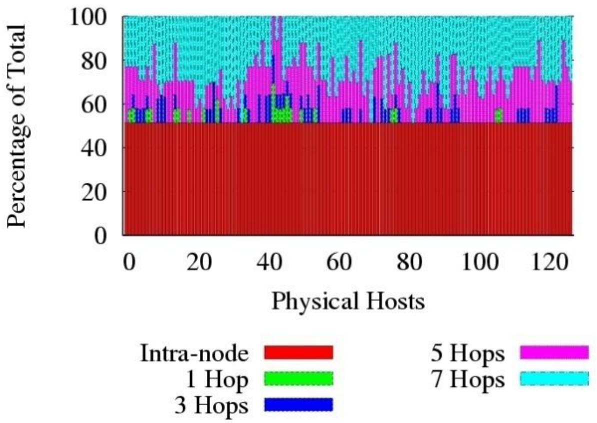
Overall performance and Split up of physical communication for MILC on Ranger



Performance for varying system sizes



Default for 2048 core run



Topo-Aware for 2048 core run

- Reduce network topology discovery time from $O(N^2_{hosts})$ to $O(N_{hosts})$
- 15% improvement in MILC execution time @ 2048 cores
- 15% improvement in Hypre execution time @ 1024 cores

H. Subramoni, S. Potluri, K. Kandalla, B. Barth, J. Vienne, J. Keasler, K. Tomko, K. Schulz, A. Moody, and D. K. Panda, Design of a Scalable InfiniBand Topology Service to Enable Network-Topology-Aware Placement of Processes, SC'12 . BEST Paper and BEST STUDENT Paper Finalist

Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- Scalable Job Startup
- Scalability for million to billion processors
  - Support for highly-efficient Inter-node communication
  - Support for highly-efficient Intra-node communication
  - Support for highly-efficient One-sided / RMA communication
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Support for GPGPUs
- Support for MICs
- QoS support for communication and I/O

MPI + CUDA - Naive

- Data movement in applications with standard MPI and CUDA interfaces

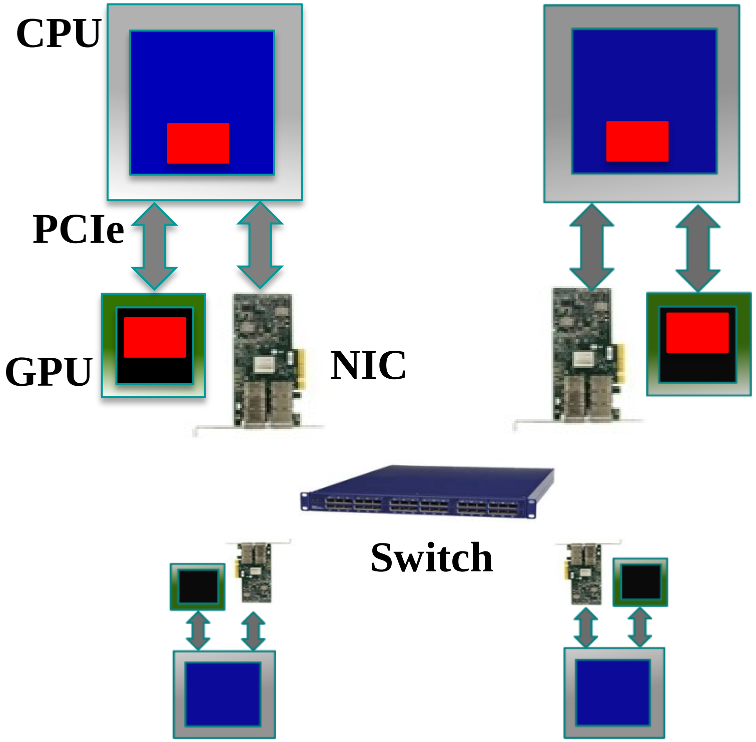
At Sender:

```
cudaMemcpy(s_hostbuf, s_devbuf, . . .);
MPI_Send(s_hostbuf, size, . . .);
```

At Receiver:

```
MPI_Recv(r_hostbuf, size, . . .);
cudaMemcpy(r_devbuf, r_hostbuf, . . .);
```

High Productivity and Low Performance



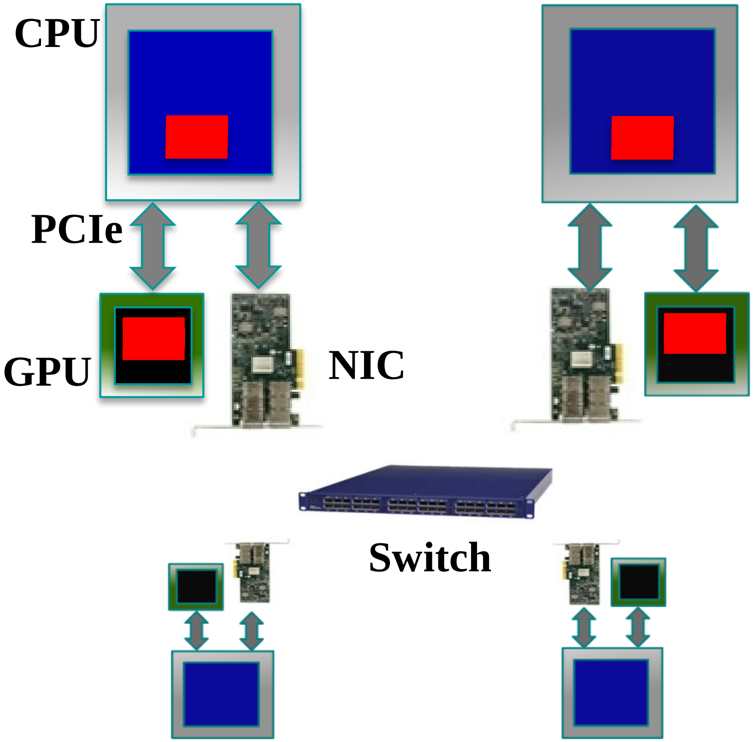
MPI + CUDA - Advanced

- Pipelining at user level with non-blocking MPI and CUDA interfaces

At Sender:

```
for (j = 0; j < pipeline_len; j++)
    cudaMemcpyAsync(s_hostbuf + j * blk, s_devbuf + j * blksz,
      …);
for (j = 0; j < pipeline_len; j++) {
      while (result != cudaSucess) {
           result = cudaStreamQuery(…);
            if(j > 0) MPI_Test(…);
       }
       MPI_Isend(s_hostbuf + j * block_sz, blksz . . .);
 }
MPI_Waitall();
```

<<Similar at receiver>>



Low Productivity and High Performance

GPU-Aware MPI Library: MVAPICH2-GPU

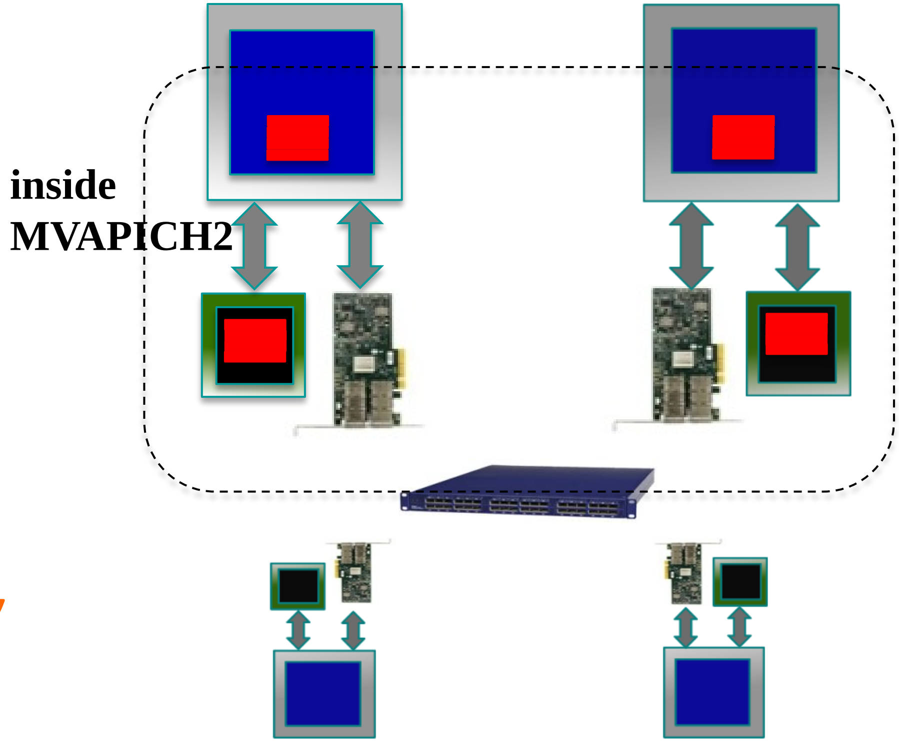
- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

```
MPI_Recv(r_devbuf, size, …);
```



inside MVAPICH2

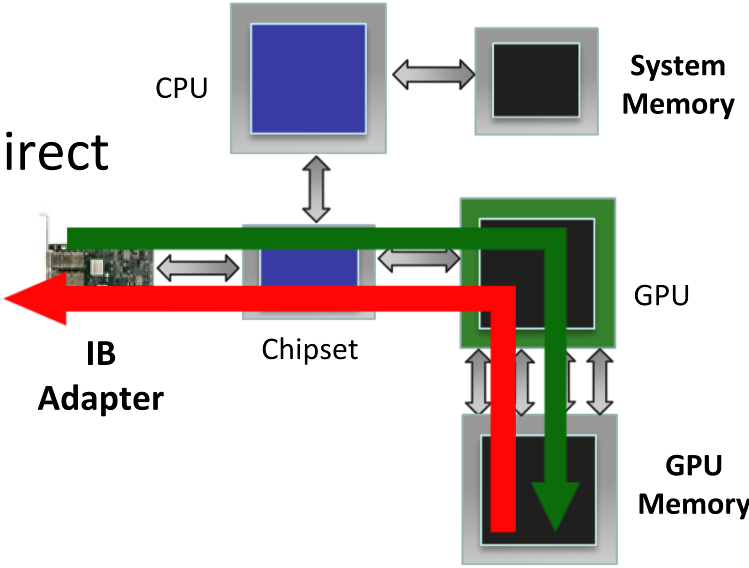
High Performance and High Productivity

GPU-Direct RDMA (GDR) with CUDA

- OFED with support for GPUDirect RDMA is developed by NVIDIA and Mellanox
- OSU has a design of MVAPICH2 using GPUDirect RDMA
  - Hybrid design using GPU-Direct RDMA
    - GPUDirect RDMA and Host-based pipelining
    - Alleviates P2P bandwidth bottlenecks on SandyBridge and IvyBridge
  - Support for communication using multi-rail
  - Support for Mellanox Connect-IB and ConnectX VPI adapters
  - Support for RoCE with Mellanox ConnectX VPI adapters

SNB E5-2670 /

IVB E5-2680V2



SNB E5-2670

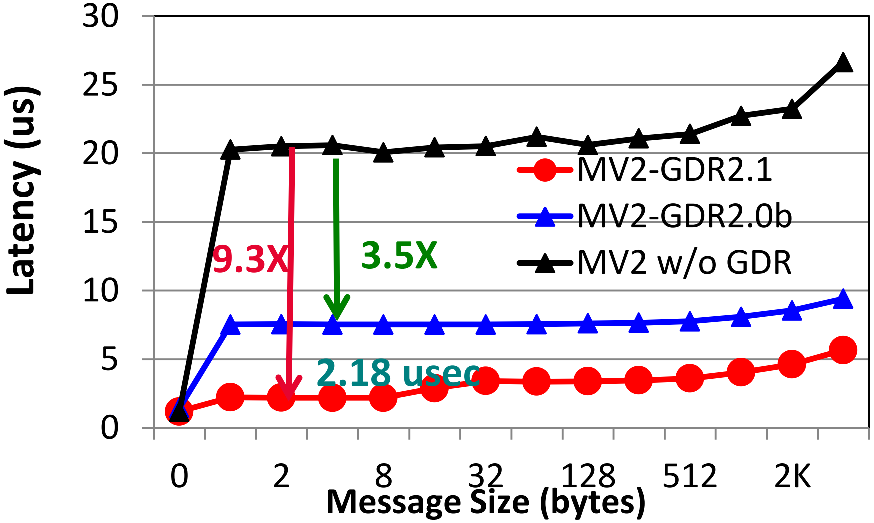
P2P write: 5.2 GB/s
P2P read: < 1.0 GB/s

IVB E5-2680V2

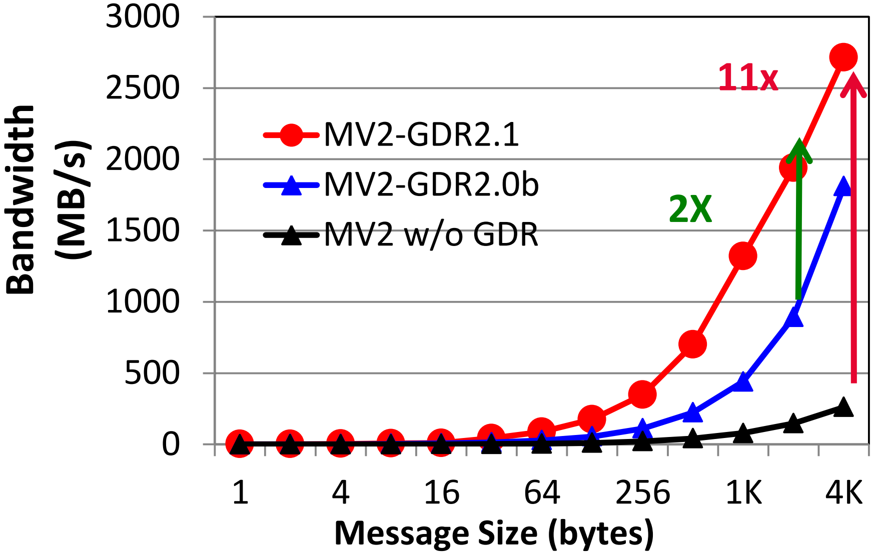
P2P write: 6.4 GB/s
P2P read: 3.5 GB/s

Performance of MVAPICH2-GDR with GPU-Direct-RDMA

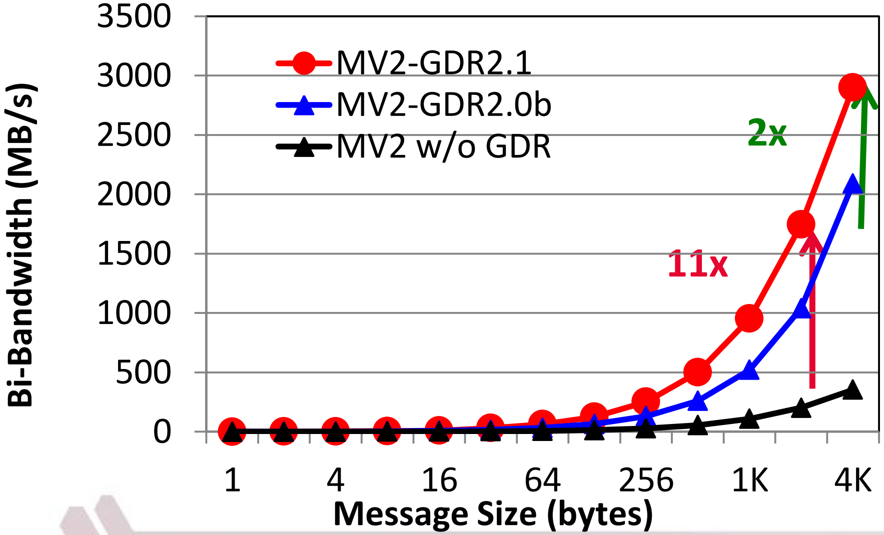
GPU-GPU Internode MPI Latency



GPU-GPU Internode MPI Uni-Bandwidth



GPU-GPU Internode Bi-directional Bandwidth



MVAPICH2-GDR-2.1

Intel Ivy Bridge (E5-2680 v2) node - 20 cores

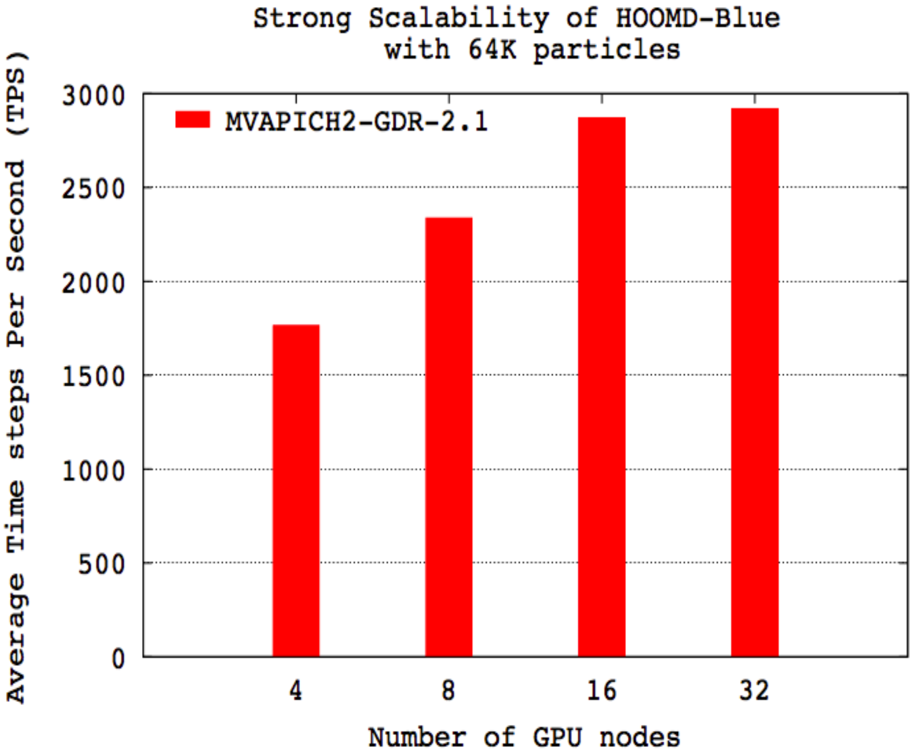
NVIDIA Tesla K40c GPU

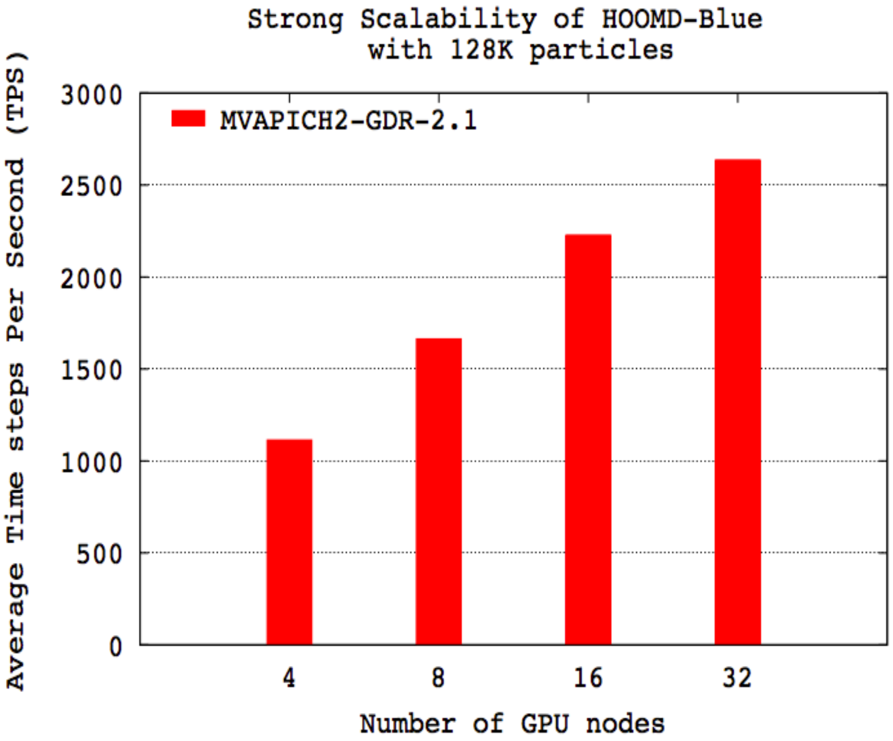
Mellanox Connect-IB Dual-FDR HCA

CUDA 7

Mellanox OFED 2.4 with GPU-Direct-RDMA

Application-Level Evaluation (HOOMD-blue)

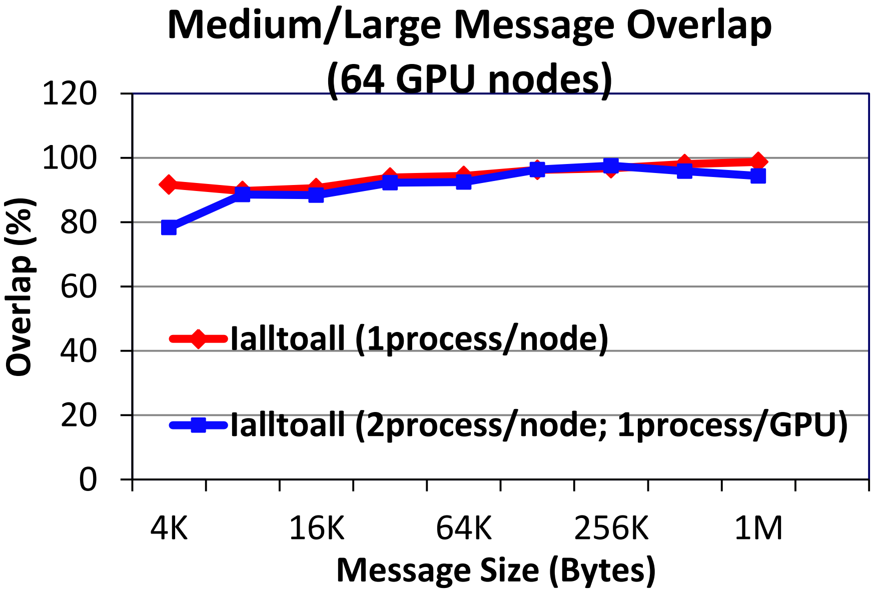


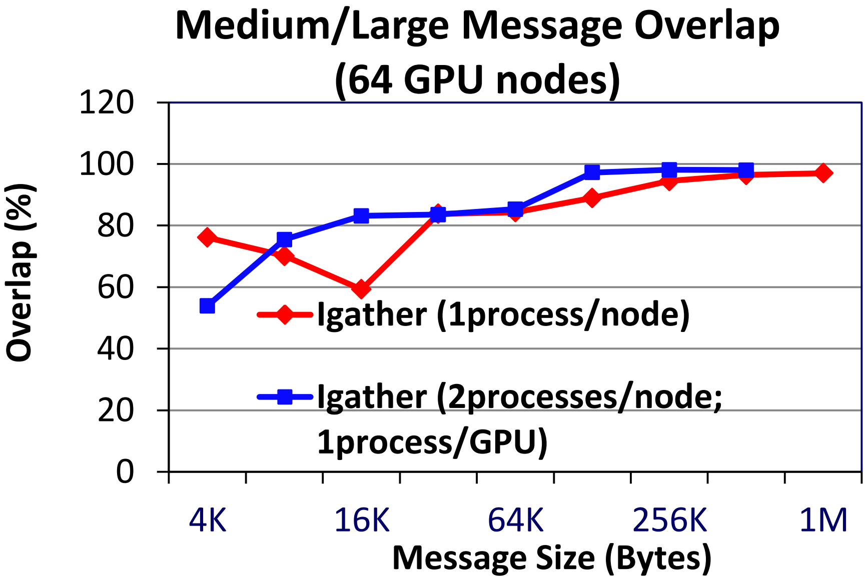


- Platform: Wilkes (Intel Ivy Bridge + NVIDIA Tesla K20c + Mellanox Connect-IB)
- HoomdBlue Version 1.0.5
  - GDRCOPY enabled: MV2_USE_CUDA=1 MV2_IBA_HCA=mlx5_0 MV2_IBA_EAGER_THRESHOLD=32768 MV2_VBUF_TOTAL_SIZE=32768 MV2_USE_GPUDIRECT_LOOPBACK_LIMIT=32768 MV2_USE_GPUDIRECT_GDRCOPY=1 MV2_USE_GPUDIRECT_GDRCOPY_LIMIT=16384

Non-Blocking Collectives (NBC) from GPU Buffers using Offload Mechanism (CORE-Direct)

- New designs are proposed to support Non-Blocking Collectives from GPU buffers to provide good overlap/latency
- Available in MVAPICH2-GDR 2.2a





Connect-X2, 2.6 GHz 12-core (IvyBridge E5-2630), dual NVIDIA K20c GPUs, Intel PCI Gen3 with MLX IB FDR switch

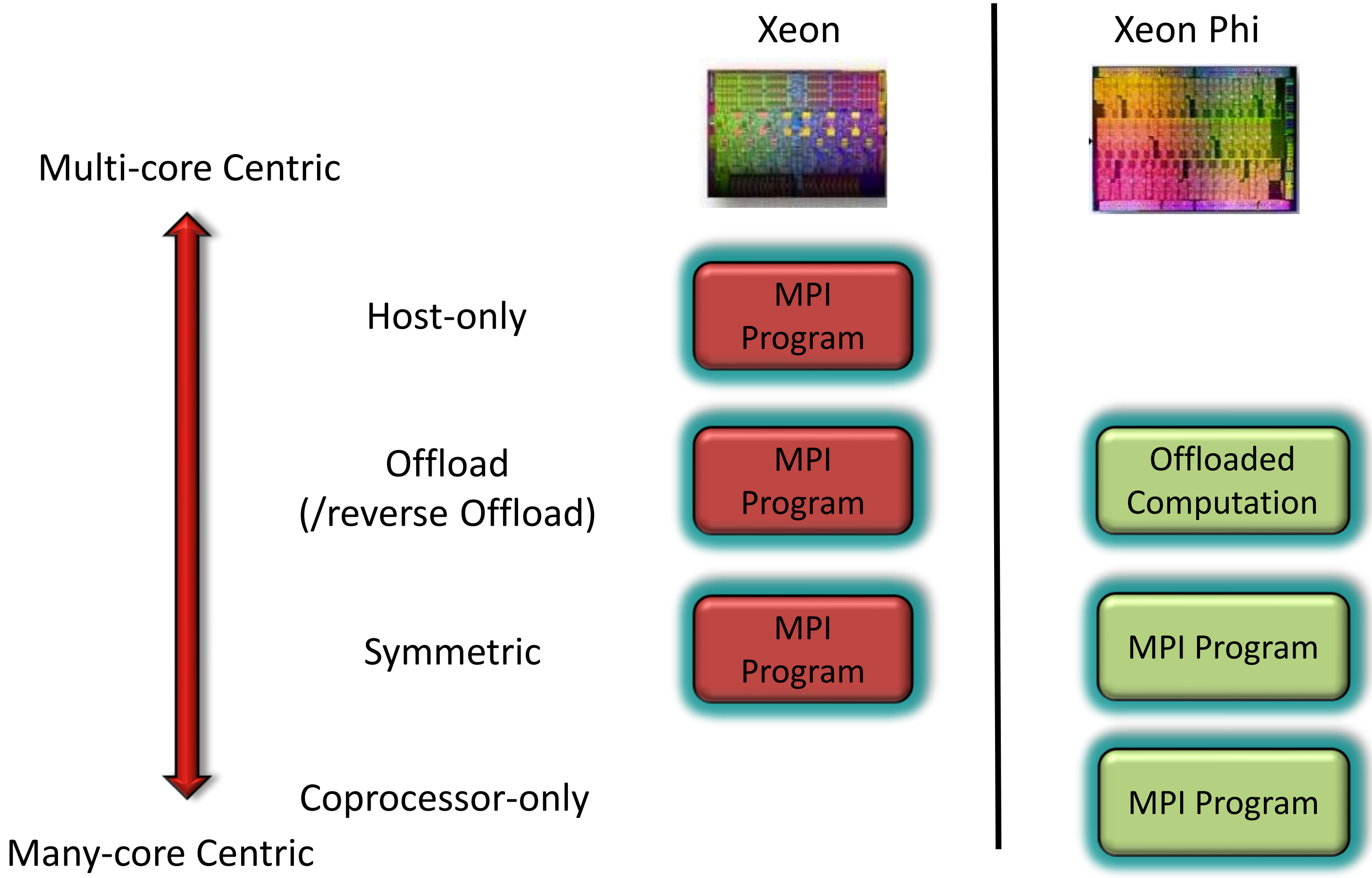
A. Venkatesh, K. Hamidouche, H. Subramoni, and D. K. Panda, “Offloaded GPU Collectives using CORE-Direct and CUDA Capabilities on IB Clusters”, HIPC, 2015

Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- Scalable Job Startup
- Scalability for million to billion processors
  - Support for highly-efficient Inter-node communication
  - Support for highly-efficient Intra-node communication
  - Support for highly-efficient One-sided / RMA communication
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Support for GPGPUs
- Support for MICs
- QoS support for communication and I/O

MPI Applications on MIC Clusters

- Flexibility in launching MPI jobs on clusters with Xeon Phi



Multi-core Centric

	Xeon	Xeon Phi
Host-only	MPI Program	
Offload (/reverse Offload)	MPI Program	Offloaded Computation
Symmetric	MPI Program	MPI Program
Coprocessor-only		MPI Program

Many-core Centric

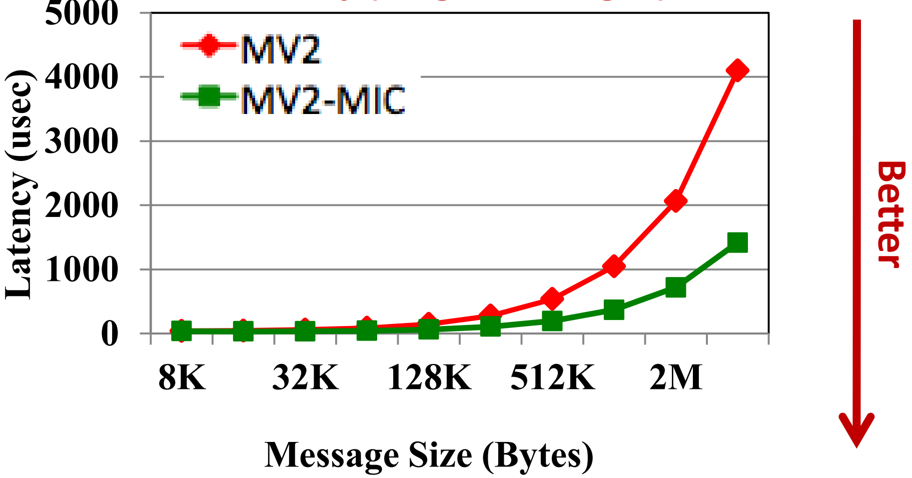
MVAPICH2-MIC 2.0: High-Performance MPI Design for Clusters with IB and MIC

- 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)

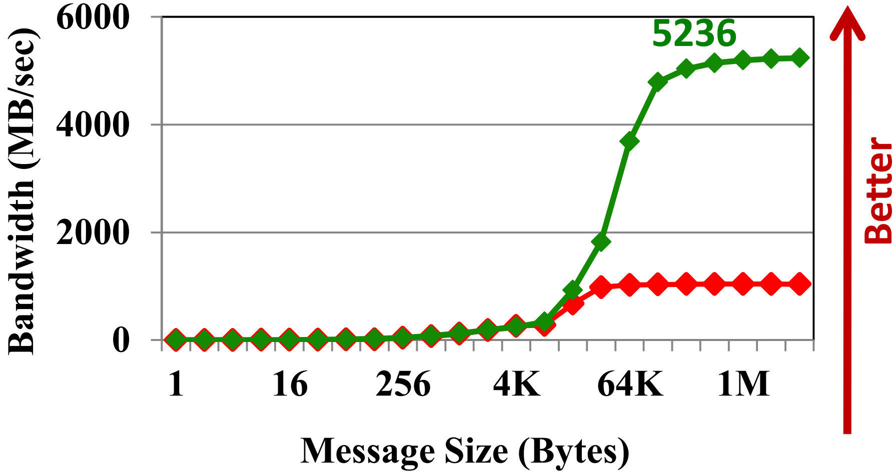
MIC-Remote-MIC P2P Communication with Proxy-based Communication

Intra-socket P2P

Latency (Large Messages)

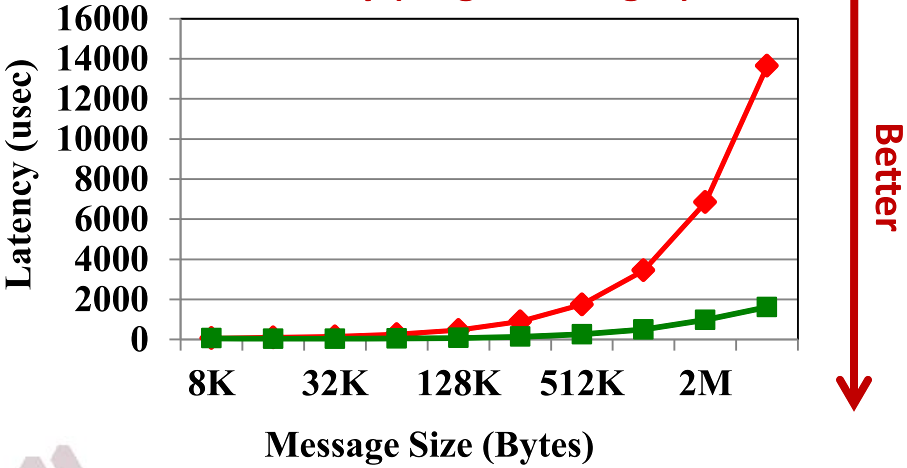


Bandwidth

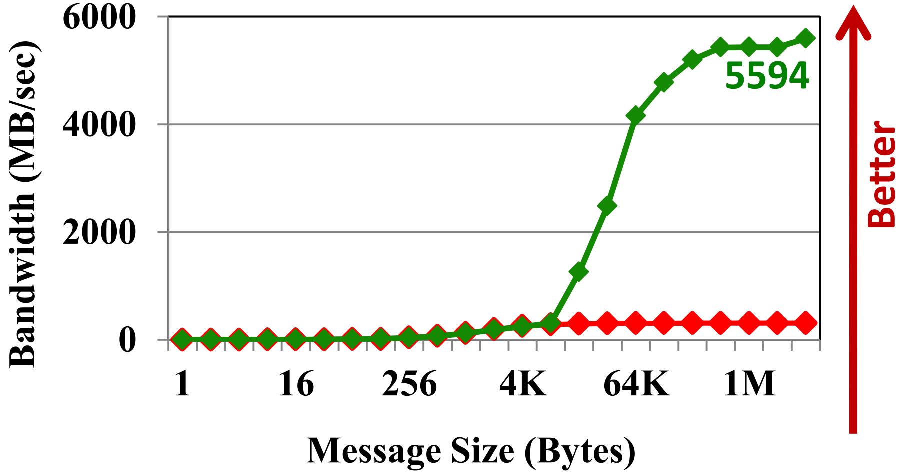


Inter-socket P2P

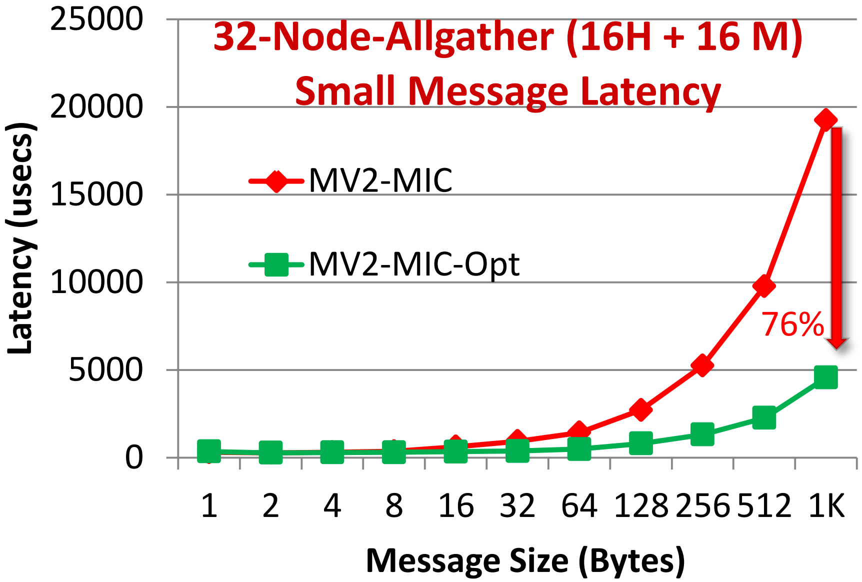
Latency (Large Messages)

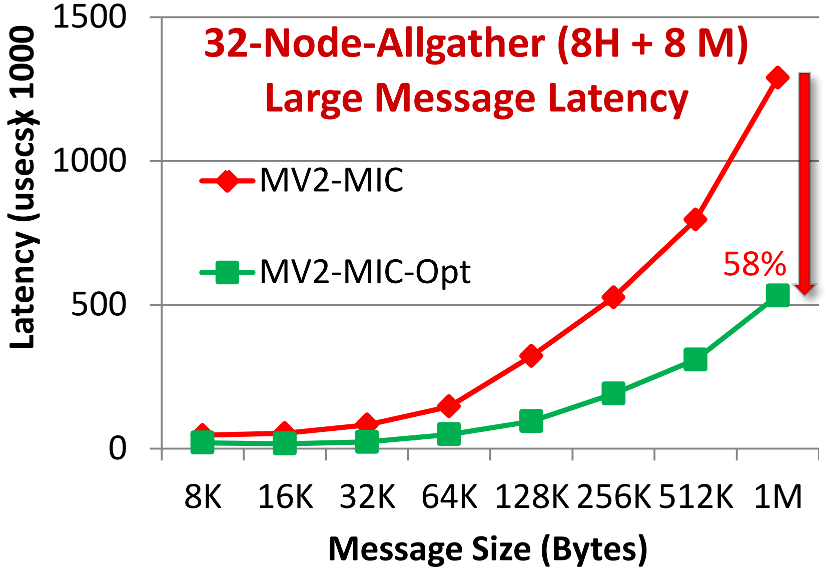


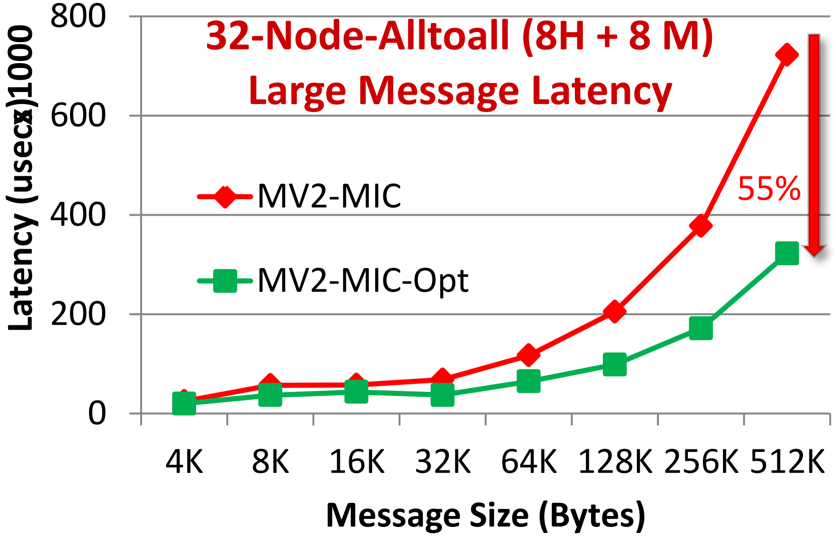
Bandwidth

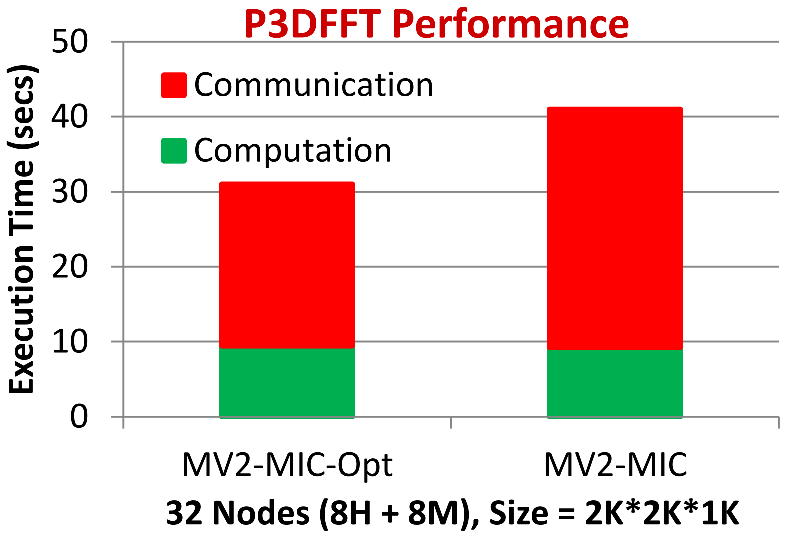


Optimized MPI Collectives for MIC Clusters (Allgather & Alltoall)


32-Node-Allgather (16H + 16 M)
Small Message Latency


32-Node-Allgather (8H + 8 M)
Large Message Latency


32-Node-Alltoall (8H + 8 M)
Large Message Latency

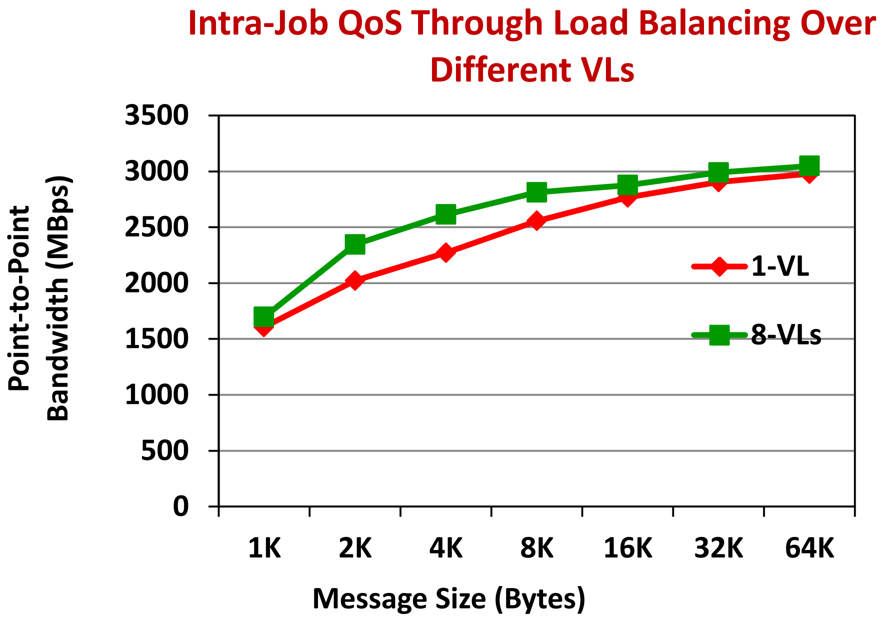

P3DFFT Performance

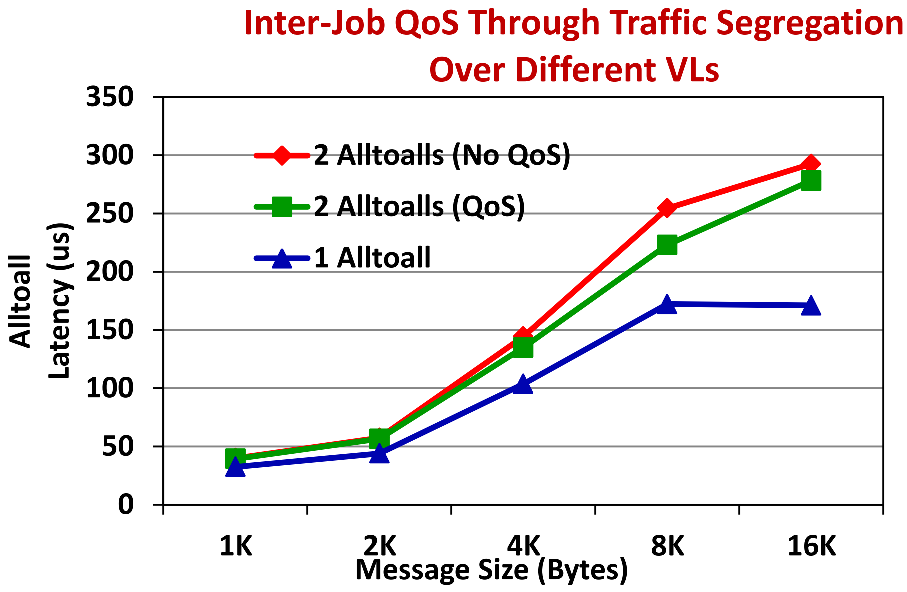
A. Venkatesh, S. Potluri, R. Rajachandrasekar, M. Luo, K. Hamidouche and D. K. Panda - High Performance Alltoall and Allgather designs for InfiniBand MIC Clusters; IPDPS'14, May 2014

Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- Scalable Job Startup
- Scalability for million to billion processors
  - Support for highly-efficient Inter-node communication
  - Support for highly-efficient Intra-node communication
  - Support for highly-efficient One-sided / RMA communication
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Support for GPGPUs
- Support for MICs
- QoS support for communication and I/O

Exploiting QoS Support in MVAPICH2

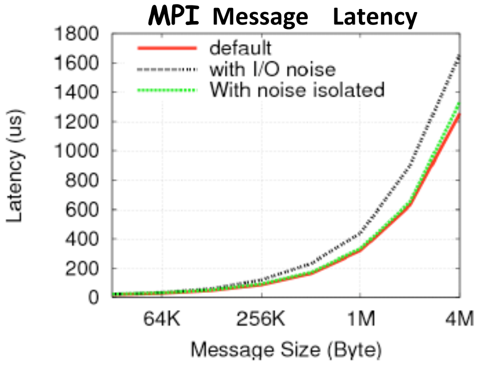


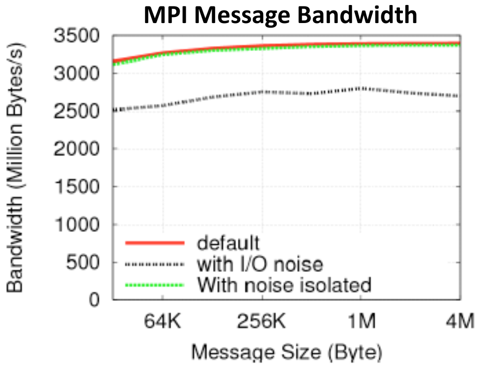


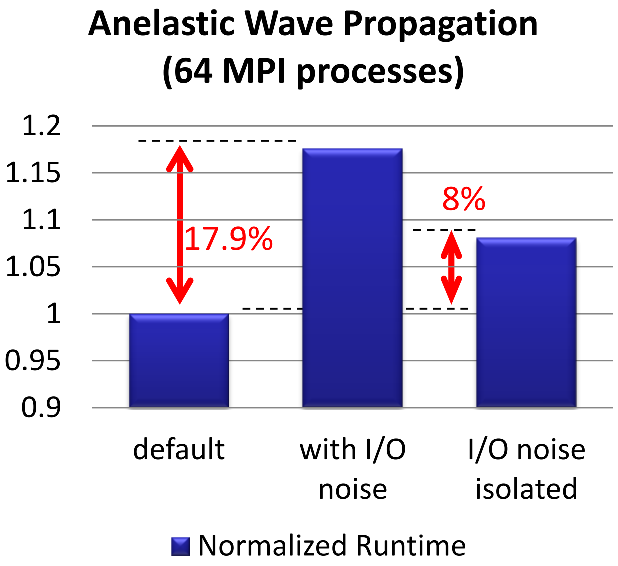
- IB is capable of providing network level differentiated service – QoS
- Uses Service Levels (SL) and Virtual Lanes (VL) to classify traffic
- Enabled at configure time using CFLAG ENABLE_QOS_SUPPORT
- Check with System administrator before enabling
  - Can affect performance of other jobs in system

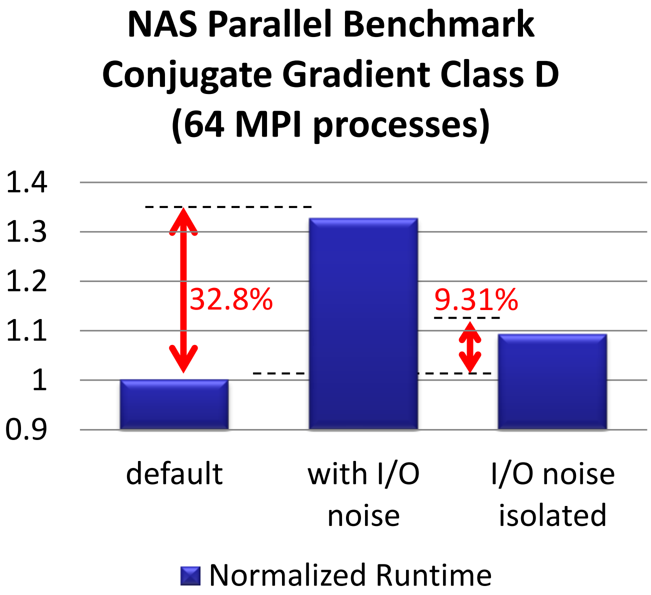
Minimizing Network Contention w/ QoS-Aware Data-Staging

- Asynchronous I/O introduces contention for network-resources
- How should data be orchestrated in a data-staging architecture to eliminate such contention?
- Can the QoS capabilities provided by cutting-edge interconnect technologies be leveraged by parallel filesystems to minimize network contention?


MPI Message Latency


MPI Message Bandwidth


Anelastic Wave Propagation (64 MPI processes)


NAS Parallel Benchmark Conjugate Gradient Class D (64 MPI processes)

- **Reduces runtime overhead from 17.9% to 8% and from 32.8% to 9.31%, in case of AWP and NAS-CG applications respectively**

R. Rajachandrasekar, J. Jaswani, H. Subramoni and D. K. Panda, Minimizing Network Contention in InfiniBand Clusters with a QoS-Aware Data-Staging Framework, IEEE Cluster, Sept. 2012

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
- Energy Awareness

Two Additional Talks

- Wednesday (1:30-2:00pm)
  - The MVAPICH2 Project: Heading Towards New Horizons in Energy-Awareness, Virtualization and Network/Job-Level Introspection
- Thursday (10:00-10:30am)
  - How to Exploit MPI, PGAS and Hybrid MPI+PGAS Programming through MVAPICH2-X?

Funding Acknowledgments

Funding Support by

Office of Science
U.S. DEPARTMENT OF ENERGY

NSF

Mellanox
TECHNOLOGIES

CISCO SYSTEMS

QLOGIC®

CRAY
THE SUPERCOMPUTER COMPANY

intel®

LINUX
NETWORX

NVIDIA.

NetApp™

Sun
microsystems

Equipment Support by

Mellanox
TECHNOLOGIES

intel®

AMD

Sun
microsystems

APPRO

advanced clustering
technologies, inc.

Microway
Technology you can count on™

NVIDIA.

QLOGIC®

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

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

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

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

MVAPICH Web Page

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

MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

panda@cse.ohio-state.edu

subramon@cse.ohio-state.edu