MVAPICH/MVAPICH2: Latest Status and Future Plans

Presentation at OFA Conference, Monterey 2012

by

Dhabaleswar K. (DK) Panda

The Ohio State University

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

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

Presentation Outline

- Overview
- MVAPICH and MVAPICH2 Features
- Sample Performance Numbers
- Future Plans and Sample Solutions

MVAPICH/MVAPICH2 Software

- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2.2), Available since 2002
  - Used by more than 1,850 organizations (HPC Centers, Industry and Universities) in 65 countries
  - More than 102,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 5th ranked 73,278-core cluster (Tsubame 2.0) at Tokyo Institute of Technology
    - 7th ranked 111,104-core cluster (Pleiades) at NASA
    - 25th ranked 62,976-core cluster (Ranger) at TACC
    - and many others
  - Available with software stacks of many InfiniBand, High-speed Ethernet and server vendors including Open Fabrics Enterprise Distribution (OFED) and Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- Partner in the upcoming U.S. NSF-TACC Stampede (10-15 PFlop) System

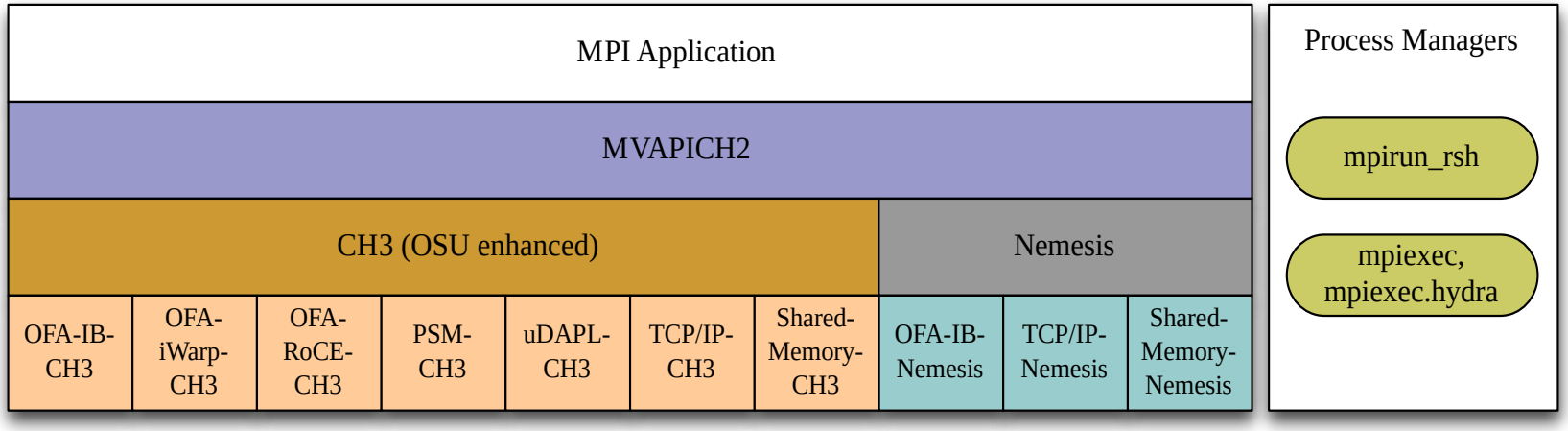
MVAPICH2 1.7 – Major Features

- Released on 10/14/11
- Major features (Compared to MVAPICH2 1.6)
  - Hybrid UD-RC/XRC support to get best performance on large-scale systems with reduced/constant memory footprint
  - HugePage support
  - Optimized Fence synchronization
  - Shared memory backed Windows for One-Sided Communication
  - Improved intra-node shared-memory communication performance
  - Minimizing number of connections and memory footprint
  - Enhancement/Optimization of algorithms and tuning for collectives (Barrier, Bcast, Gather, Allgather, Reduce, Allreduce, Alltoall and Allgatherv)
  - Fast process migration using RDMA
  - Supporting large data transfer (>2GB)
  - Integrated with Enhanced LiMIC2 (v0.5.5) to support Intra-node large message (>2GB) transfers
  - Improved connection management
  - Support for Chelsio T4 adapter
  - Improved pt-to-pt communication and Multi-core-aware collective support (QLogic PSM interface)
  - Based on MPICH2 1.4.1p1
  - Hwloc v1.2.2

Latest MVAPICH2 1.8RC1 – Features

- Released on 03/22/12
- Major features (Compared to MVAPICH2 1.7)
  - Support for MPI communication from NVIDIA GPU device memory (Details will be presented on Wednesday)
    - High performance RDMA-based inter-node point-to-point communication (GPU-GPU, GPU-Host and Host-GPU)
    - High performance intra-node point-to-point communication for multi-GPU adapters/node (GPU-GPU, GPU-Host and Host-GPU)
    - Taking advantage of CUDA IPC (available in CUDA 4.1) in intra-node communication for multiple GPU adapters/node
    - Optimized and tuned collectives for GPU device buffers
    - MPI datatype support for point-to-point and collective communication from GPU device buffers
  - New shared memory design for enhanced intra-node small message performance
  - Support suspend/resume functionality with mpirun_rsh
  - Enhanced support for CPU binding with socket and numanode level granularity
  - Checkpoint-Restart and run-through stabilization support in OFA-IB-Nemesis interface
  - Enhancing OFA-IB-Nemesis interface to handle IB errors gracefully
  - Enhanced integration with SLURM and PBS
  - Reduced memory footprint of the library
  - Enhanced one-sided communication design with reduced memory requirement
  - Enhancements and tuned collectives (Bcast and Alltoallv)
  - Support iWARP interoperability between Intel NE020 and Chelsio T4 Adapters
  - RoCE enable environment variable name is changed from MV2_USE_RDMAOE to MV2_USE_RoCE
  - Update to hwloc v1.4

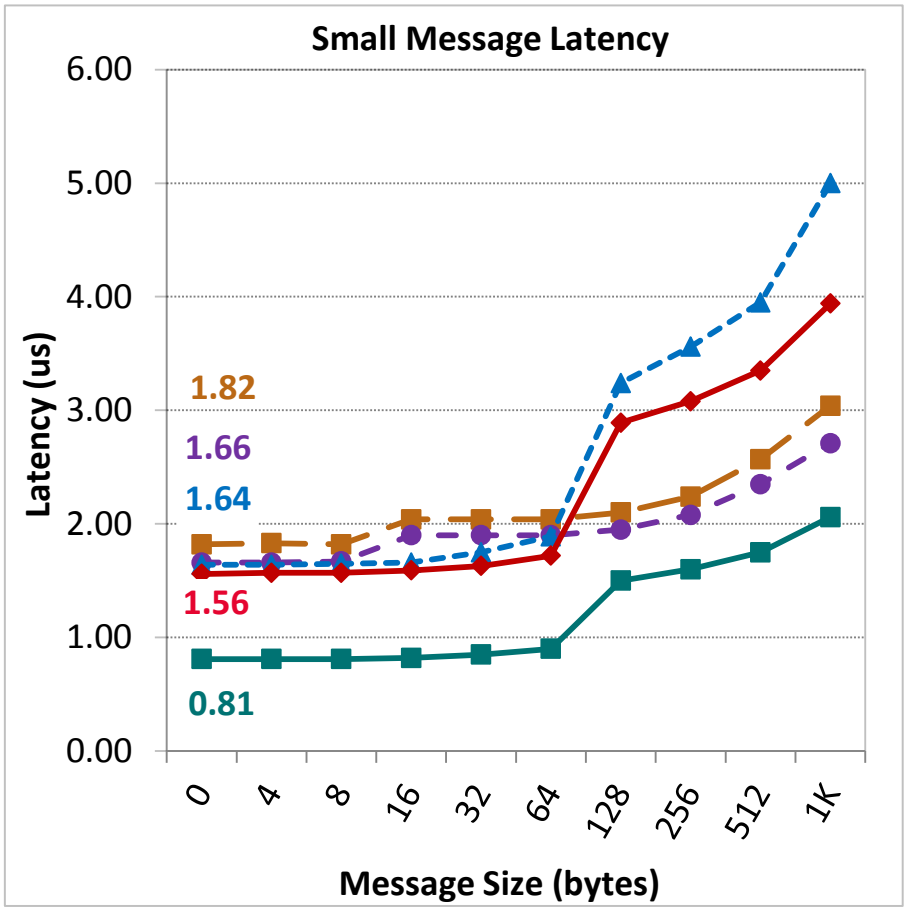
MVAPICH2 Architecture (Latest Release 1.8RC1)

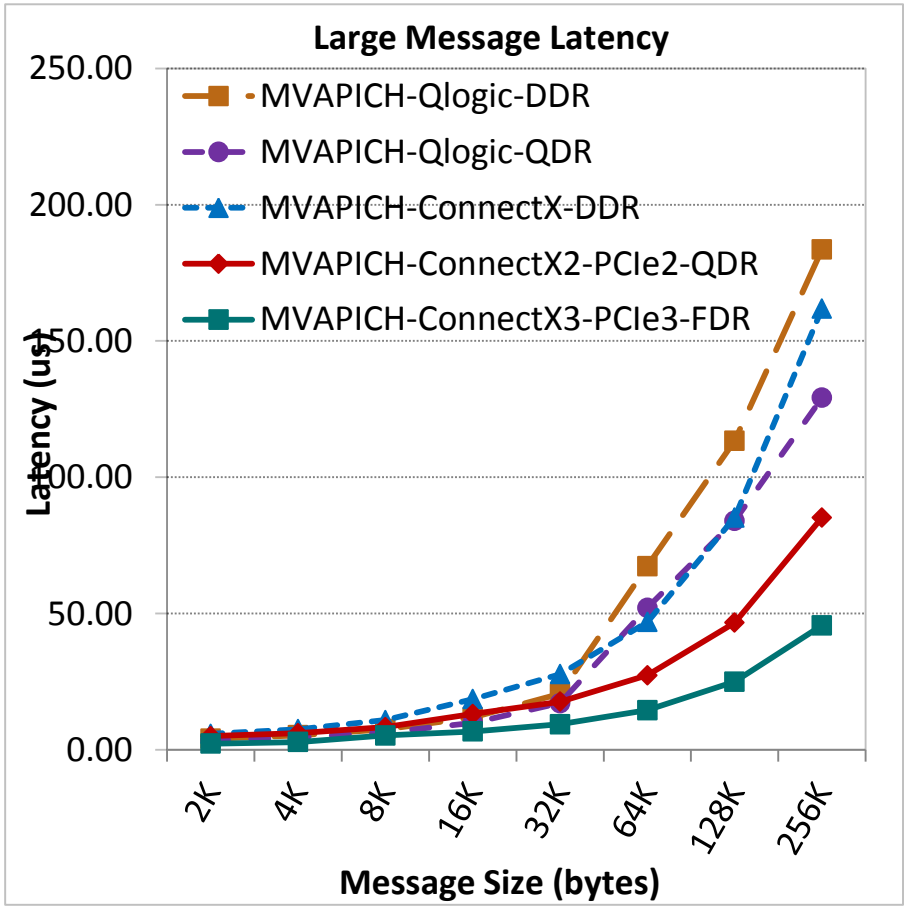


All Different PCI interfaces

Major Computing Platforms: IA-32, EM64T, Nehalem, Westmere, Sandybridge, Opteron, Magny, ..

One-way Latency: MPI over IB

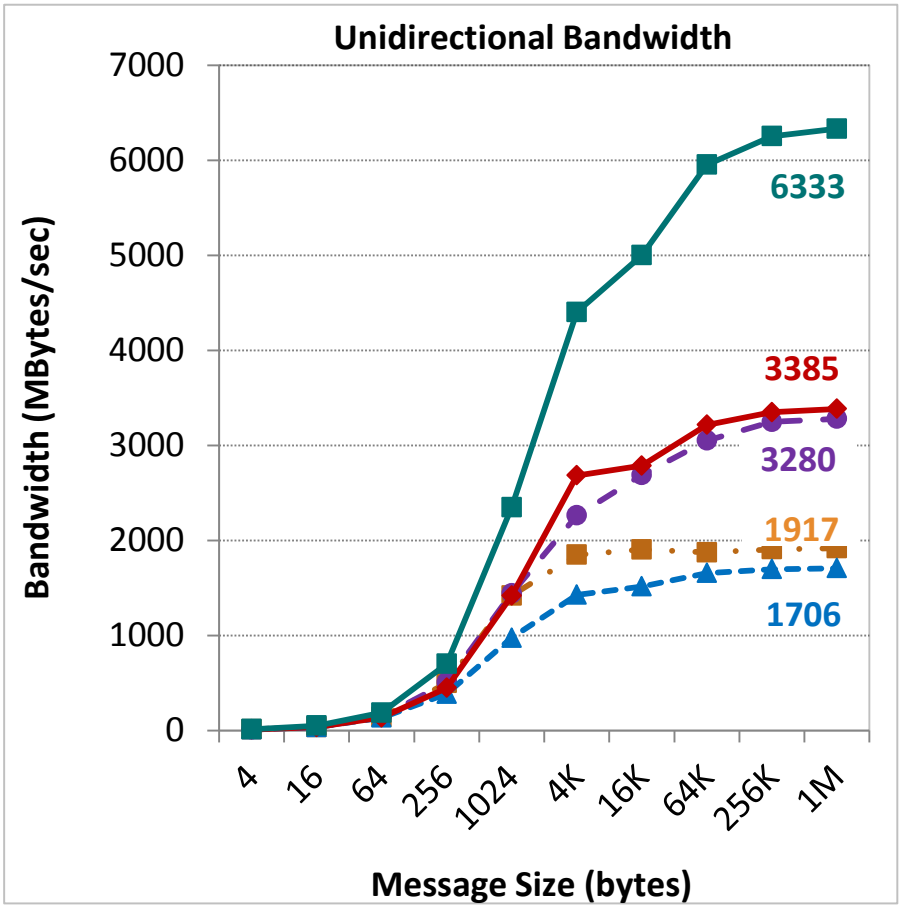


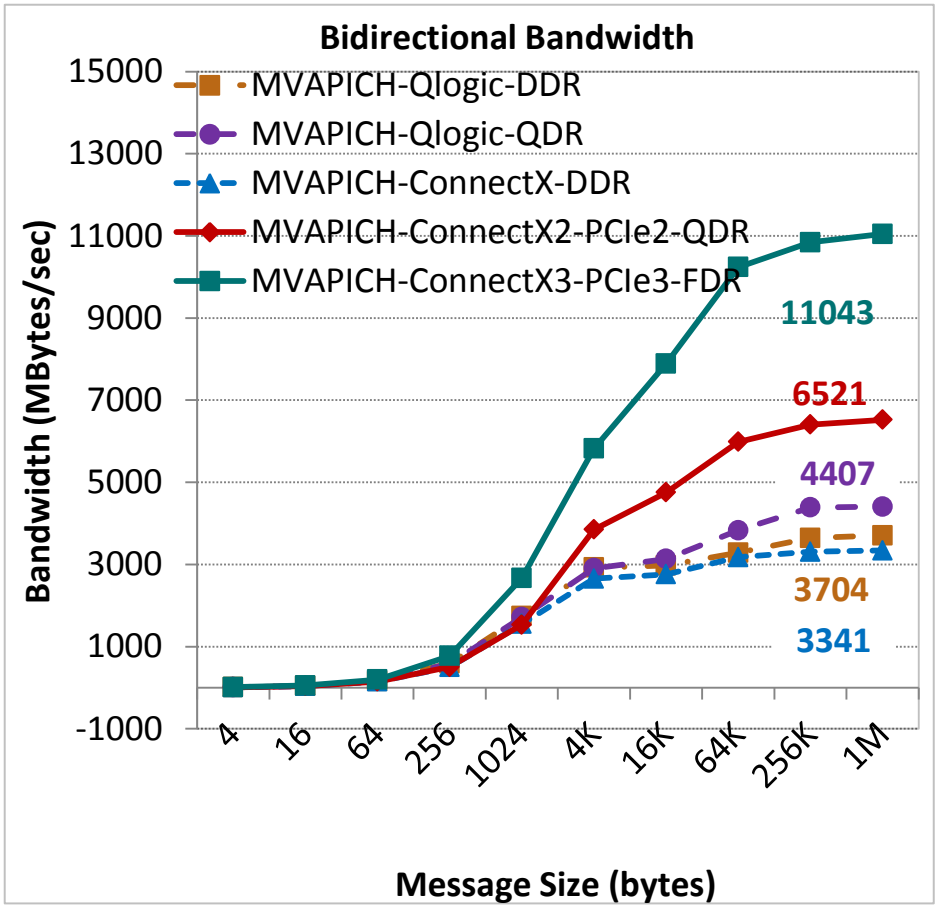


DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

FDR - 2.6 GHz Octa-core (Sandybridge) Intel PCI Gen3 without IB switch

Bandwidth: MPI over IB

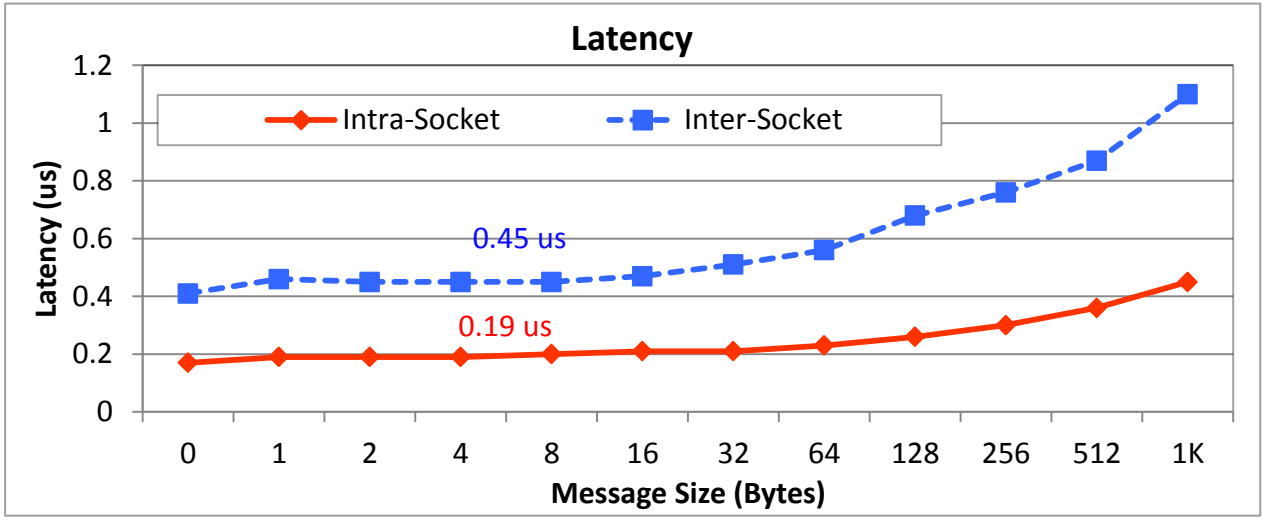




DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

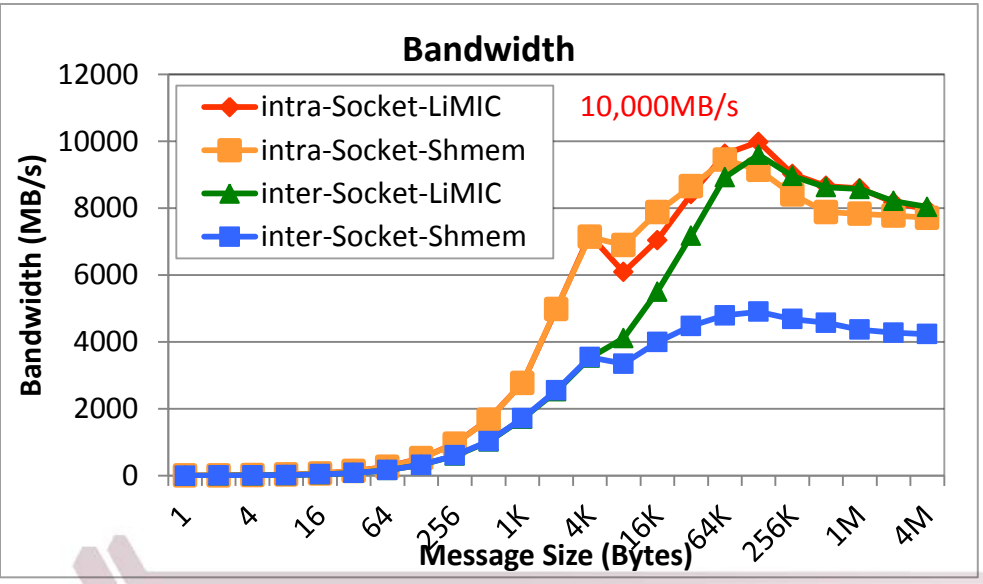
FDR - 2.6 GHz Octa-core (Sandybridge) Intel PCI Gen3 without IB switch

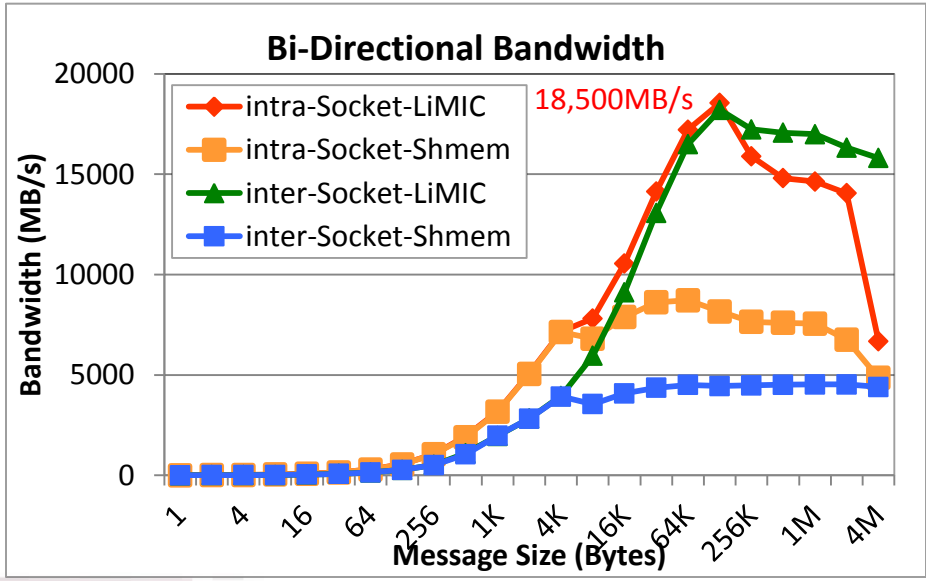
MVAPICH2 Two-Sided Intra-Node Performance (Shared memory and Kernel-based Zero-copy Support)



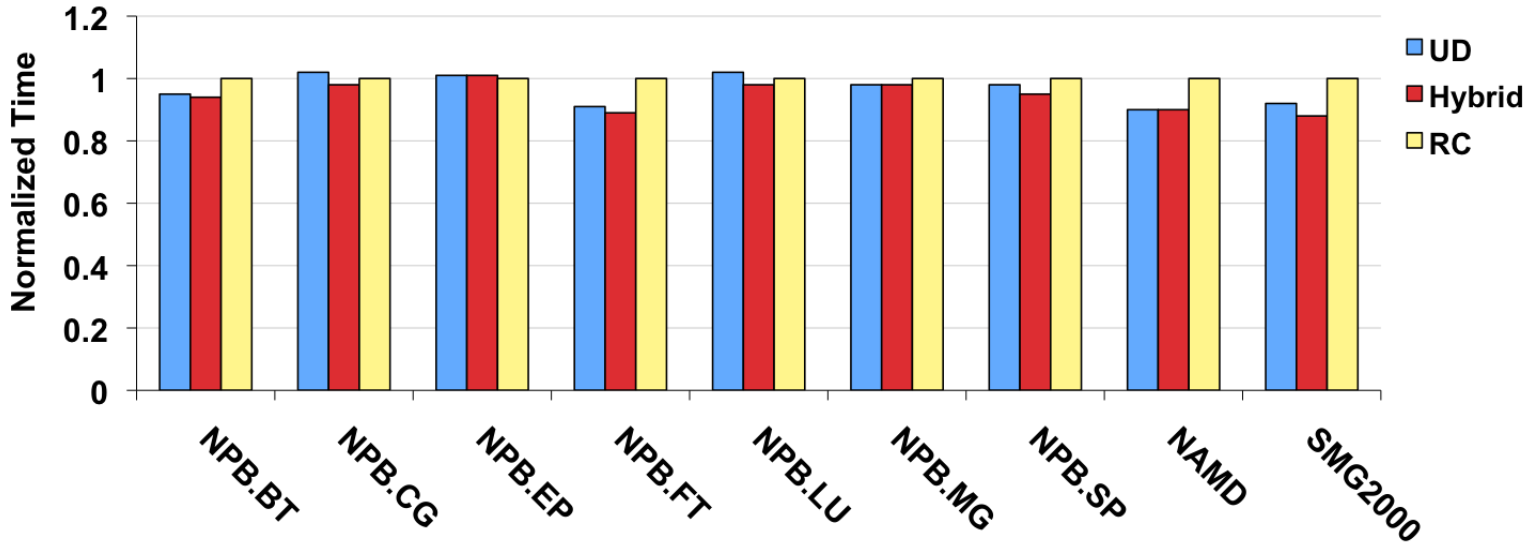
Latest MVAPICH2 1.8RC1

Intel Westmere





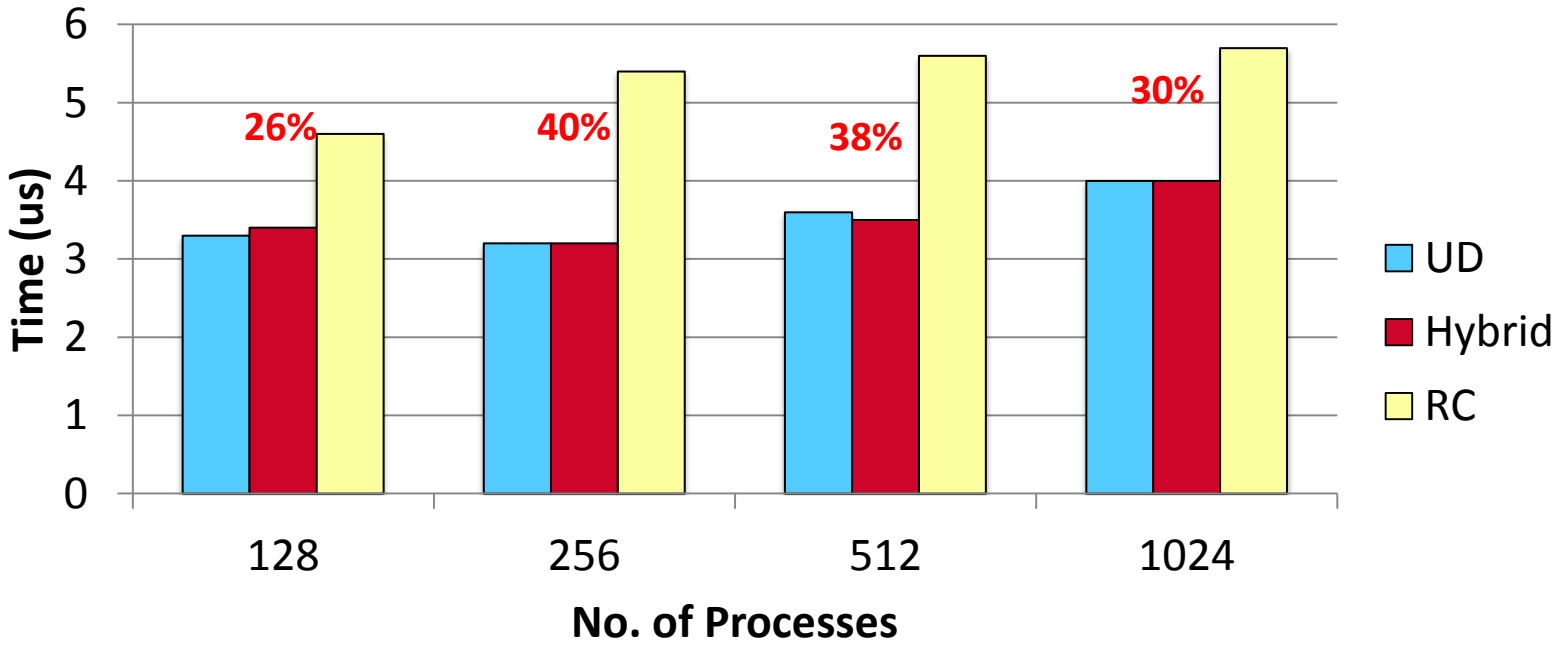
Hybrid Transport Design (UD-RC/XRC)



- Supports all different modes: RC/XRC, UD, UD+RC or UD+XRC
- Available in MVAPICH (since 1.1)
- Available in MVAPICH2 (since 1.7) with more advanced, tuned and adaptive designs
- Provides flexibility to tune performance of critical applications on large-scale clusters with various modes

M. Koop, T. Jones and D. K. Panda, MVAPICH-Aptus: Scalable High-Performance Multi-Transport MPI over InfiniBand, IPDPS '08

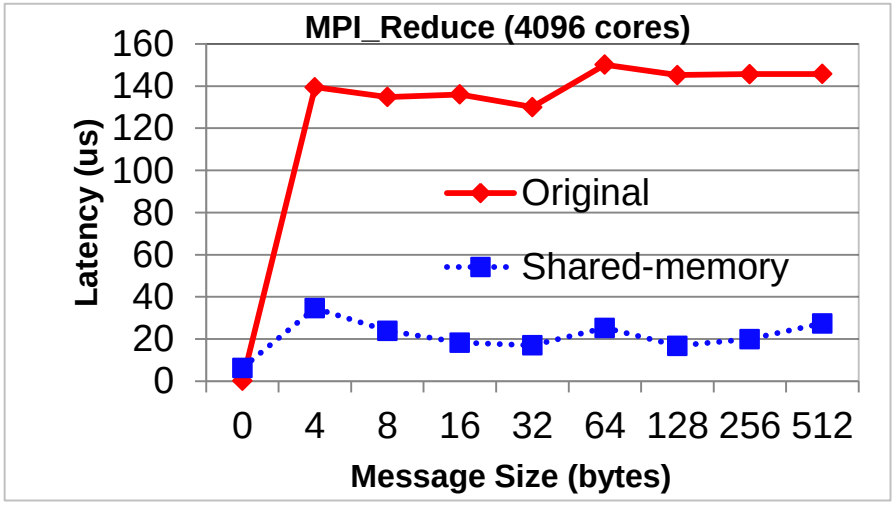
HPCC Random Ring Latency with MVAPICH2 (UD/RC/Hybrid)



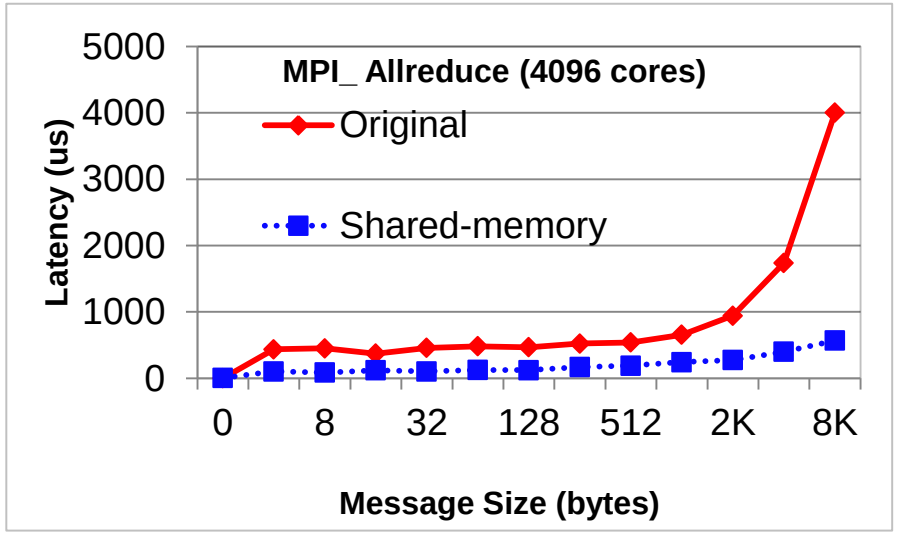
- Experimental Platform : LLNL Hyperion Cluster (Intel Harpertown + QDR IB)
- RC: Default Parameters
- UD: MV2_HYBRID_ENABLE_THRESHOLD=1,
  MV2_HYBRID_MAX_RC_CONNECTIONS=0
- Hybrid: MV2_HYBRID_ENABLE_THRESHOLD=1

Shared-memory Aware Collectives

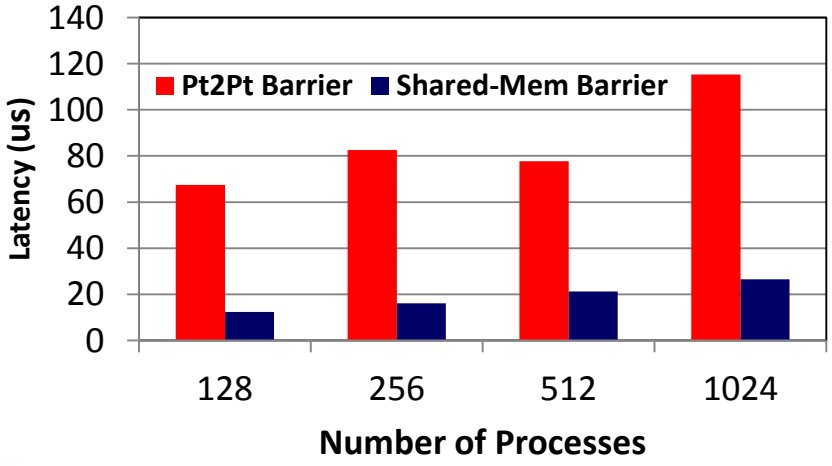
- MVAPICH2 Reduce/Allreduce with 4K cores on TACC Ranger (AMD Barcelona, SDR IB)



MV2_USE_SHMEM_REDUCE=0/1



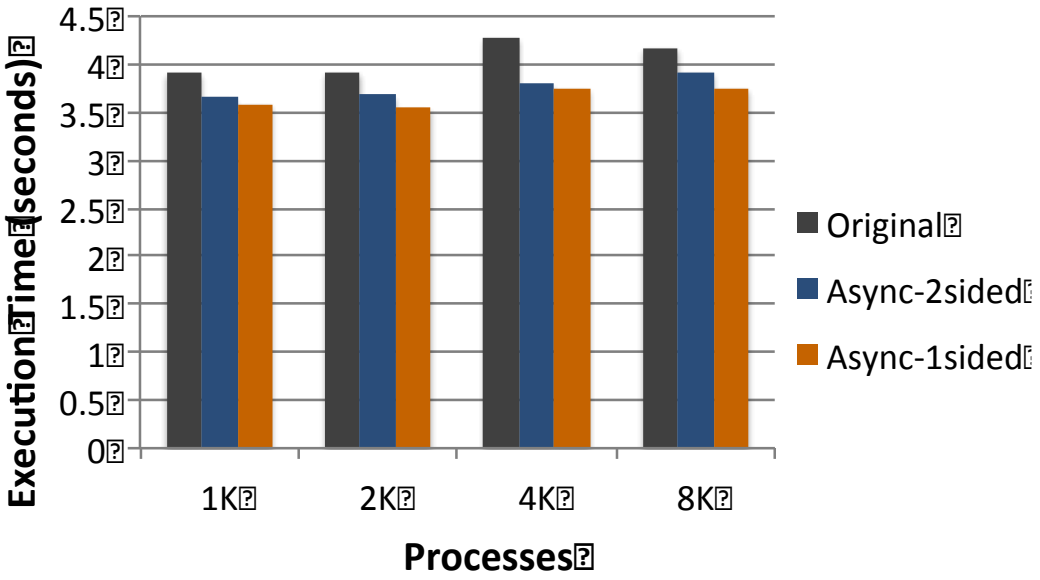
MV2_USE_SHMEM_ALLREDUCE=0/1

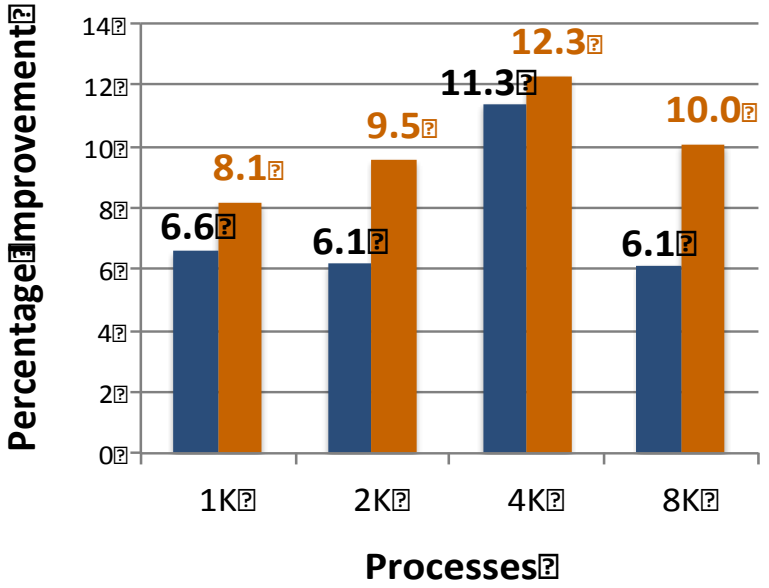


- MVAPICH2 Barrier with 1K Intel Westmere cores , QDR IB

MV2_USE_SHMEM_BARRIER=0/1

One-Sided Design: Delivering Overlap and Performance Benefits for AWP-ODC

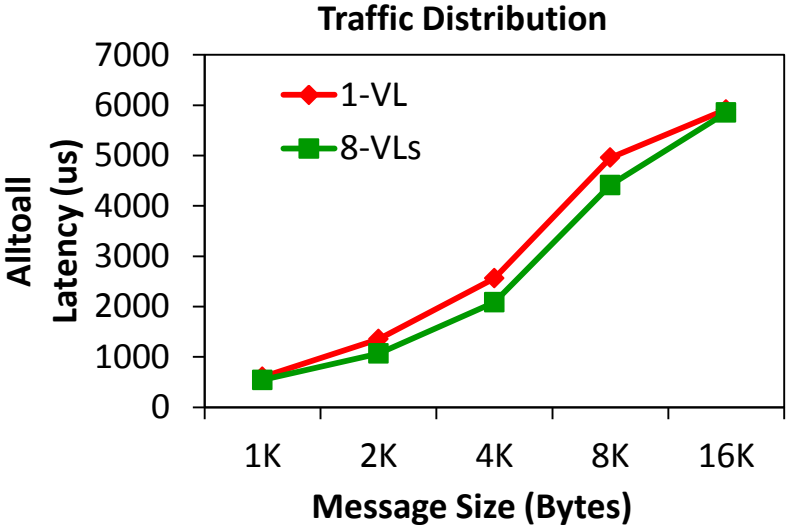


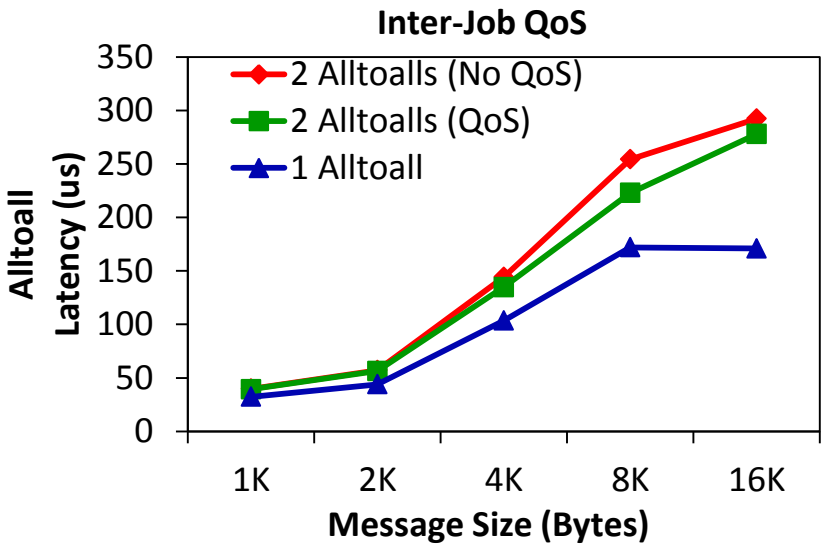


- AWP-ODC, an earthquake simulation application - experiments on TACC Ranger cluster – 64x64x64 data grid per process – 25 iterations – 32KB messages
- On 4K processes – 11% with Async-2sided and 12% with Async-1sided (RMA)
- On 8K processes – 6% with Async-2sided and 10% with Async-1sided (RMA)

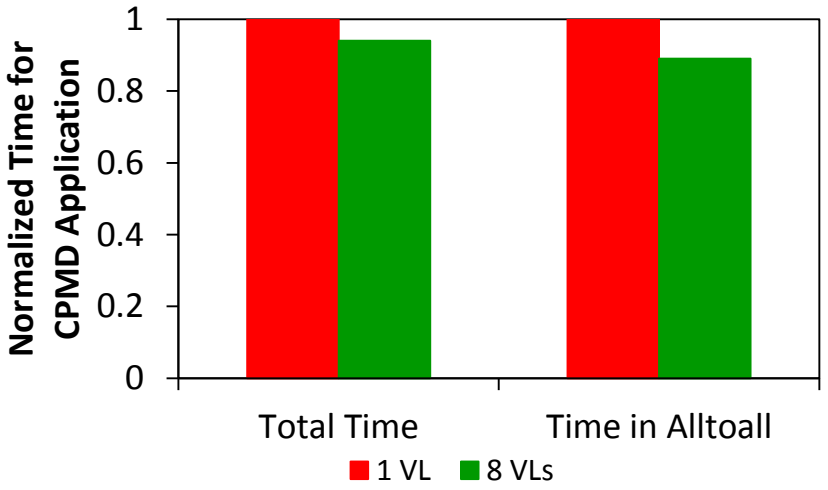
Joint work with OSU, SDSC and TACC
Gordon-Bell Finalist Paper for Supercomputing 2010

Multiple VLs, Inter-Job QoS and Support for 3D Torus





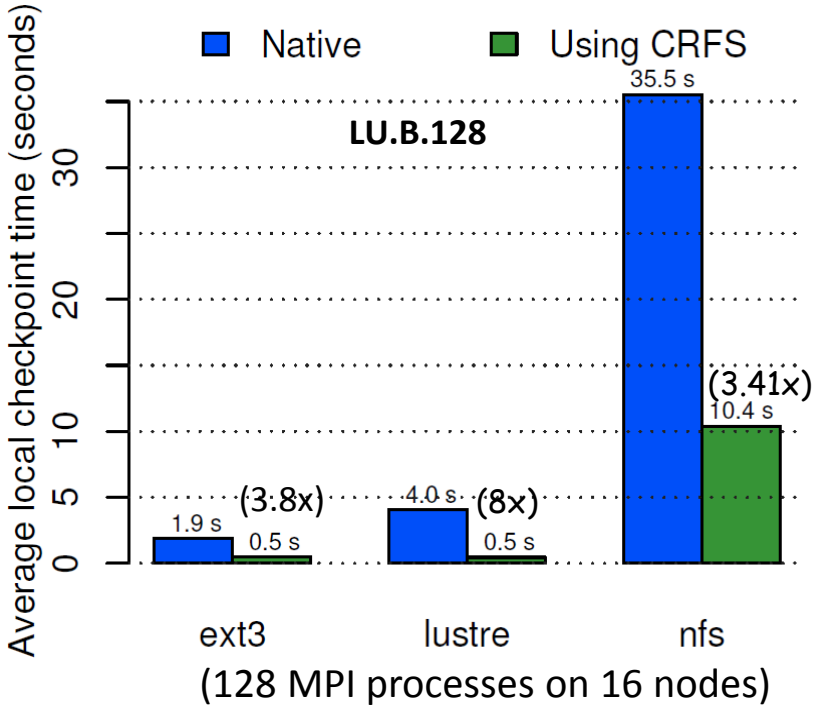
- Micro-benchmarks use 8 communicating pairs
  - One QP per pair
- Performance improvement over One VL case
  - Alltoall – 20 %
  - Application – 11%
- 12% performance improvement with Inter-Job QoS
- Multiple VLs used to break deadlocks in IB Torus networks
  - MPI queries OpenSM to retrieve correct SL
  - Enabled at configure time (--enable-3dtorus-support)



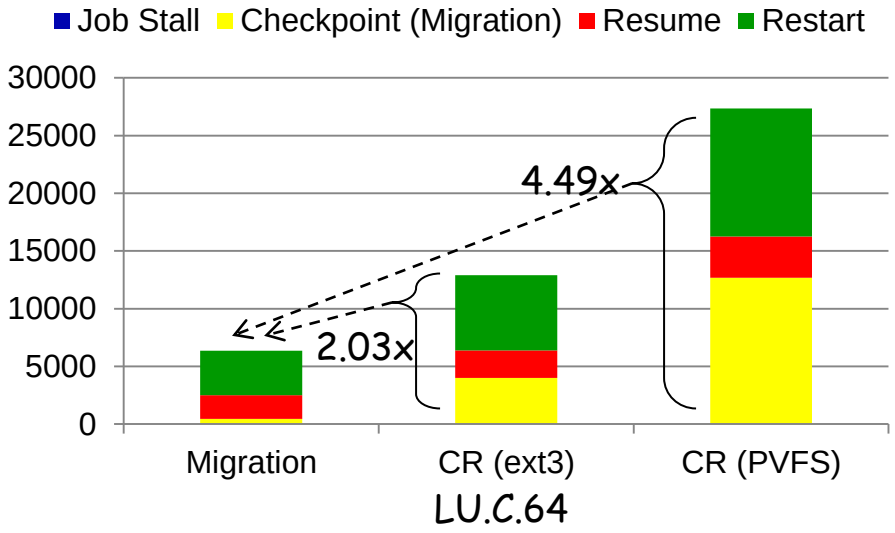
H. Subramoni, P. Lai, S. Sur and D. K. Panda, Improving Application Performance and Predictability using Multiple Virtual Lanes in Modern Multi-Core InfiniBand Clusters , Int'l Conference on Parallel Processing (ICPP '10), Sept. 2010.

Fault-Tolerance Support in MVAPICH2

- **Checkpoint/ Restart with Aggregation**
  - Checkpoint writes aggregated to reduce filesystem contention



- **RDMA-Based Process Migration**
  - Fault-predication supported proactive failure avoidance mechanism.
  - Also allows MPI processes to be relocated on large-scale systems for topology-aware compaction



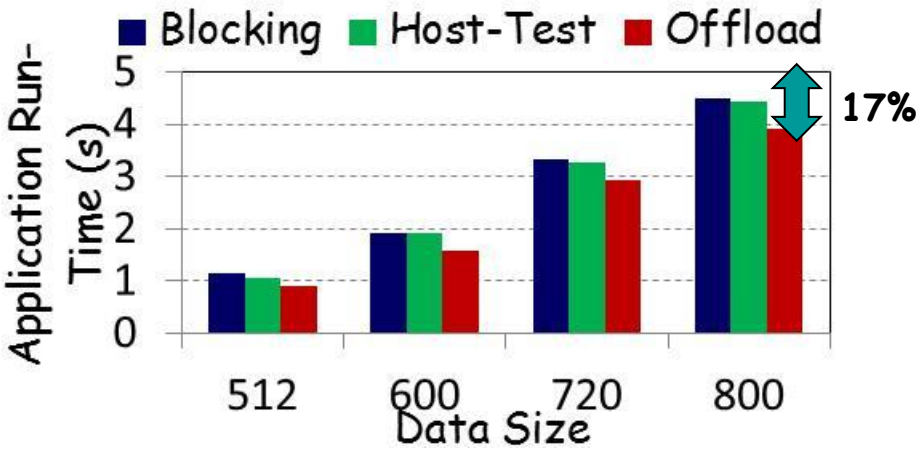
X. Ouyang, R. Rajachandrasekar, X. Besseron, H. Wang, J. Huang and D. K. Panda, CRFS: A Lightweight User-Level Filesystem for Generic Checkpoint/Restart, Int'l Conference on Parallel Processing (ICPP '11)

X. Ouyang, S. Marcarelli, R. Rajachandrasekar and D. K. Panda, RDMA-Based Job Migration Framework for MPI over InfiniBand, Int'l Conference on Cluster Computing (Cluster '10), Sept. 2010.

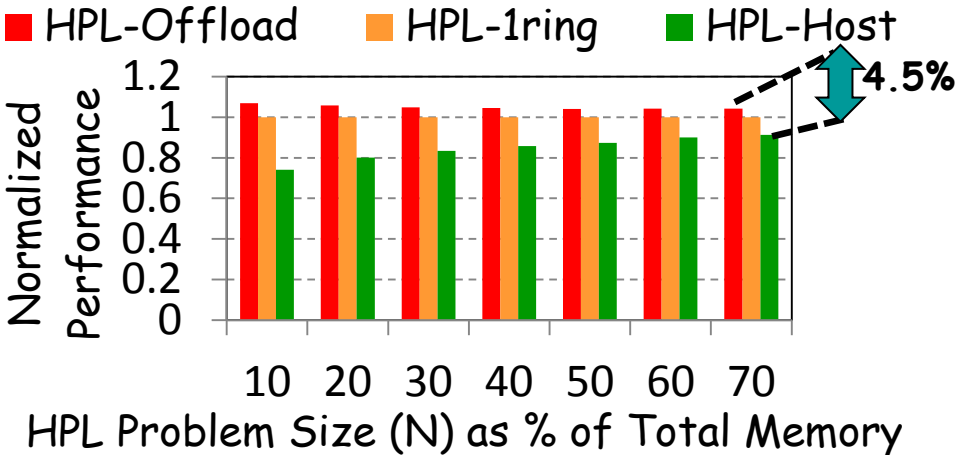
MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 500K-1M cores
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - To be presented in the following talk
- Enhanced Optimization for GPU Support and Accelerators
  - Extending the GPGPU support (To be presented on Wednesday morning)
  - Support for Intel MIC (A paper will be presented at Intel-TACC Symposium in April '12)
- Taking advantage of Collective Offload framework
  - Including support for non-blocking collectives (MPI 3.0)
- Extended topology-aware collectives
- Power-aware collectives
- Enhanced Multi-rail Designs
- Automatic optimization of collectives
  - LiMIC2, XRC, Hybrid (UD-RC/XRC) and Multi-rail
- Support for MPI Tools Interface
- Checkpoint-Restart and migration support with incremental checkpointing
- Fault-tolerance with run-through stabilization (being discussed in MPI 3.0)
- QoS-aware I/O and checkpointing
- Automatic tuning and self-adaptation for different systems and applications

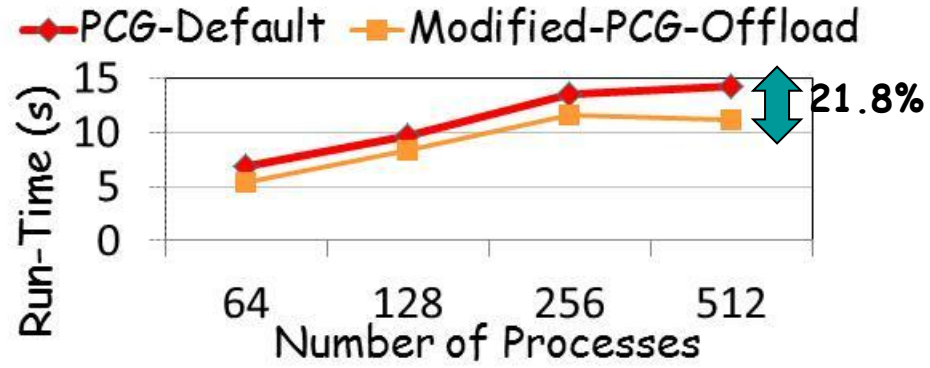
Application Benefits with Non-Blocking Collectives based on CX-2 Collective Offload



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



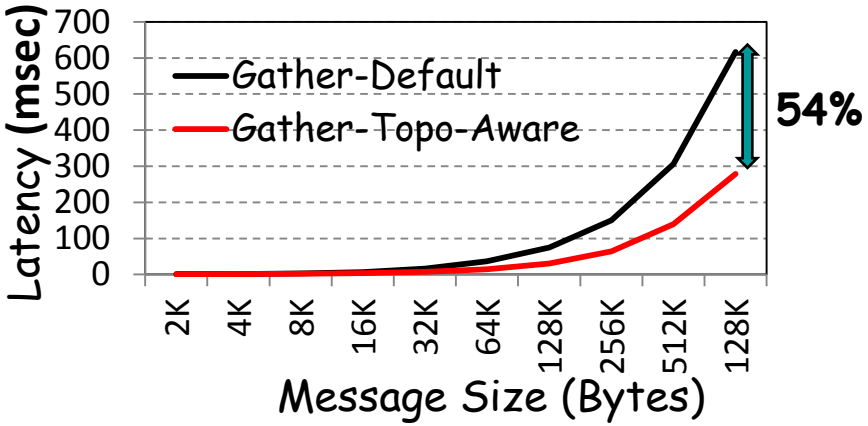
Modified HPL with Offload-Bcast does up to 4.5% better than default version (512 Processes)

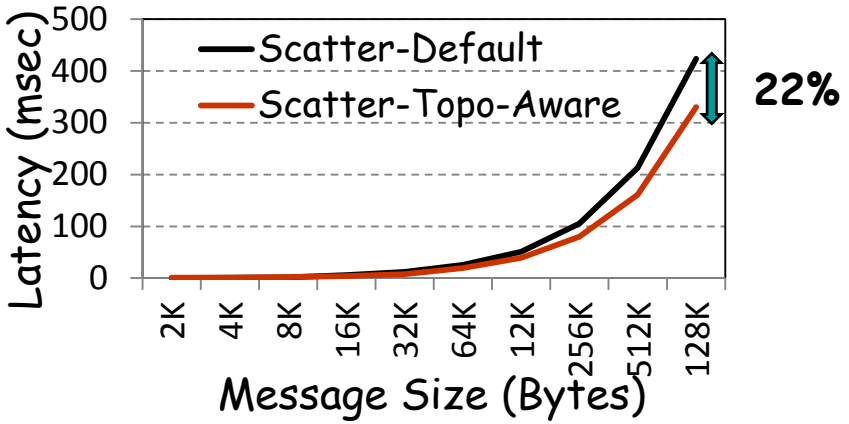


Modified Pre-Conjugate Gradient Solver with Offload-Allreduce does up to 21.8% better than default version

- K. Kandalla, H. Subramoni, K. Tomko, D. Pekurovsky, S. Sur and D. K. Panda, 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, H. Subramoni, J. Vienne, K. Tomko, S. Sur and D. K. Panda, Designing Non-blocking Broadcast with Collective Offload on InfiniBand Clusters: A Case Study with HPL, HotI 2011
- K. Kandalla, U. Yang, J. Keasler, T. Kolev, A. Moody, H. Subramoni, K. Tomko, J. Vienne and D. K. Panda, Designing Non-blocking Allreduce with Collective Offload on InfiniBand Clusters: A Case Study with Conjugate Gradient Solvers, IPDPS '12

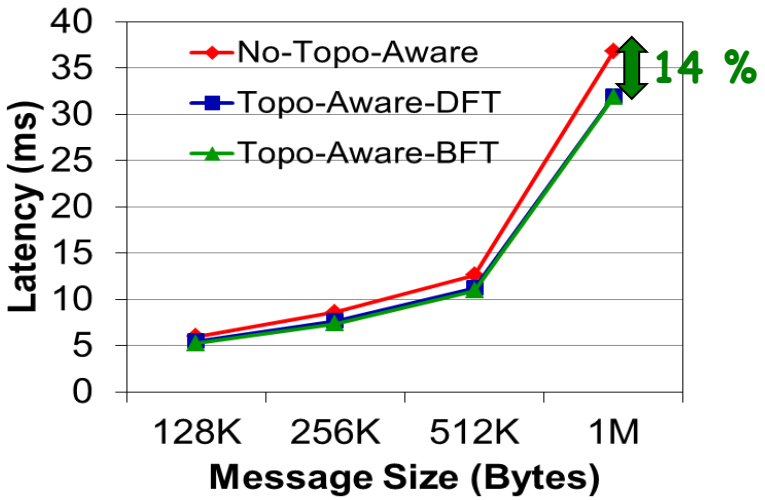
Topology-Aware Collectives

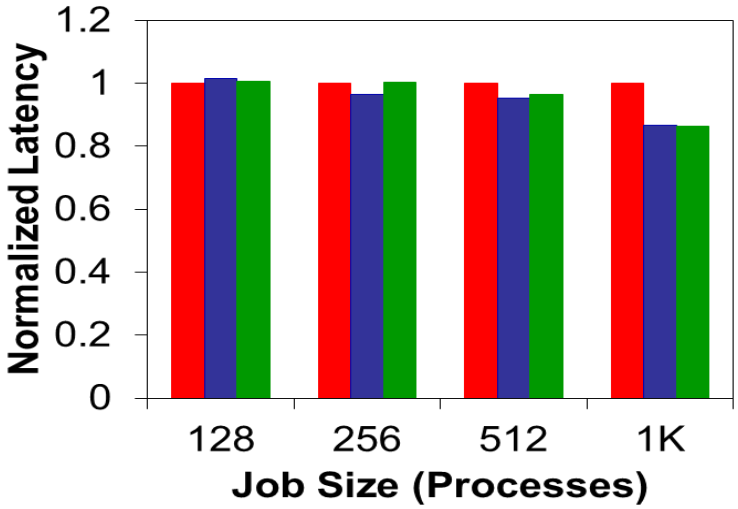




Default (Binomial) Vs Topology-Aware Algorithms with 296 Processes

K. Kandalla, H. Subramoni, A. Vishnu and D. K. Panda, Designing Topology-Aware Collective Communication Algorithms for Large Scale Infiniband Clusters: Case Studies with Scatter and Gather, CAC '10



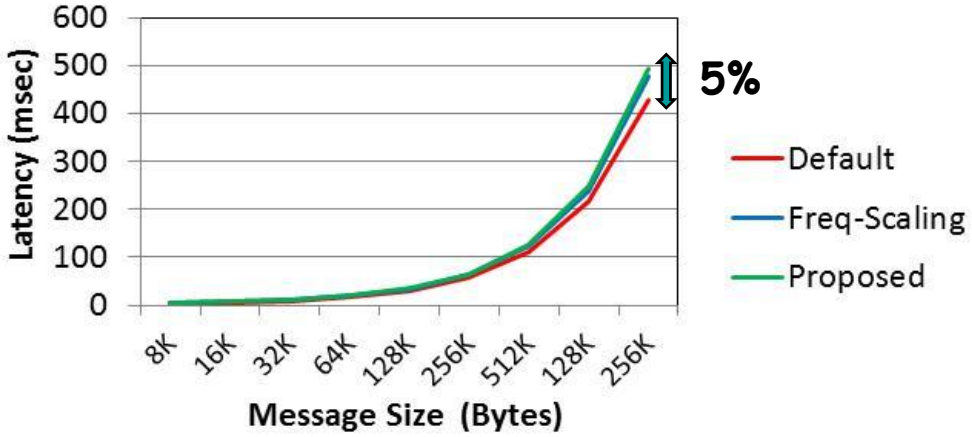


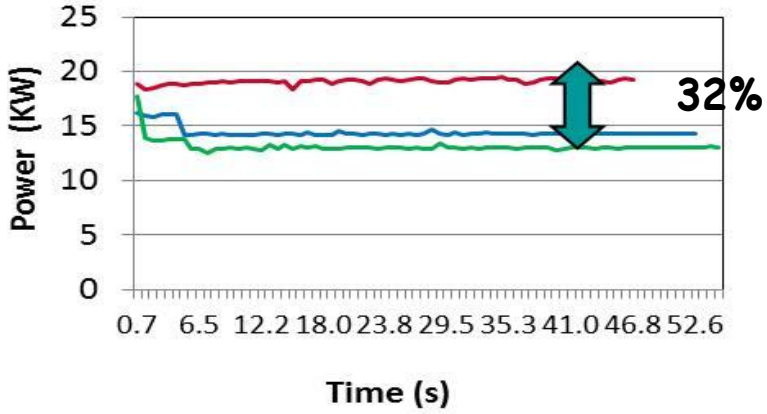
Impact of Network-Topology Aware Algorithms on Broadcast Performance

H. Subramoni, K. Kandalla, J. Vienne, S. Sur, B. Barth, K.Tomko, R. McLay, K. Schulz, and D. K. Panda, Design and Evaluation of Network Topology-/Speed- Aware Broadcast Algorithms for InfiniBand Clusters, Cluster '11

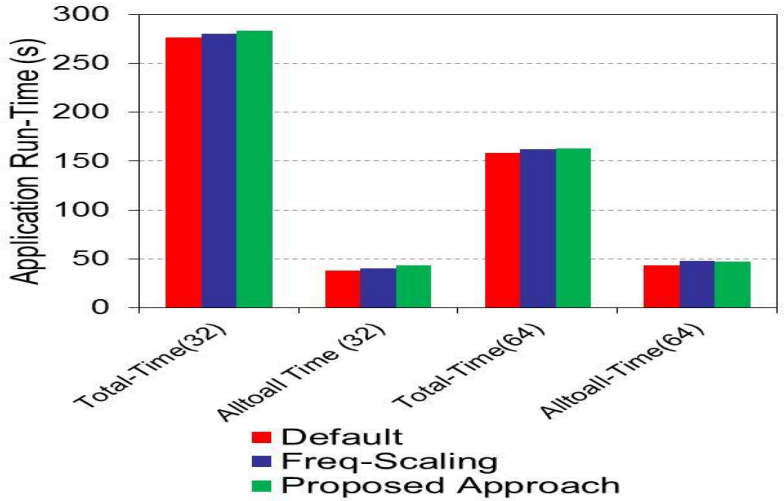
Power and Energy Savings with Power-Aware Collectives

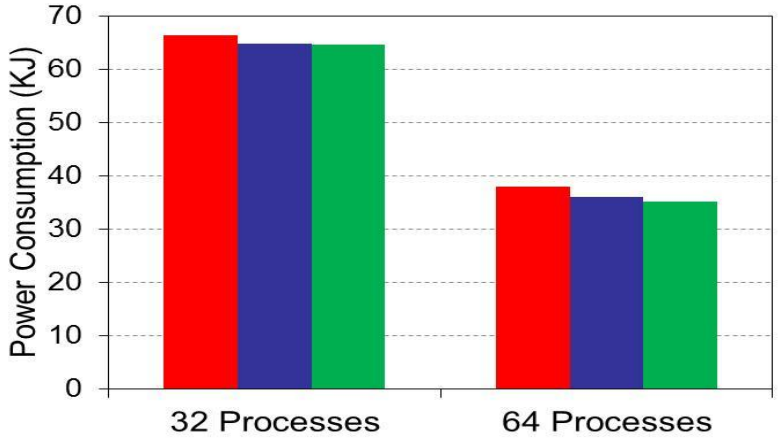
Performance and Power Comparison : MPI_Alltoall with 64 processes on 8 nodes





CPMD Application Performance and Power Savings:





K. Kandalla, E. Mancini, Sayantan Sur and D. K. Panda, Designing Power Aware Collective Communication Algorithms for InfiniBand Clusters, ICPP '10

More Details on MVAPICH2 Features, Performance Optimization, Tuning and Future Plans for Exascale Computing

- Two presentations at HPC Advisory Council Lugano Conference (March 13th and 14th, 2012)
  - http://www.hpcadvisorycouncil.com/events/2012/Switzerland-Workshop/Presentations/Day_1/3_OSU.pdf
  - http://www.hpcadvisorycouncil.com/events/2012/Switzerland-Workshop/Presentations/Day_2/4_OSU.pdf
- Videos of these presentations available from insideHPC.com

Funding Acknowledgments

Funding Support by

Equipment Support by

Personnel Acknowledgments

Current Students

- J. Chen (Ph.D.)
- V. Dhanraj (M.S.)
- N. Islam (Ph.D.)
- J. Jose (Ph.D.)
- K. Kandalla (Ph.D.)
- M. Luo (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Potluri (Ph.D.)
- R. Rajachandrasekhar (Ph.D.)
- M. Rahman (Ph.D.)
- A. Singh (Ph.D.)
- H. Subramoni (Ph.D.)

Past Students

- P. Balaji (Ph.D.)
- D. Buntinas (Ph.D.)
- S. Bhagvat (M.S.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- P. Lai (Ph. D.)
- J. Liu (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- S. Pai (M.S.)
- G. Santhanaraman (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Current Post-Docs

- J. Vienne
- H. Wang

Current Programmers

- M. Arnold
- D. Bureddy
- J. Perkins

Past Post-Docs

- X. Besseron
- H.-W. Jin
- E. Mancini
- S. Marcarelli

Past Research Scientist

- S. Sur

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