Follow us on

https://twitter.com/mvapich

2024 OFA Virtual Workshop

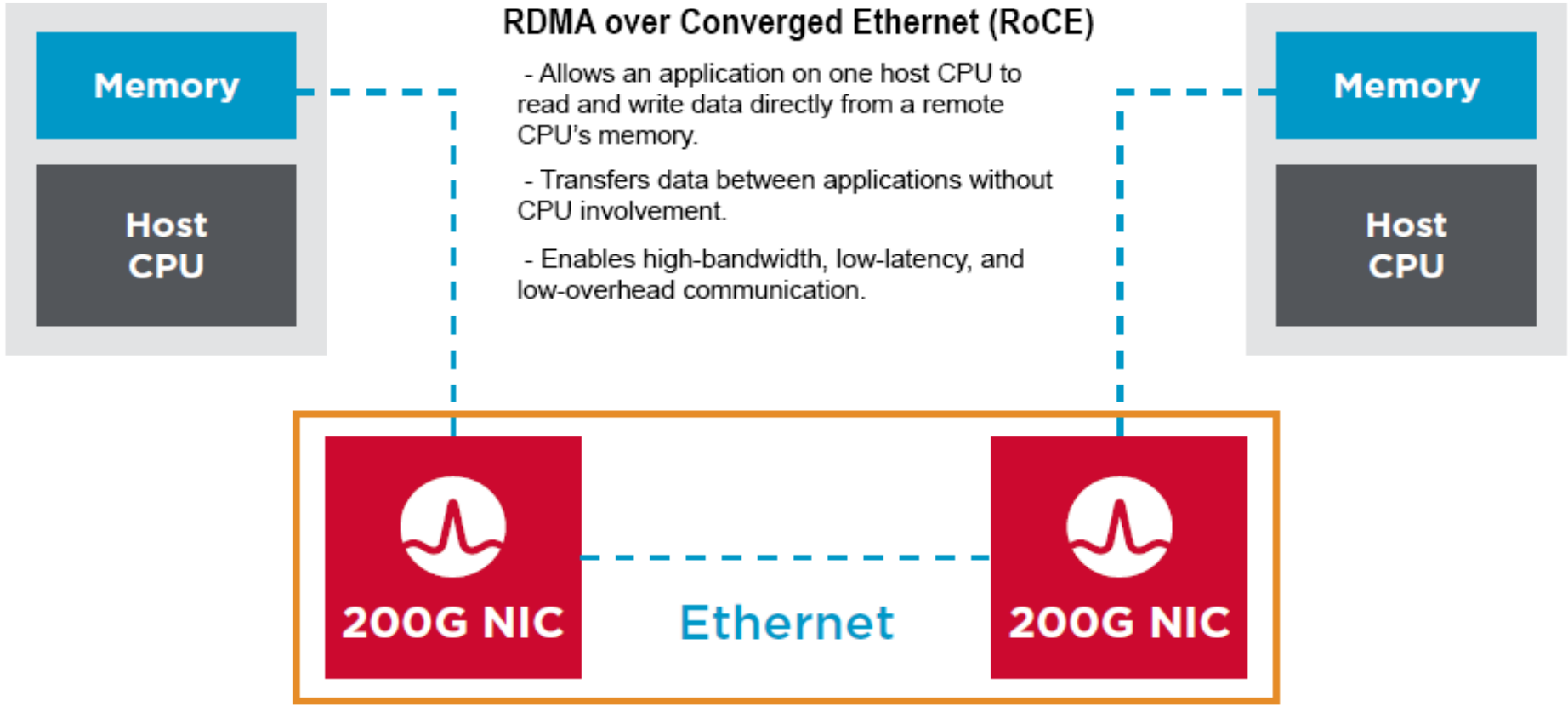
HIGH PERFORMANCE & SCALABLE MPI LIBRARY OVER BROADCOM ROCE

Mustafa Abduljabbar▫, Shulei Xu▫, Noah Pavuk▫, and Hemal Shah*

* Core Switching Group (CSG), Broadcom Inc

▫ Network Based Computing Laboratory (NOWLAB), The Ohio State University

Introduction



https://techdocs.broadcom.com/us/en/storage-and-ethernet-connectivity/ethernet-nic-controllers/bcm957xxx/adapters/RDMA-over-Converged-Ethernet.html

Why RoCE for HPC?

- Enhanced Performance: RoCE delivers significantly lower latency and higher throughput compared to traditional Ethernet, empowering HPC applications to achieve peak performance.
- Optimized Efficiency: RoCE offloads RDMA operations to specialized hardware, reducing CPU overhead and freeing up valuable processing resources for critical computational tasks.
- Cost-Effective Solution: RoCE leverages existing Ethernet infrastructure.
- Scalable and Flexible: RoCE supports a range of Ethernet speeds and Layer 3 routing.

Goal: Highly optimized MPI for Broadcom RoCEv2

1. MVAPICH-CPU release: Optimizing MPI communication operations on new generation Broadcom adapters
   - We provide support for newer generation Broadcom network adapters (Thor 200 Gbps) in MVAPICH2 and optimize the communication protocols (RC, UD, Hybrid)
   - Focus will be towards point-to-point operations (two-sided) and frequently used collective operations (such as Allreduce and Alltoall).
   - Benefits of these designs will be studied at the applications level.
   - These design changes will be incorporated into the future MVAPICH release.
2. MVAPICH-GPU release: Exploring the use of Peer Direct capabilities in new Broadcom adapters for high-performance data transfers to/from GPU device memory
   - Broadcom has introduced support for Peer Direct RDMA to enable high-performance communication operations from device memory.
   - We study and evaluate the performance of Broadcom's Peer Direct with Thor adapters.
   - We explore designs in MVAPICH2-GDR for accelerating relevant portions of device-based communication operations using Peer Direct technology with Thor adapters. The focus will be on point-to-point intra-node, inter-node, and commonly used collectives (Allreduce and Alltoall).
   - The designs will be incorporated into the future MVAPICH2-GDR release.

Overview of the MVAPICH Project

- High Performance open-source MPI Library
- Support for multiple interconnects
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, OPX, Broadcom RoCE, Intel Ethernet, Rockport Networks, Slingshot 10/11
- Support for multiple platforms
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and AMD)
- Started in 2001, first open-source version demonstrated at SC '02
- Supports the latest MPI-3.1 standard
- http://mvapich.cse.ohio-state.edu
- Additional optimized versions for different systems/environments:
  - MVAPICH2-X (Advanced MPI + PGAS), since 2011
  - MVAPICH2-GDR with support for NVIDIA (since 2014) and AMD (since 2020) GPUs
  - MVAPICH2-MIC with support for Intel Xeon-Phi, since 2014
  - MVAPICH2-Virt with virtualization support, since 2015
  - MVAPICH2-EA with support for Energy-Awareness, since 2015
  - MVAPICH2-Azure for Azure HPC IB instances, since 2019
  - MVAPICH2-X-AWS for AWS HPC+EFA instances, since 2019
- Tools:
  - OSU MPI Micro-Benchmarks (OMB), since 2003
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015

- Used by more than 3,375 organizations in 91 countries
- More than 1.77 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '23 ranking)
  - 11th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 29th , 448, 448 cores (Frontera) at TACC
  - 46th, 288,288 cores (Lassen) at LLNL
  - 61st, 570,020 cores (Nurion) in South Korea and many others
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)
- Partner in the 29th ranked TACC Frontera system
- Empowering Top500 systems for more than 18 years

Overview

- Introduction
- **Performance Characterization**
  - **MPI performance overheads vs. IB level**
- Latency and Message Rate Optimization
- Performance Evaluation
  - Micro-benchmark level
  - Application level
- MVAPICH 3.0 Performance Evaluation

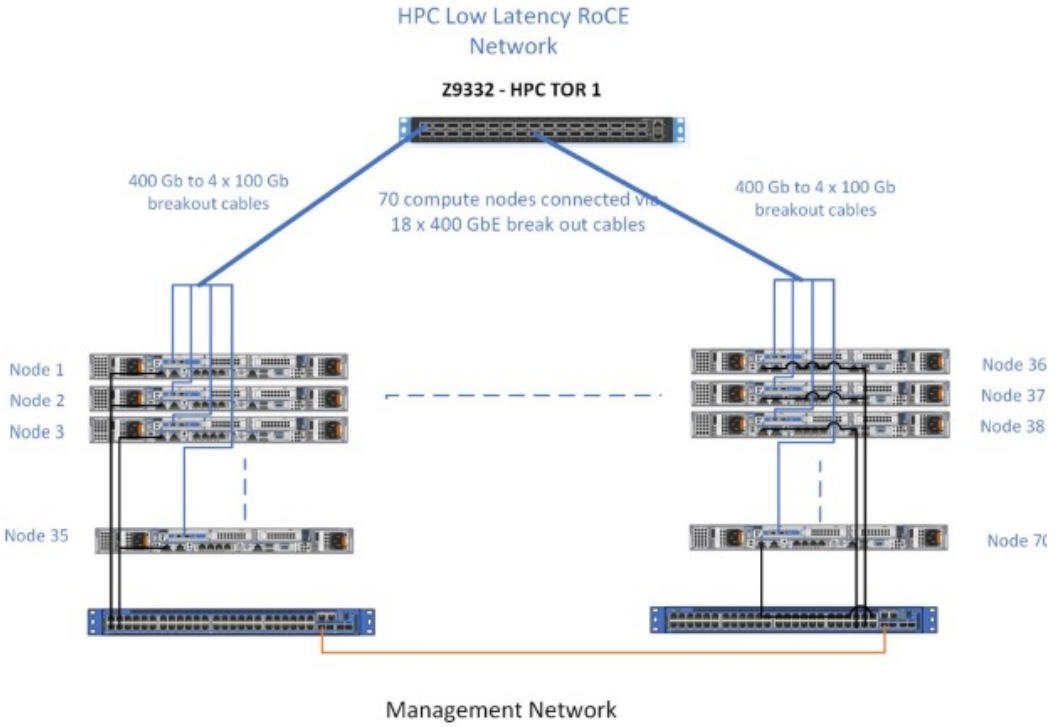
Configuration & Runtime

- MVAPICH2 Runtime:
  - `RC: MV2_USE_UD_HYBRID=0 MV2_USE_ONLY_UD=0`
  - `UD: MV2_USE_UD_HYBRID=0 MV2_USE_ONLY_UD=1`
- UCX 1.12.1:
  - `./configure --prefix=<UCX_INSTALL_PATH>`
- OpenMPI 4.1.4 (w/ UCX 1.12.1):
  - `./configure –prefix=<INSTALL_PATH> --with-ucx=<UCX_INSTALL_PATH>`
- OpenMPI Runtime:
  - `mpirun -np <NP> -npernode <PPN> -hostfile hosts --mca pml ucx –x UCX_TLS=self,sm,rc_v /path/to/cp2k.popt -i /path/to/inputfile`

Cluster Setup

Blue Bonnet Cluster*

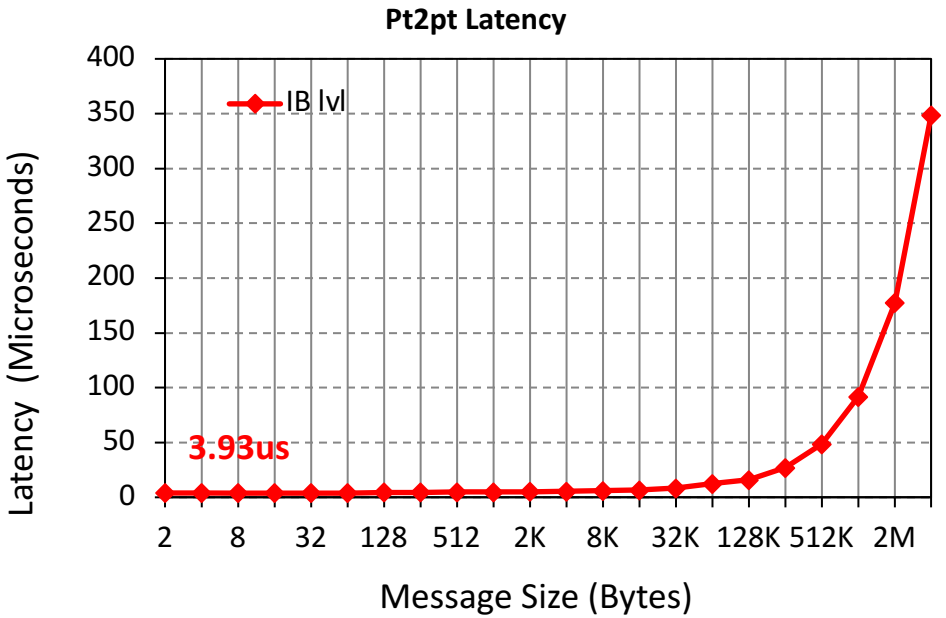
	R6525
Chassis Configuration	8 x 2.5" SAS/SATAChassis
Processor Configuration	2 Sockets of AMD EPYC 7713 2.0 GHz 64C processors
Memory	16 x 16 GB @ 3200 MB/s DDR4 = 256 GB
Storage - OS Boot	2 x 480 GB SSD SATA Mix Use (RAID 1)
Storage Controller	PERC H345
Network Cards	**Add-in-Card for HPC traffic :** Broadcom 57508 Dual Port 100GbE QSFP Adapter, PCIe Low Profile (Thor) **Integrated LOM:** 2 x 1 GbE Base-T Broadcom **Optional OCP 3.0 Card:** Broadcom 57414 Dual Port 10/25 GbE SFP28
iDRAC	iDRAC9 Express
PCIe Riser	Riser Config 2, 1 x 16 Gen4 LP PCIe slot (CPU1), 2 x 16 LP PCIe slot (CPU2)
Power Supply	Redundant PSU (1+1) 800 W
RDMA switch Fabric	Dell PowerSwitch Z9332 (400 GbE) (Broadcom Tomahawk3)
Management Fabric	Dell PowerSwitch S3248 (1 GbE)

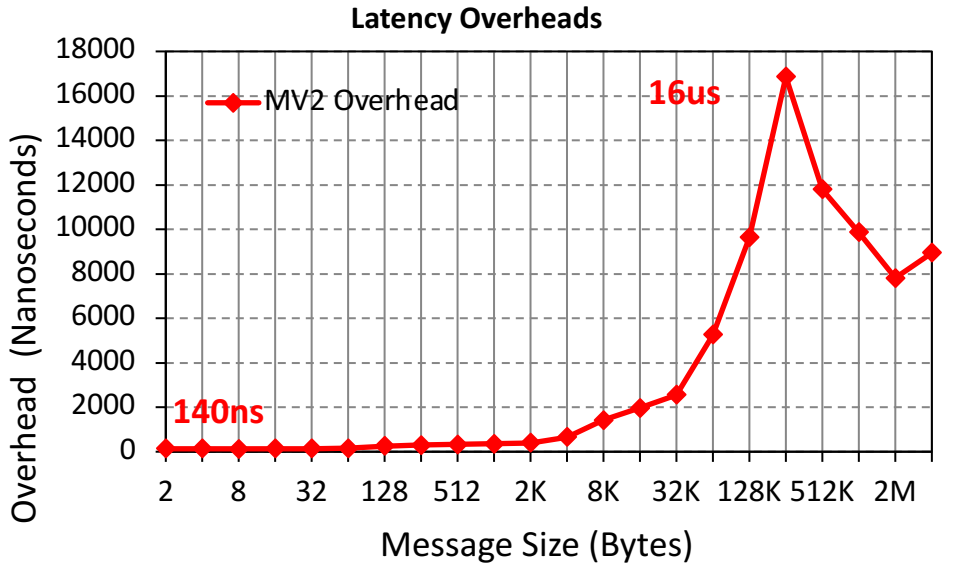


- Nodes split across 2 physical racks
- Single switch topology with all 70 nodes connected via breakout cable

*Courtesy of DELL Technology

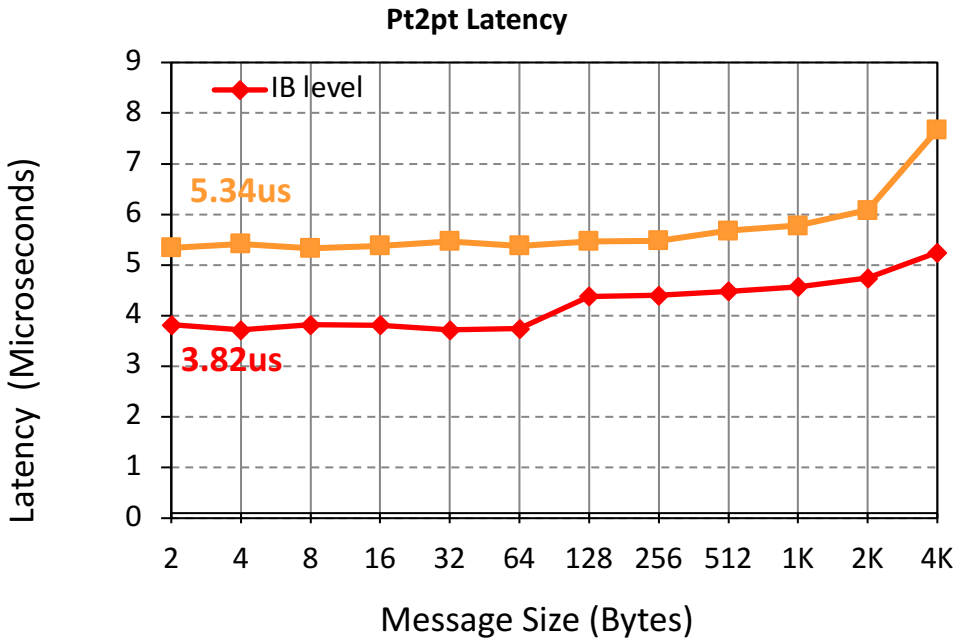
MVAPICH2.3.x-Broadcom RDMA RC Latency on RoCE 100 GbE

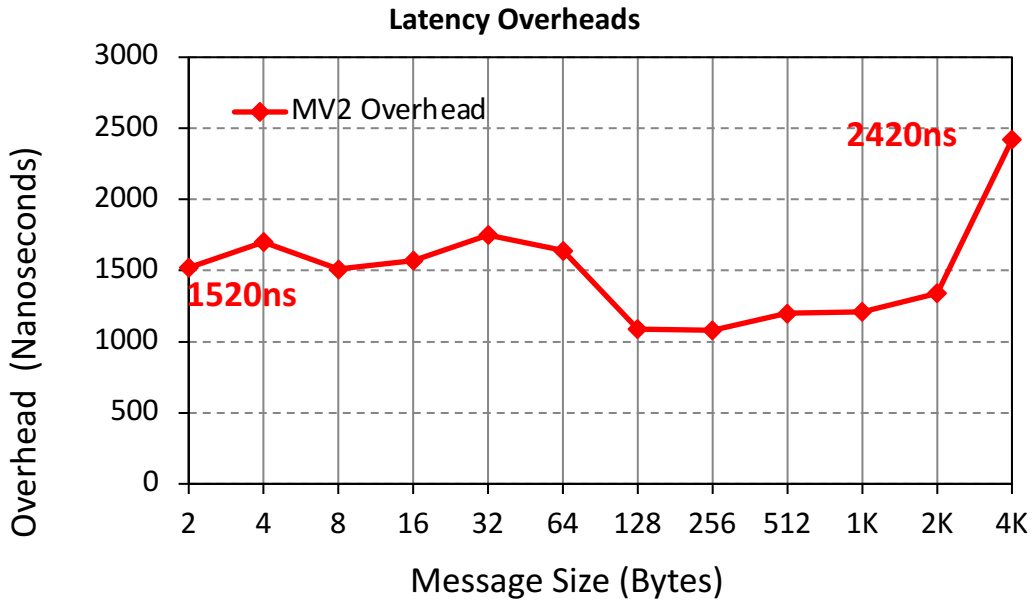




- IBPerf latency performance on the left side, MVAPICH overhead on the right side.
- MVAPICH not shown on the left due to invisible scale difference
- Overhead is in the order of **nanoseconds** to a few microseconds
  - Thanks to our optimized pt-to-pt parameter tuning

MVAPICH2.3.x-Broadcom RDMA UD Latency on RoCE 100 GbE





- IBPerf latency performance on the left side, MVAPICH overhead on the right side.
- UD overhead of 1.5 – 2.4 us is add on top of IB level latency
  - This has some impact on the application-level when switching to UD at a large scale.
  - Issue is unique to Broadcom RoCE + 2.3 series, based on our experience.

Overview

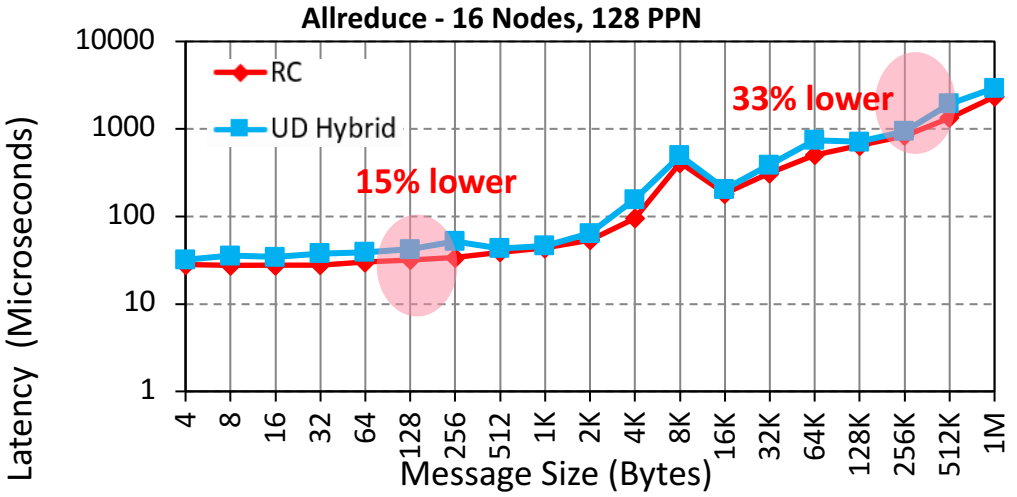
- Introduction
- Performance Characterization
  - MPI performance overheads vs. IB level
- **Latency and Message Rate Optimization**
- Performance Evaluation
  - Micro-benchmark level
  - Application level
- MVAPICH 3.0 Performance Evaluation

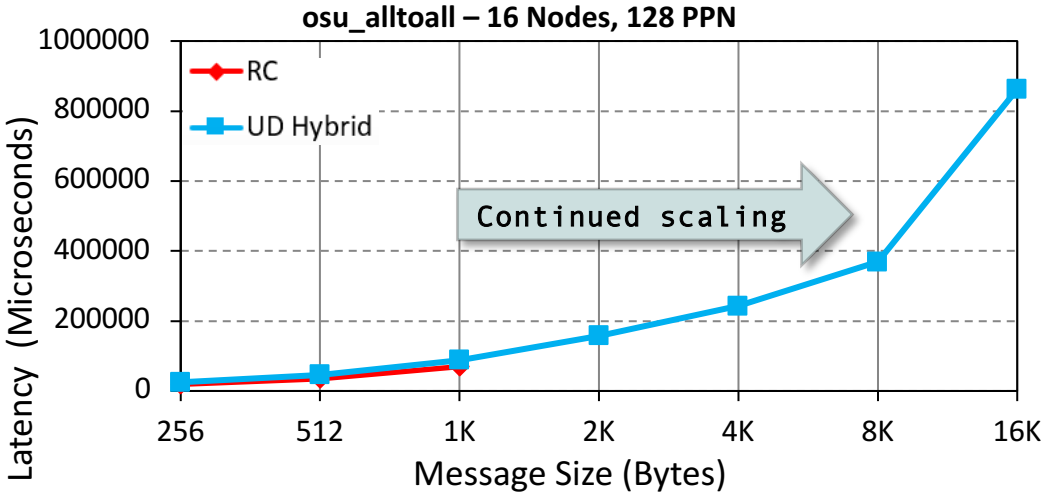
Latency and Message Rate Optimization

- **Add corresponding point-to-point & collective tuning tables**
  - **For up to 64 nodes x 128 PPN = 8192 processes**
  - **Based on Dell Bluebonnet (CPU) system and Rattler2 (GPU) system**
- Enhanced UD+RC hybrid transport mode tuned for Broadcom adapter
- Optimized default CPU mapping policy
- Support for asynchronous threading progress
- Startup Optimization
- Point-to-point Message Coalescing
- SGL packetized eager communication

UD/RC Hybrid Transport Protocol Analysis

- RC has better performance vs. UD in most cases
- UD Hybrid becomes exclusive on large scales (e.g. alltoall with >= 16 nodes)
- Tuned hybrid transport mode
  - Use RC for small scale & message sizes
  - Use UD for the other cases





Latency and Message Rate Optimization (Cont'd)

- Add corresponding point-to-point & collective tuning tables
- Enhanced UD+RC hybrid transport mode tuned for Broadcom adapter
- **Optimized default CPU mapping policy**
  - **Make hybrid spread CPU mapping policy as default**
  - Example showing in right table:

```
-------------CPU AFFINITY-------------
RANK: 0  CPU_SET:    0   1   2   3
RANK: 1  CPU_SET:    4   5   6   7
RANK: 2  CPU_SET:    8   9  10  11
RANK: 3  CPU_SET:   12  13  14  15
RANK: 4  CPU_SET:   16  17  18  19
RANK: 5  CPU_SET:   20  21  22  23
RANK: 6  CPU_SET:   24  25  26  27
RANK: 7  CPU_SET:   28  29  30  31
-------------------------------------
```

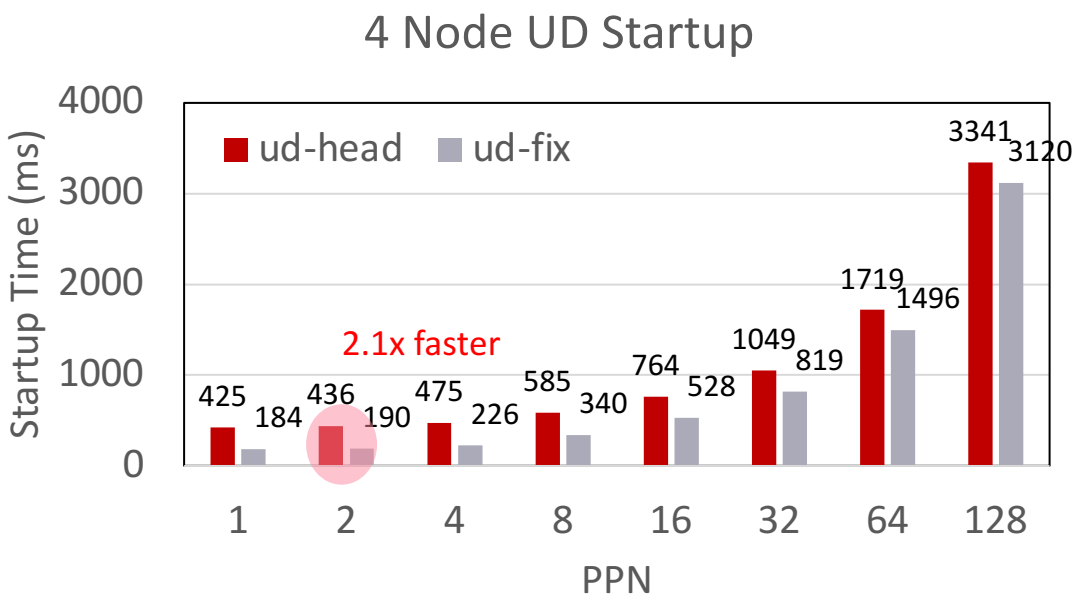
Hybrid-Spread Affinity Policy

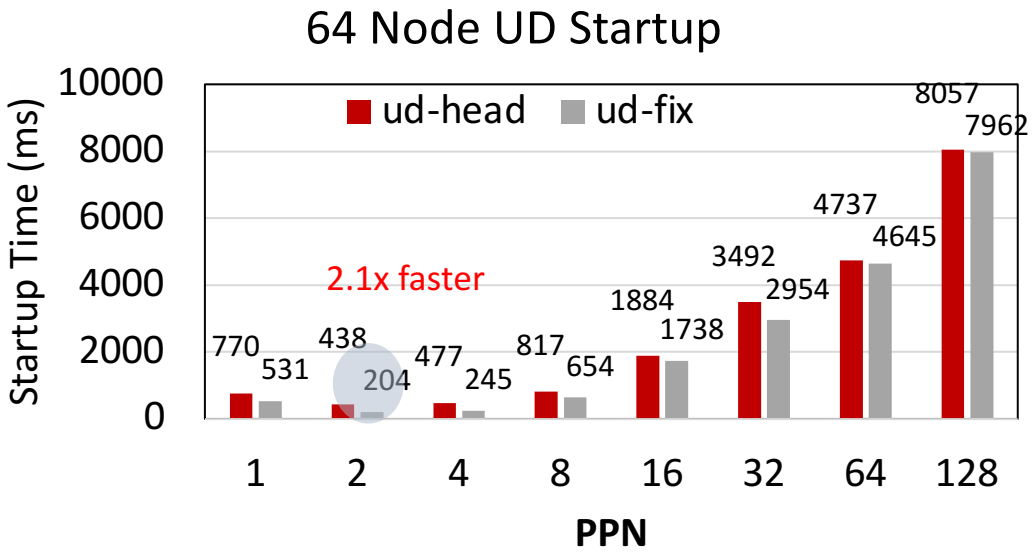
- **Support to enable affinity with asynchronous progress thread**
  - Allow to manage communication in parallel with application computation to achieve better computation/communication overlap
  - Controlled by runtime parameter: **MV2_OPTIMIZED_ASYNC_PROGRESS=1**
- UD Startup Optimization
- Point-to-point Message Coalescing
- SGL packetized eager communication

Latency and Message Rate Optimization (Cont'd)

- Add corresponding point-to-point & collective tuning tables
- Enhanced UD+RC hybrid transport mode tuned for Broadcom adapter
- Optimized default CPU mapping policy
- Support to enable affinity with asynchronous progress thread
- **UD Startup Optimization**
  - Optimize the specific function calls with highest overhead by analysis of the UD startup profiling data
- **Point-to-point Message Coalescing**
  - Combine small messages to reduce send/recv calls
- **SGL packetized eager communication**
  - Use scatter-gather list (SGL) to packetize the eager send requests

UD Startup Optimization

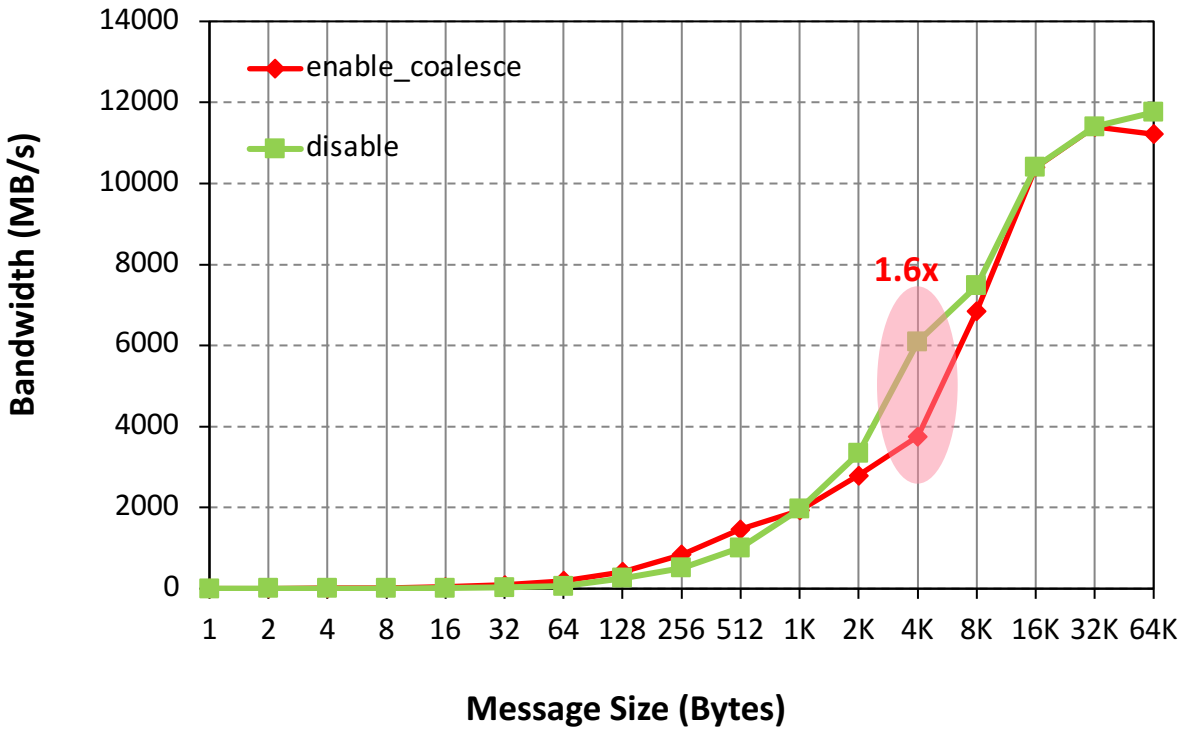




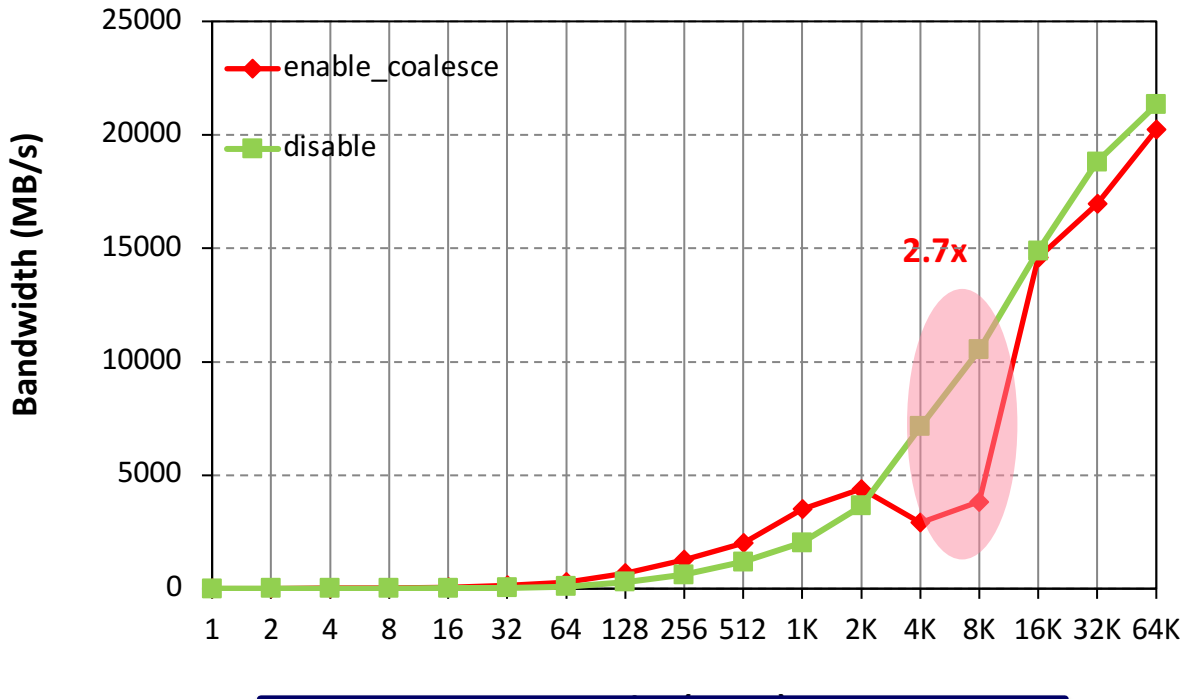
- Provide up to 2.3x faster UD startup in small 4 nodes scale
- Provide up to 2.1x faster UD startup in large 64 nodes scale

Pt-to-Pt Message Coalescing Performance

osu_bw



osu_bibw

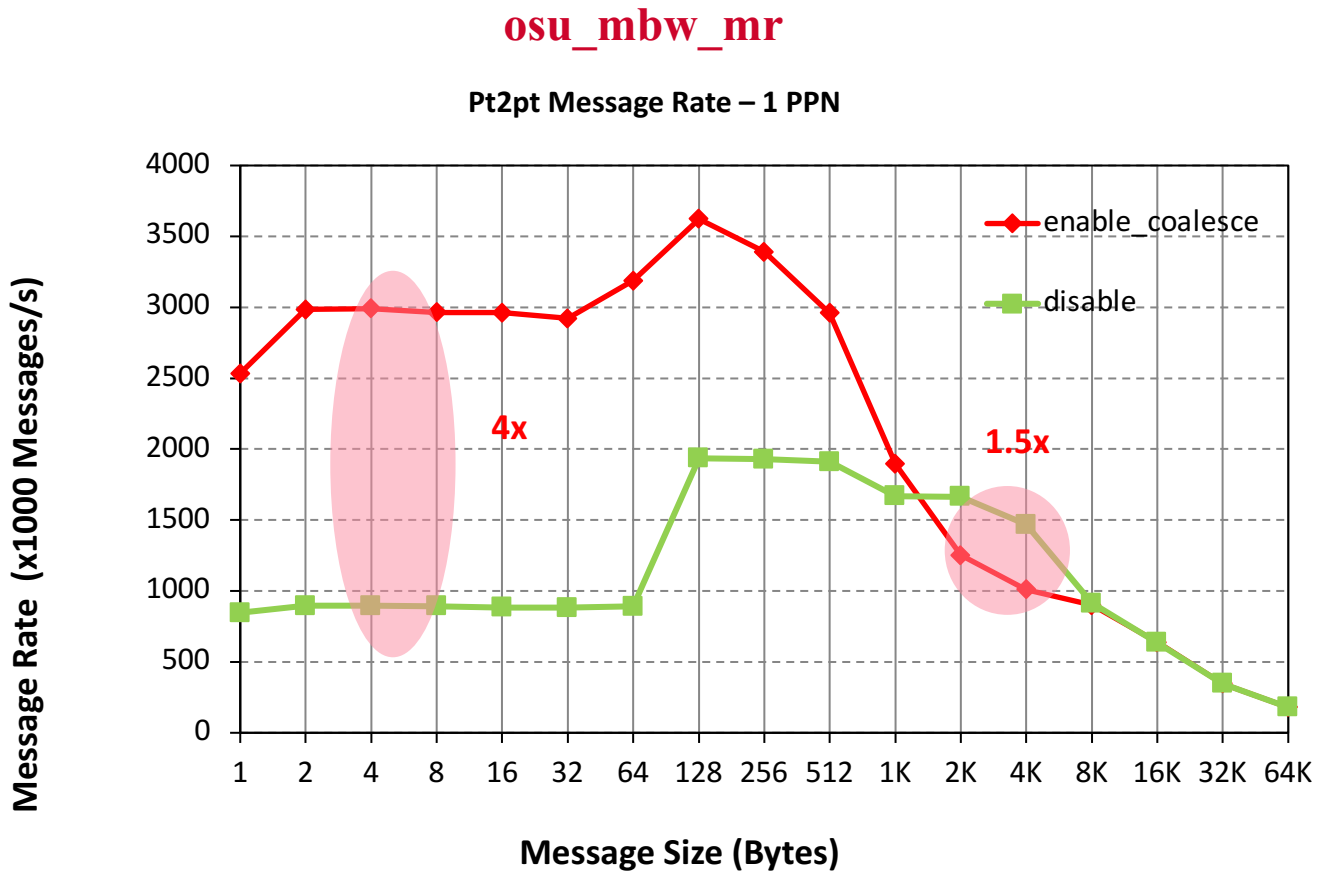


- Enabling/disabling coalescing has an impact on BW and BiBW
  - Coalescing effective up to 1K message size
- Up to 1.6x higher bandwidth, 2.7x higher bi-bandwidth with medium sized messages

The hybrid policy takes advantage of message coalescing below 1KB size and disable it for larger sizes

Pt2pt Message Coalescing – Single-Pair Message Rate

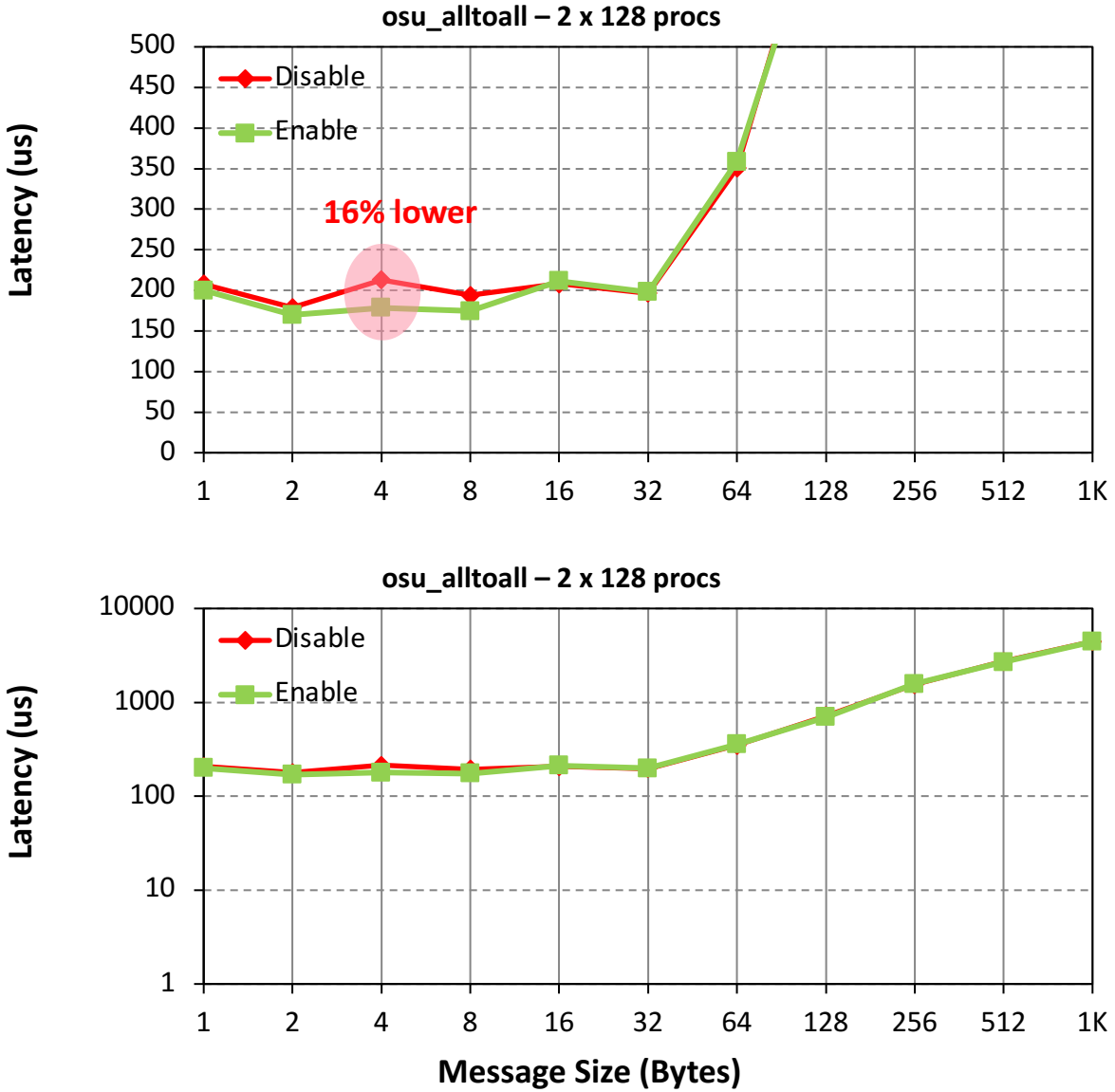
- **Test Name:** Single-Pair Bandwidth and Message Rate Test
- **Evaluation Focus:** Aggregate uni-directional bandwidth and message rate
- **Participants:** 1 process per node
- **Sending Process Behavior:**
  - Sends a fixed number of messages (window size)
  - Sends messages back-to-back to the paired receiving process
  - Waits for a reply from the receiver
- Iterations: Repeated for 1000 iterations



1. Enabling/disabling coalescing has an impact on MR
2. Up to 1.5x higher bandwidth with limits

SGL packetized eager communication – 100 GbE

- Reduce up to 16% alltoall latency for 4 bytes messages size
- Enable by adding `MV2_USE_EAGER_SGL=1` runtime parameter
- Enabled by default for up to 1KB message sizes (`MV2_USE_EAGER_SGL_LIMIT=1k`)

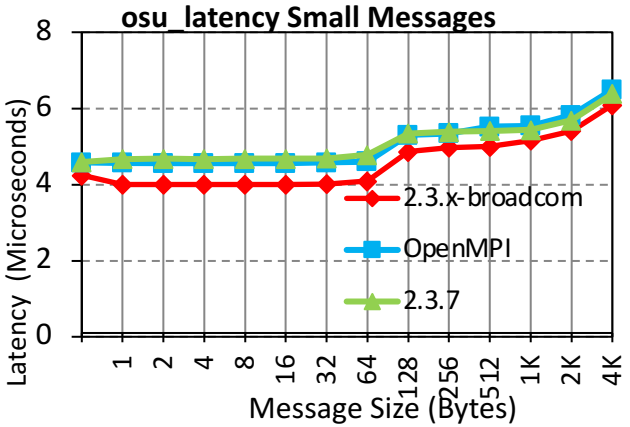


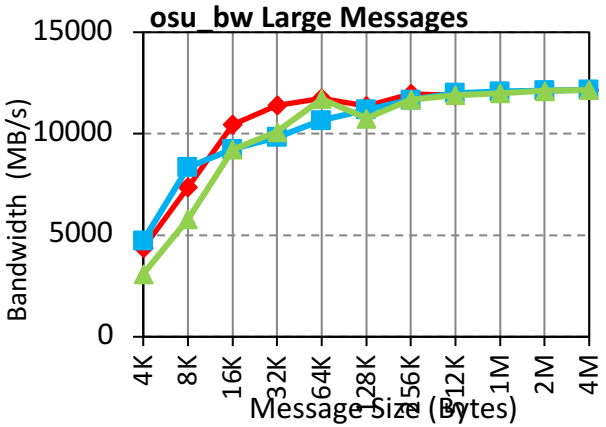
Overview

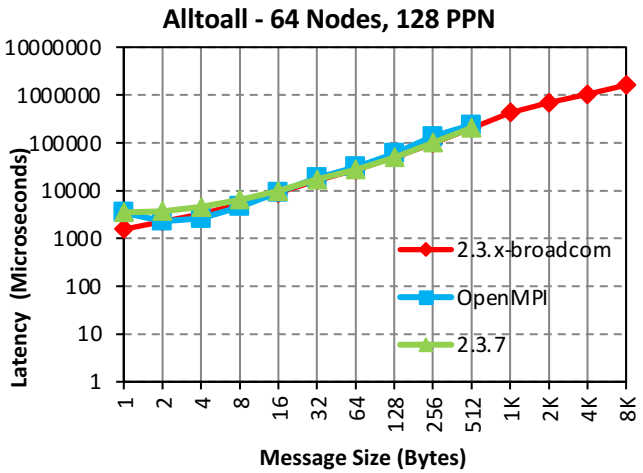
- Introduction
- Performance Characterization
- Latency and Message Rate Optimization
- **Performance Evaluation**
  - **Micro-benchmark level**
  - **Application level**
- MVAPICH 3.0 Performance Evaluation

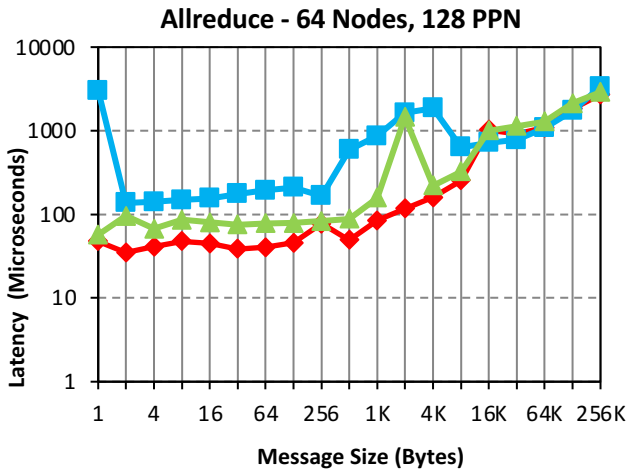
Performance Evaluation – CPU Microbenchmarks

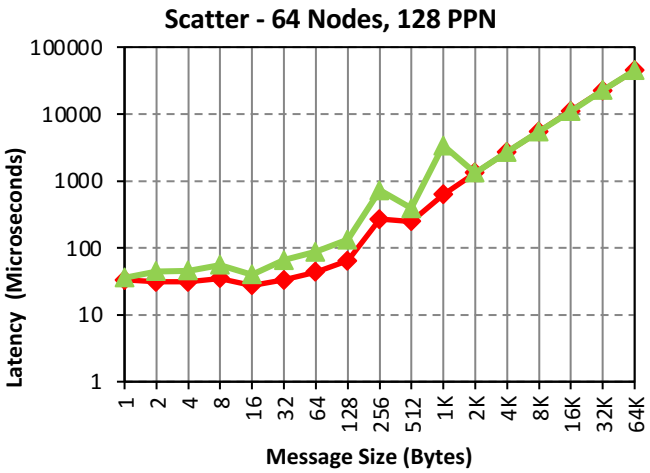
- Experiment results from Dell Bluebonnet
- Up to 20% reduction in small message point-to-point latency
- From 0.1x to 2x increase in bandwidth
- Up to 12.4x lower MPI_Allreduce latency
- Up to 5x lower MPI_Scatter latency





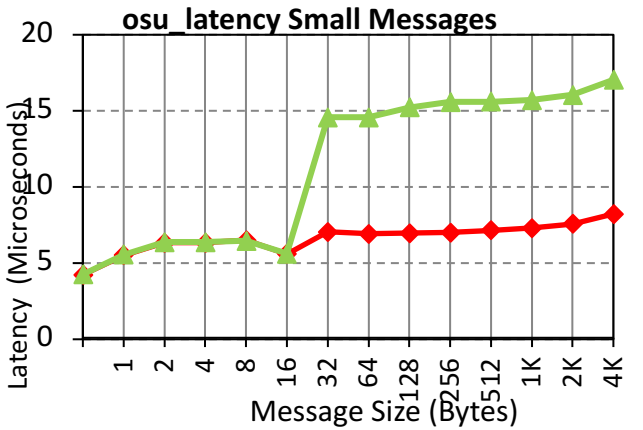


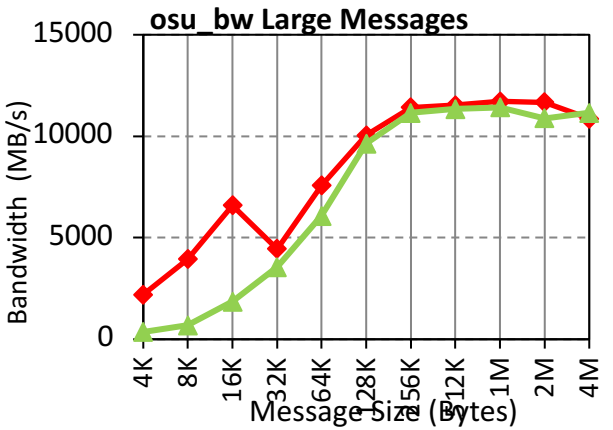


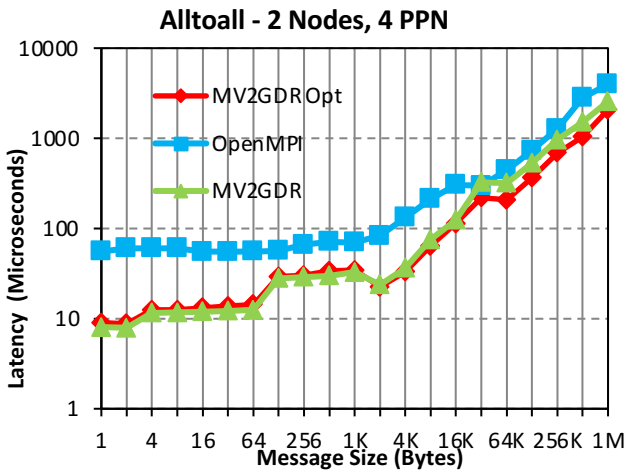


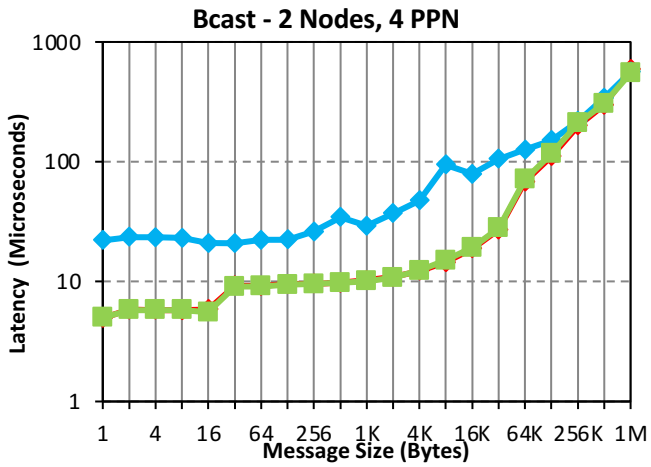
Performance Evaluation – GPU Microbenchmarks

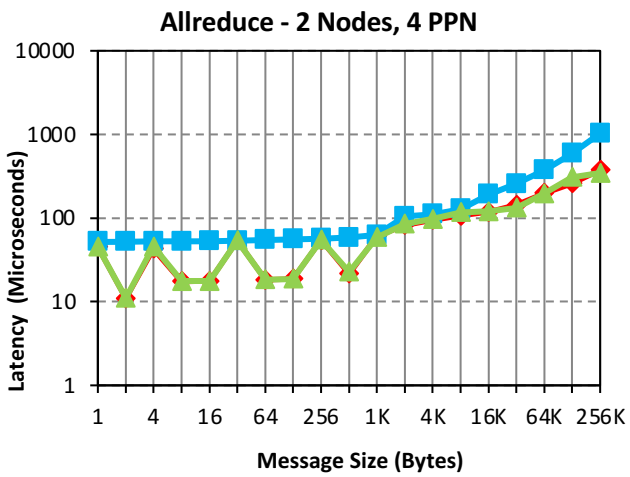
- Experiment results from Rattler2 DELL cluster (A100 GPUs)
- Compared to non-optimized version, up to 2 – 3x reduction in medium to large message point-to-point latency
- Up to 2.6x increase in bandwidth
- Up to 35% reduction in alltoall latency



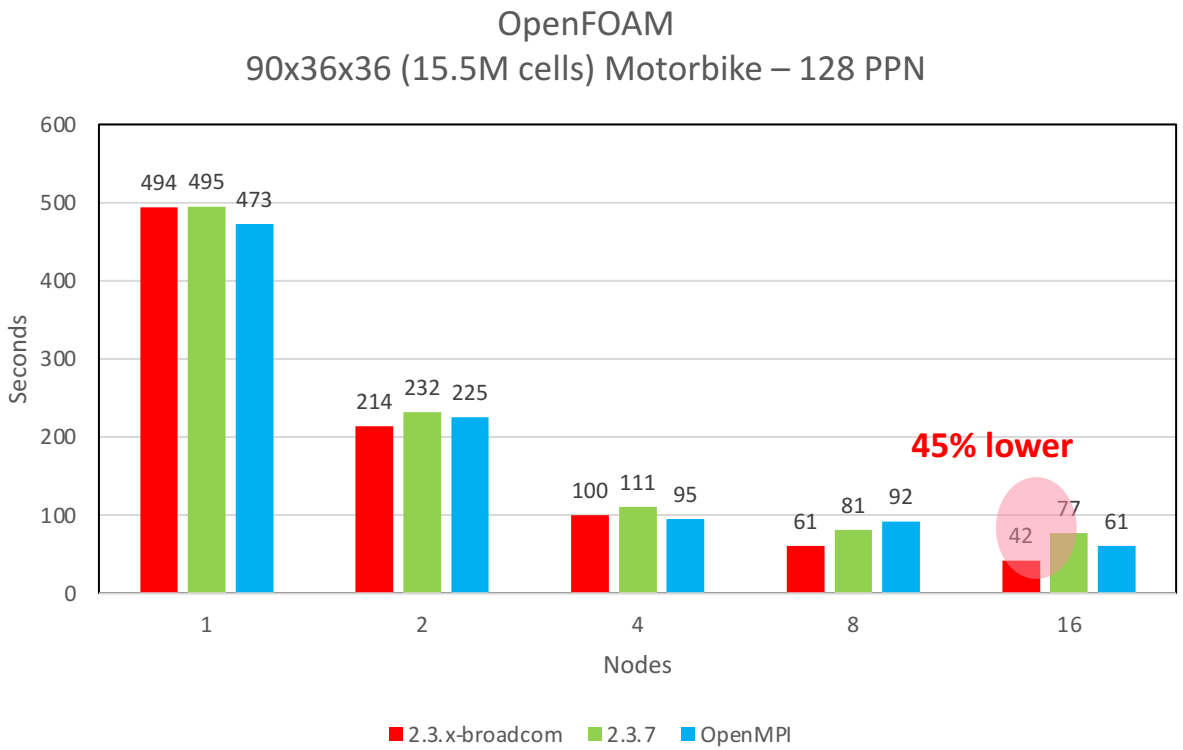


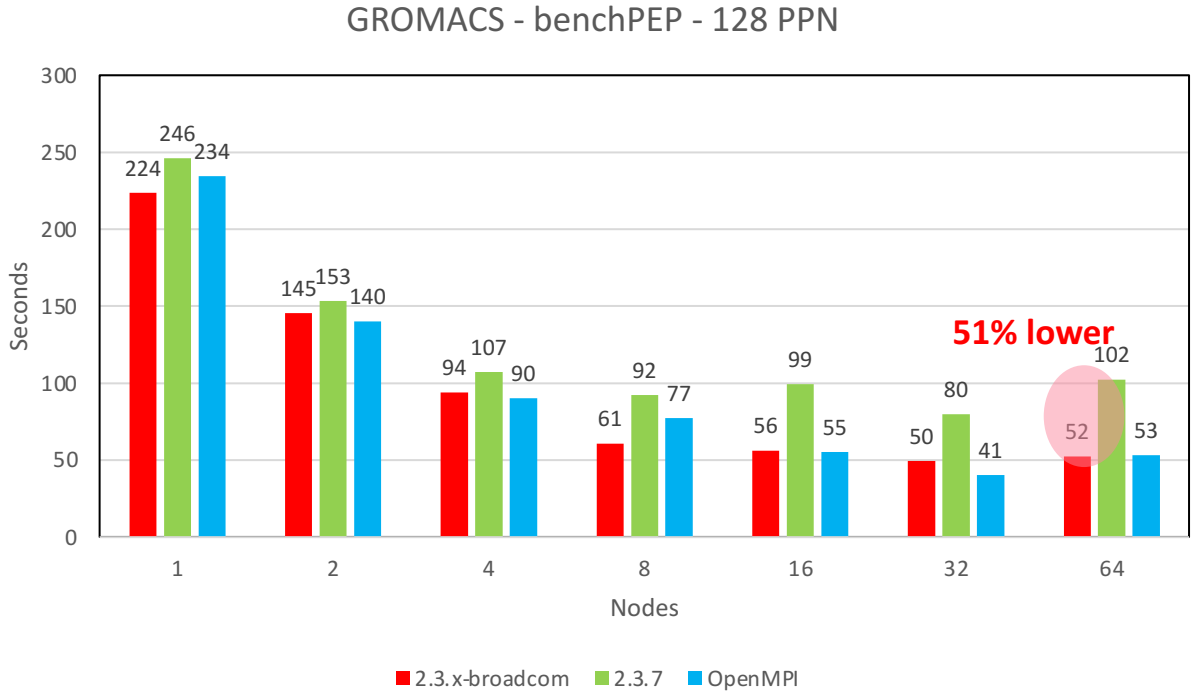






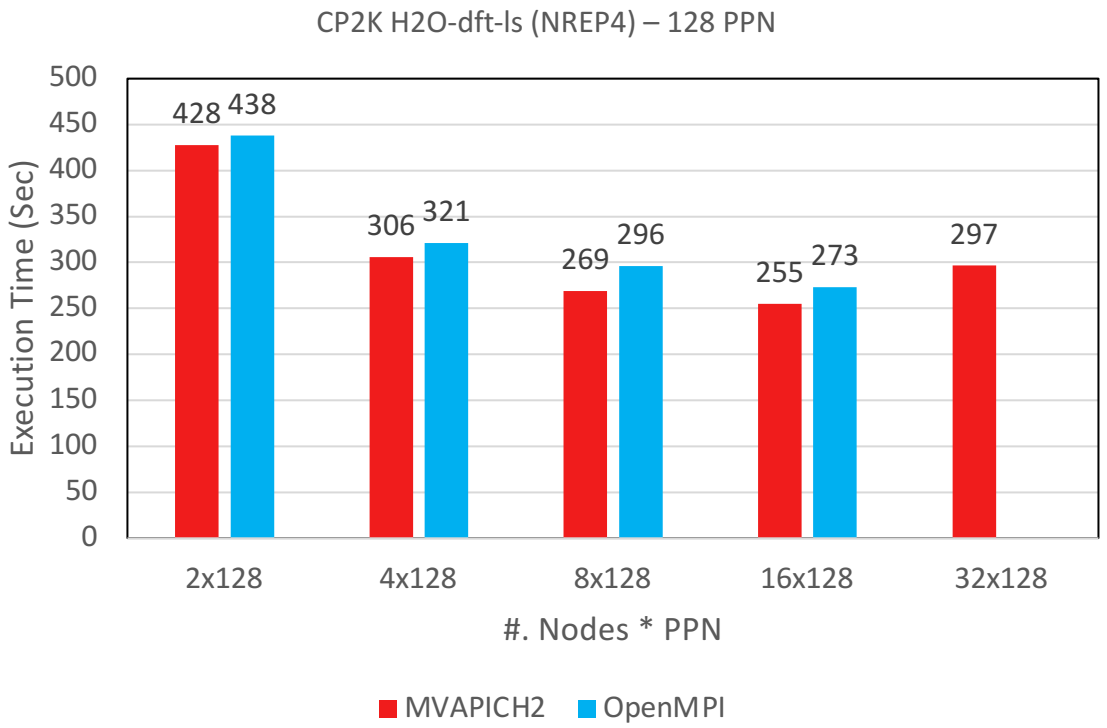
Performance Evaluation – Applications

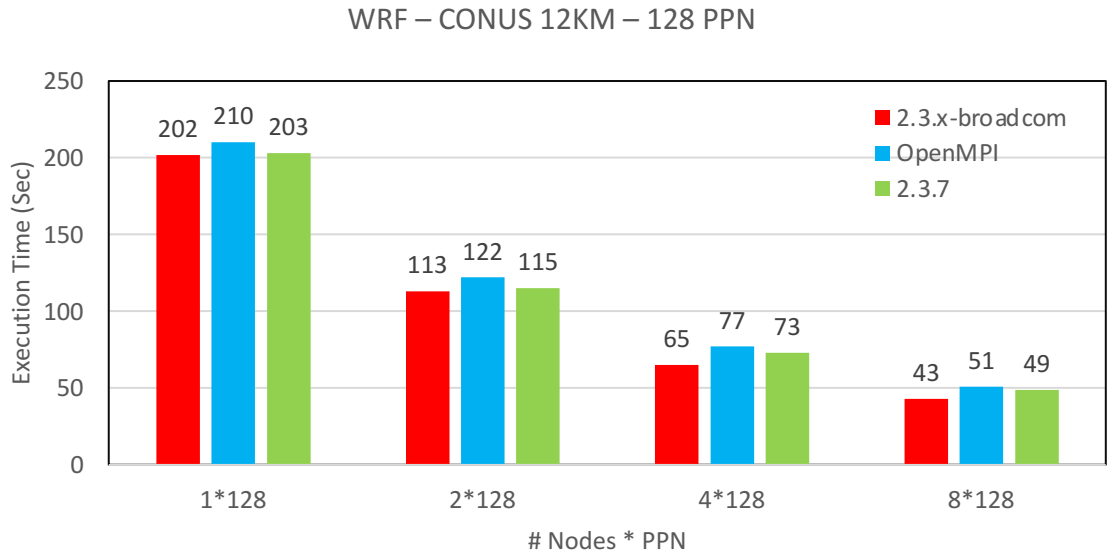


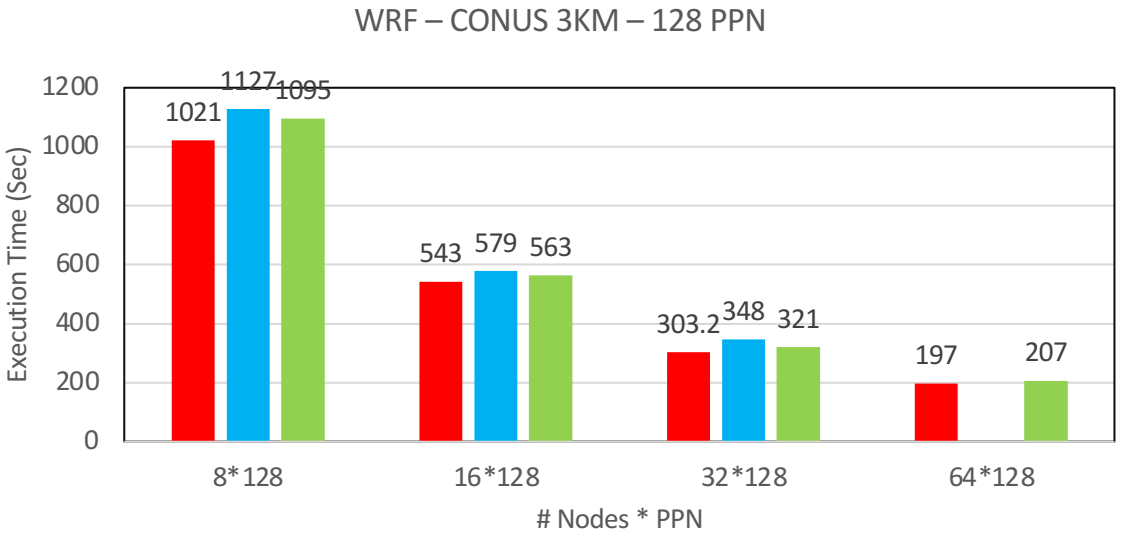


- Reduce up to 45% execution time of OpenFOAM Motorbike on 16 nodes 128 PPN scale
- Reduce up to 51% execution time of GROMACS benchPEP on 64 nodes 128 PPN scale

Performance Evaluation – Applications





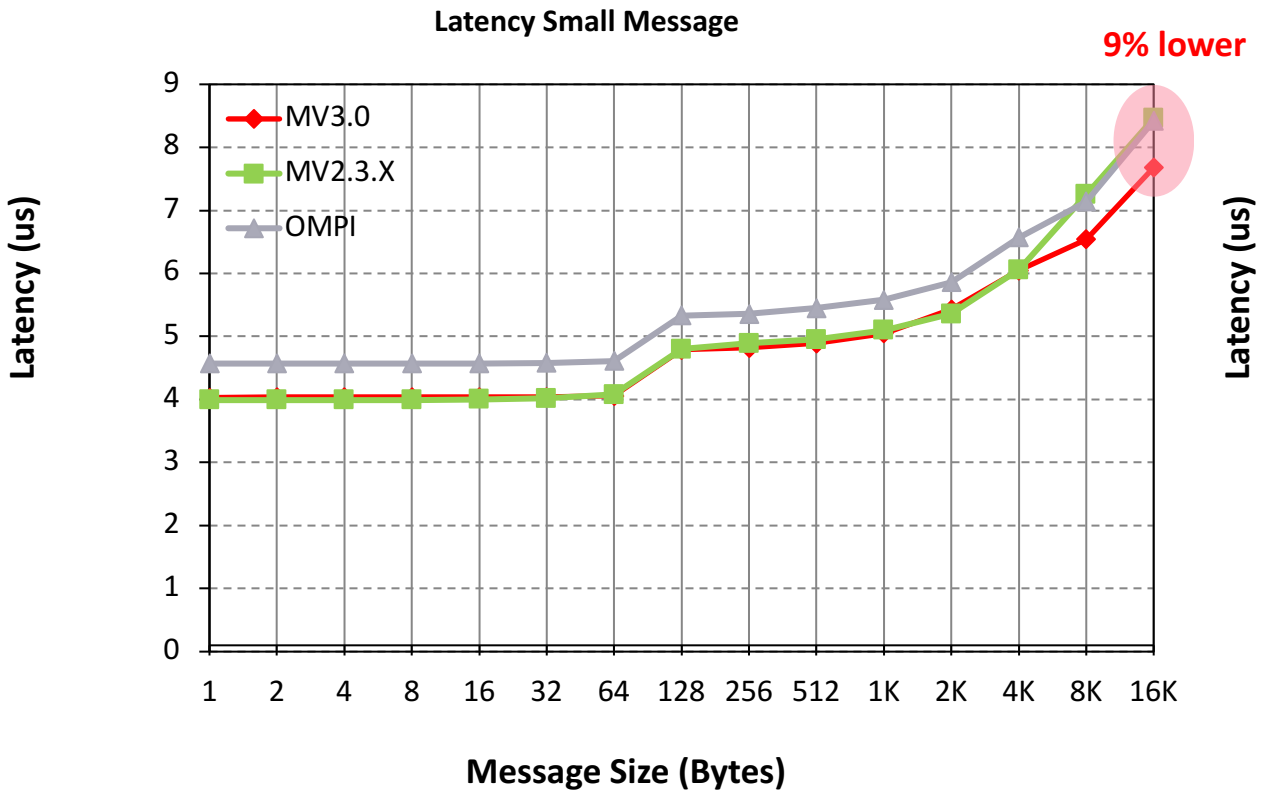


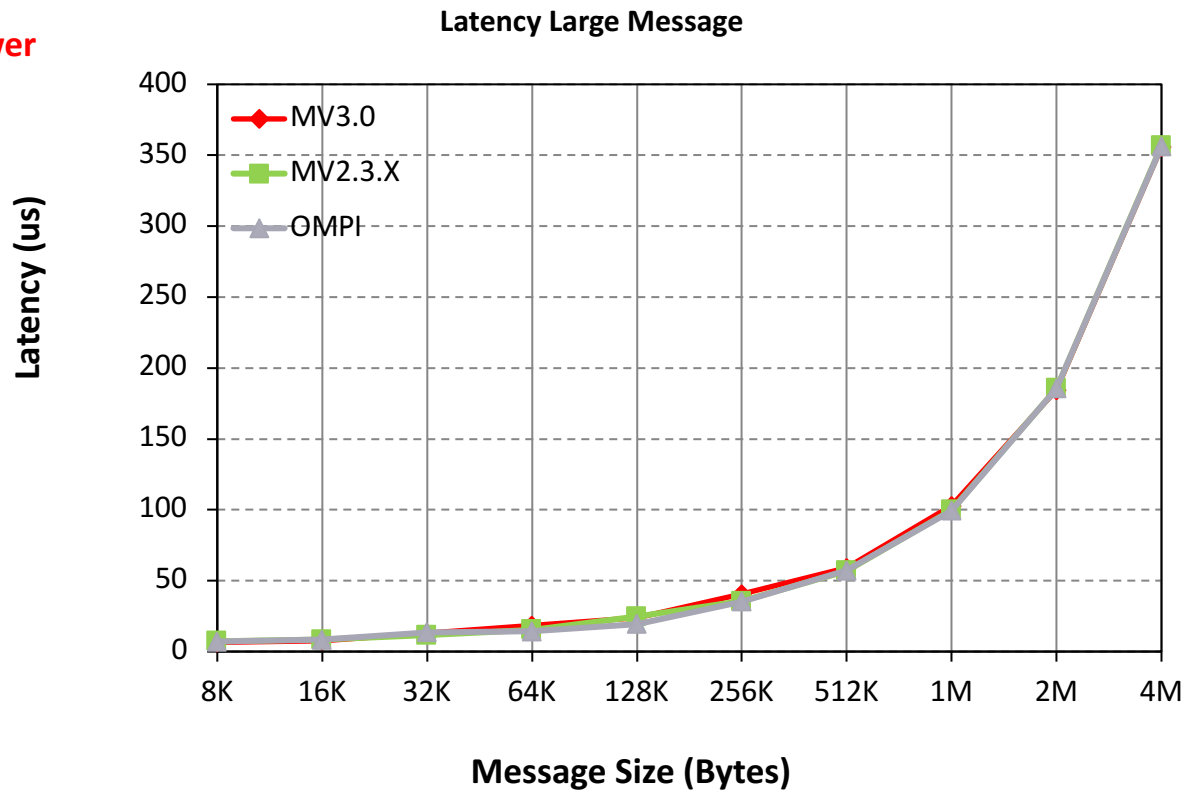
- Reduce up to 15% execution time of CP2K H2O-dft-ls (NREP4)
- Reduce up to 7% execution time of WRF CONUS 3KM

Overview

- Introduction
- Performance Characterization
- Latency and Message Rate Optimization
- Performance Evaluation
  - Micro-benchmark level
  - Application level
- **MVAPICH 3.0 Performance Evaluation**

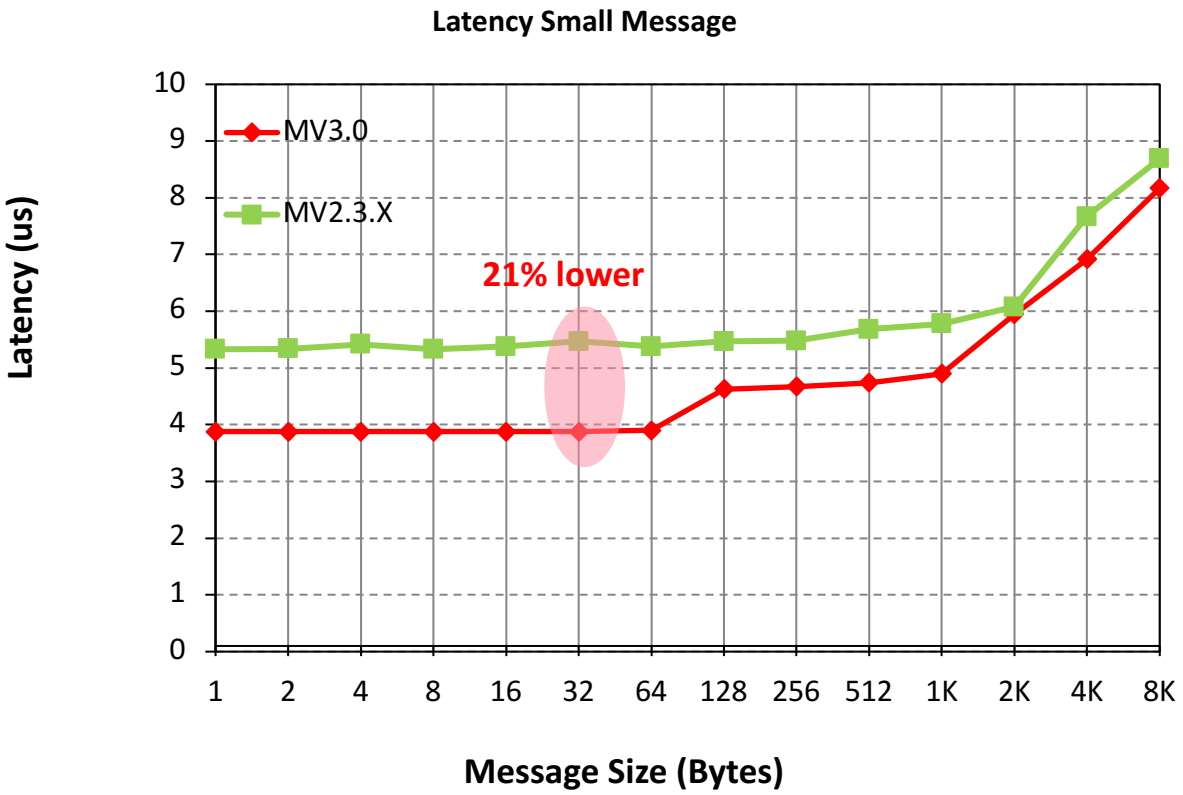
MVAPICH-3.0 Pt-to-Pt Latency (RC) on FW 227 (RHEL 8.8)

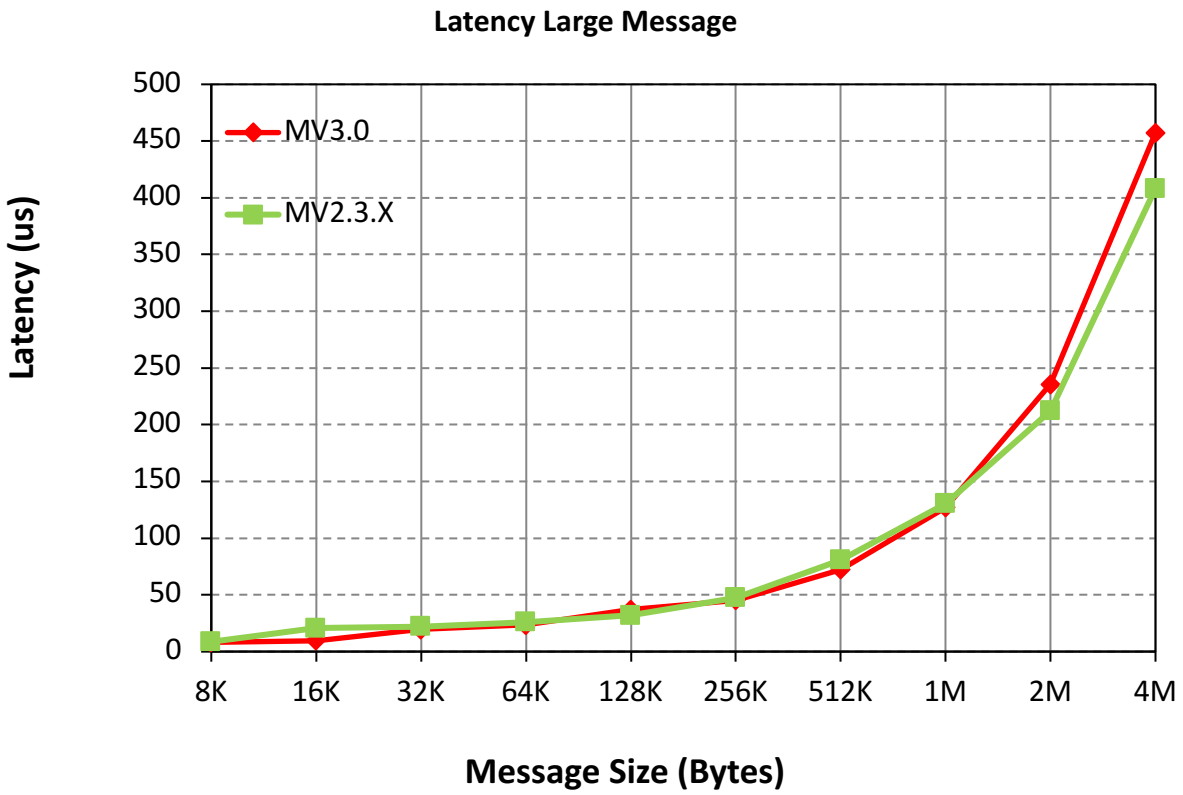




- MVAPICH 3.0 provides competitive point-to-point performance
- Reduce 9% latency with 16KB message size

MVAPICH-3.0 Pt-to-Pt Latency (UD) on FW 227 (RHEL 8.8)





- MVAPICH 3.0 provides competitive point-to-point UD performance
- Reduce 28% latency with 16KB message size

Conclusion & Future Work

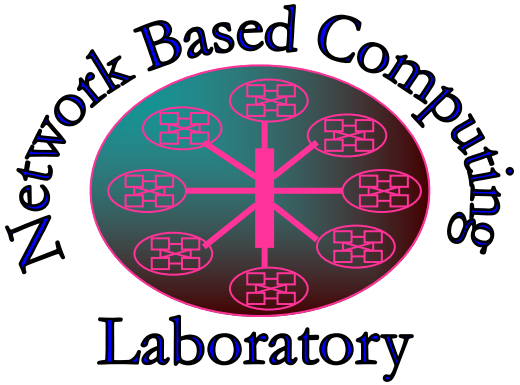
- Conclusion:
  - We have analyzed MPI overheads vs. IB level performance on Broadcom adapter
  - We have tuned MVAPICH2 1) coalescing, 2) SGL eager usage, 3) binding policies, 4) UD start-up, 5) UD/RC thresholds and 6) collective algorithms for Broadcom Thor families
  - The bottom-up approach targeting microbenchmark latency and message rate resulted in significant microbenchmark and application-level gains
- Future Work:
  - Optimize additional applications
  - Integrate existing optimizations with MVAPICH-3.0 on Broadcom systems
  - In progress: MVAPICH-2.3.8 (with enhanced RoCEv2 support)
  - Optimize MVAPICH for UltraEthernet

Ultra Ethernet Consortium - Modernizing RDMA

Classic RDMA	Ultra Ethernet
In-order packet delivery	**Out-of-order placement, in-order message completion**
Go-back-n ➔ inefficient	**Selective Ack and retransmit**
No multipathing	**Packet-level multipathing**
DCQCN ➔ hard to tune	**Scalable and Simplified Congestion control**

Higher fabric utilization at ultra-high scale with automated config and tuning

THANK YOU!



Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/