MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiBD
High-Performance
Big Data

HiDL
High-Performance
Deep Learning

High Performance & Scalable MPI library over Broadcom RoCE

Presentation at the 11th Annual MVAPICH User Group (MUG) Conference (MUG '23)

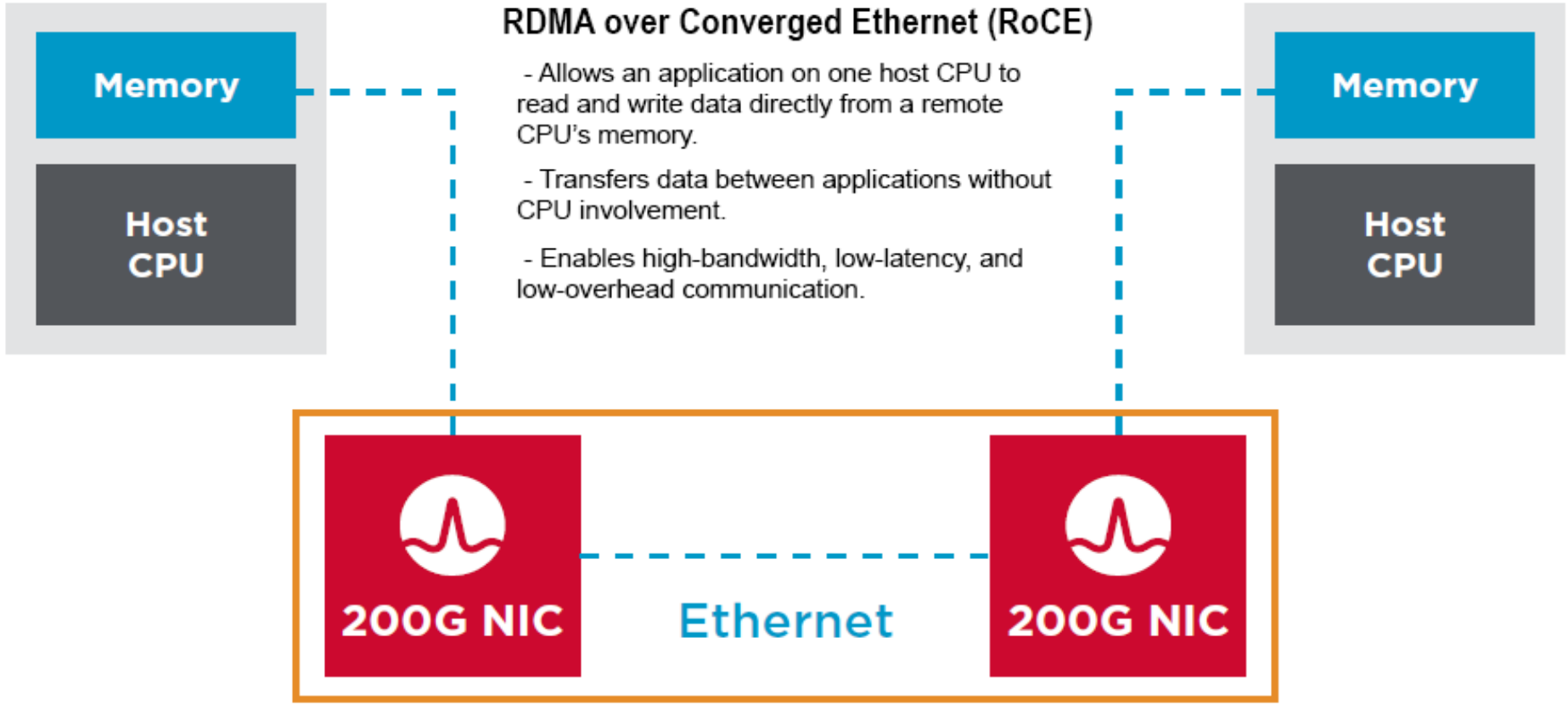
Presenter: Shulei Xu

xu.2452@osu.edu
The Ohio State University

Follow us on

https://twitter.com/mvapich

Introduction



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

Motivation & Problem Statement

- Broadcom adapter support RDMA over Converged Ethernet (RoCE)
- MVAPICH2 2.3.x series has been supporting RoCE for many years
- **Goal:** Optimize MVAPICH2 (2.3.7 release) through a funded collaboration with Broadcom
- Evaluate and study the performance improvements of the optimizations

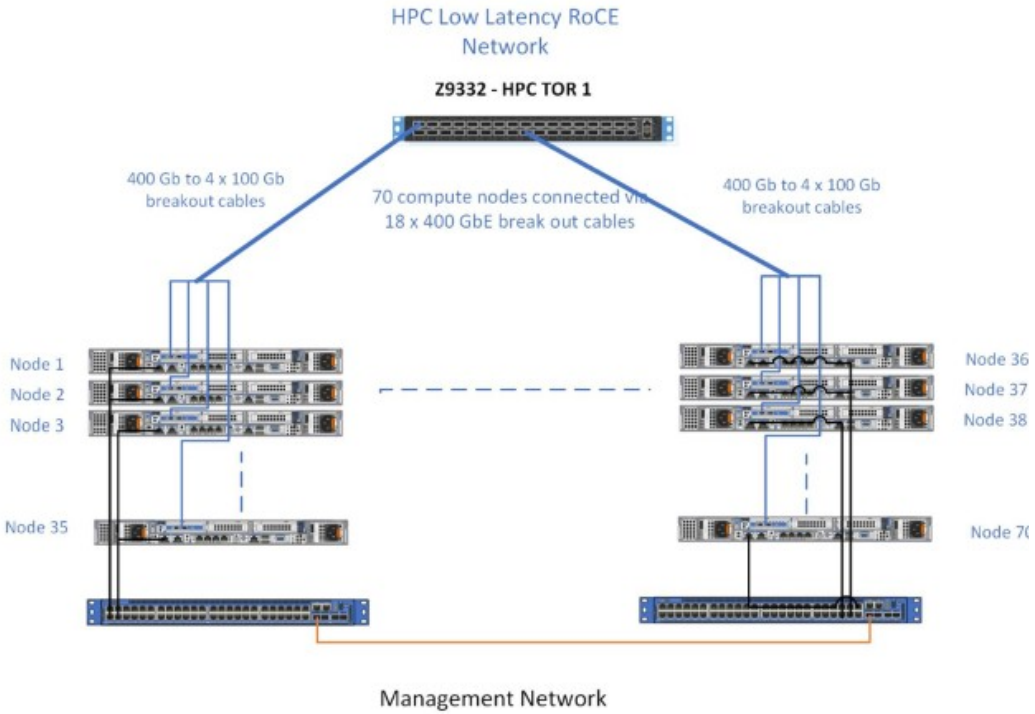
Configuration & Middleware Setup

- 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 1.14.0:
  - `./configure --prefix=<UCX_INSTALL_PATH>`
- OpenMPI 4.1.5 (w/ UCX 1.14.0):
  - `./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

CPU and NIC Setup

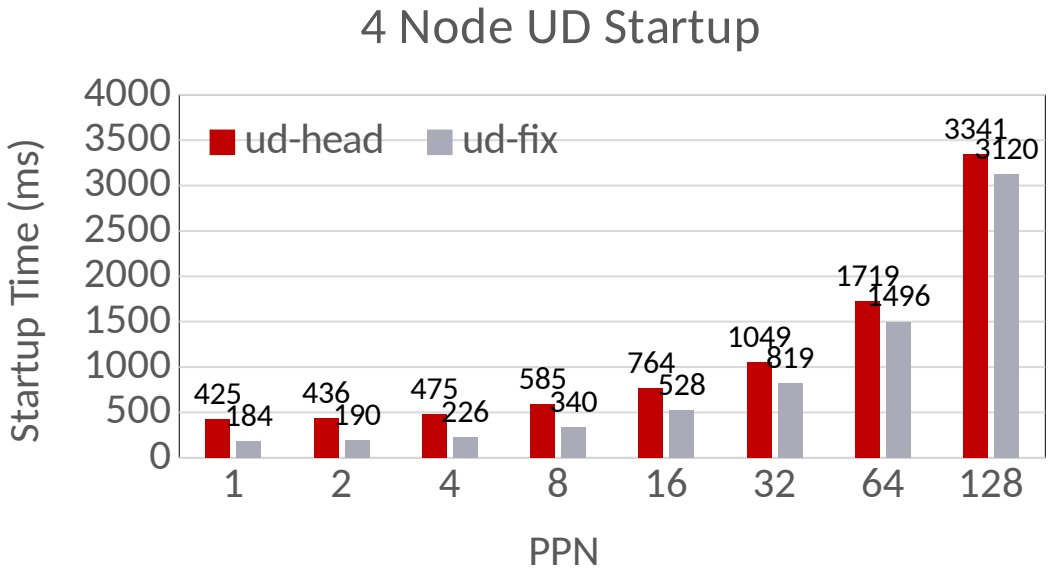
Feature	Specification
Model	AMD EPYC 7713
Cores	64
Socket	2
Base clock speed	2.9 GHz
Boost clock speed	3.6 GHz
L1 cache	32 KB per core
L2 cache	512 KB per core
L3 cache	32 MB per core
Memory support	Up to 32 TB of DDR4-3200
PCIe lanes	128
TDP	280W

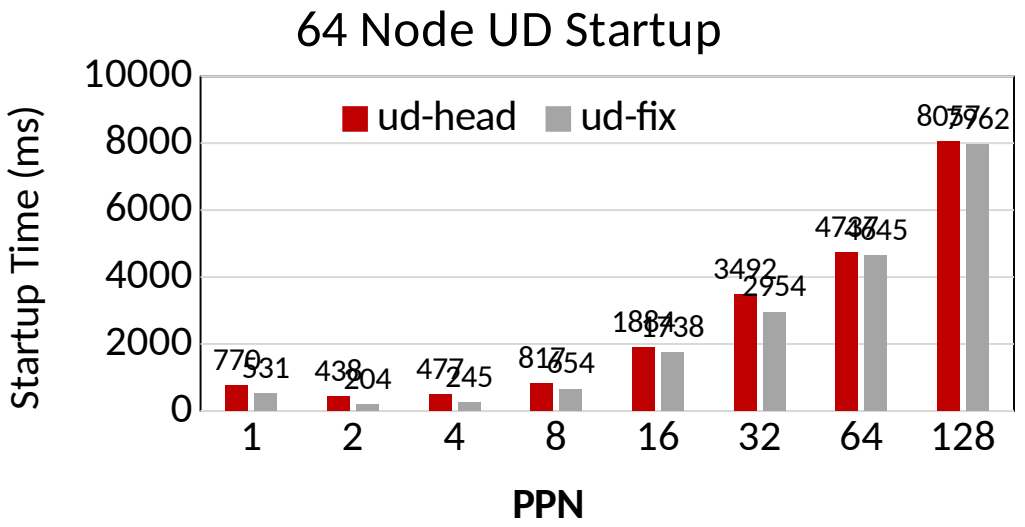
Feature	Specification
Model	Broadcom Thor RoCE HCA
Driver	bnxt_en
Driver Version	1.10.2-226.0.141.0
Firmware Version	226.0.145.1/pkg 22.61.10.71
Number of ports	1
Speed	100 Gb/s
Supported link modes	25000baseCR/Full 50000baseCR2/Full 100000baseCR4/Full 50000baseCR/Full 100000baseCR2/Full 200000baseCR4/Full
Active link mode	100000baseCR4/Full

Performance Optimizations

- High-level description of optimization efforts:
  - Enhanced point-to-point & collective tuning
  - Enhanced UD+RC hybrid transport mode tuned for Broadcom ROCE adapter
  - Optimized default CPU mapping policy
    - Make hybrid spread CPU mapping policy as default
  - Added support for asynchronous threading progress
  - Added UD Startup Optimization

UD Startup Optimization

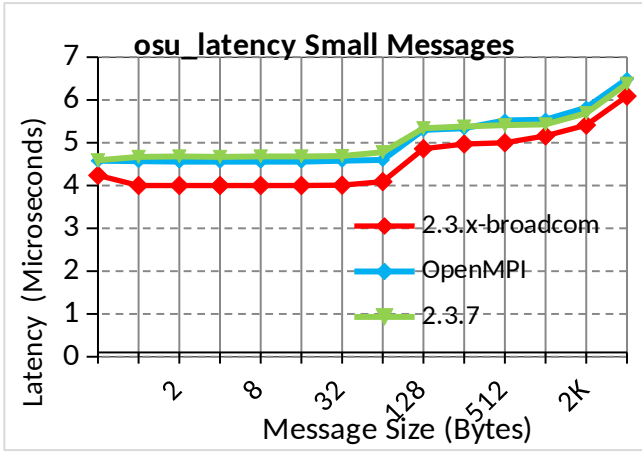


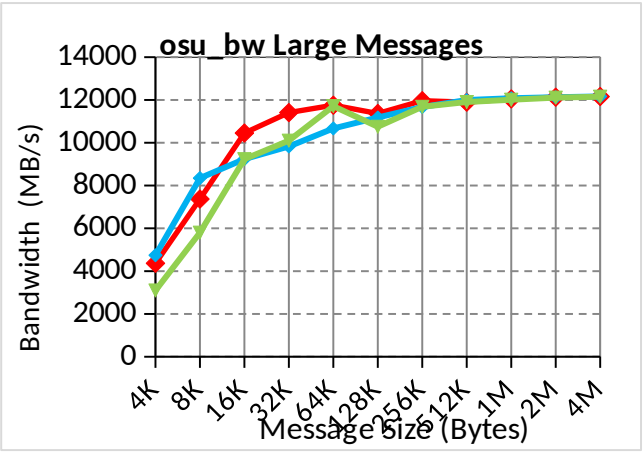


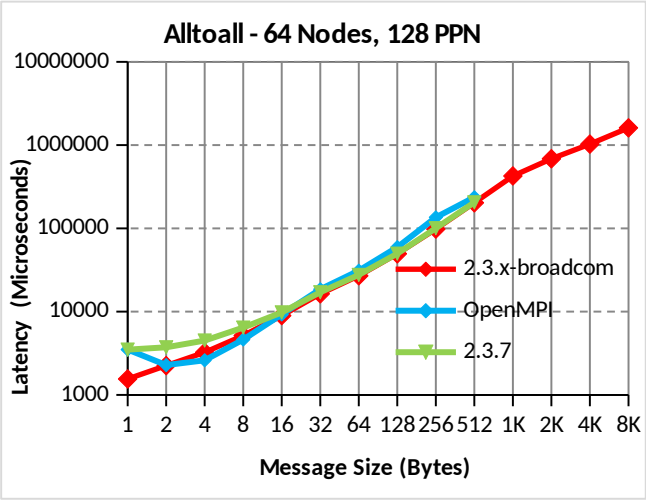
- Provide up to 2.3x faster UD startup in small 4 nodes scale
- Provide up to 2.1x faster UD startup in large 64 nodes scale

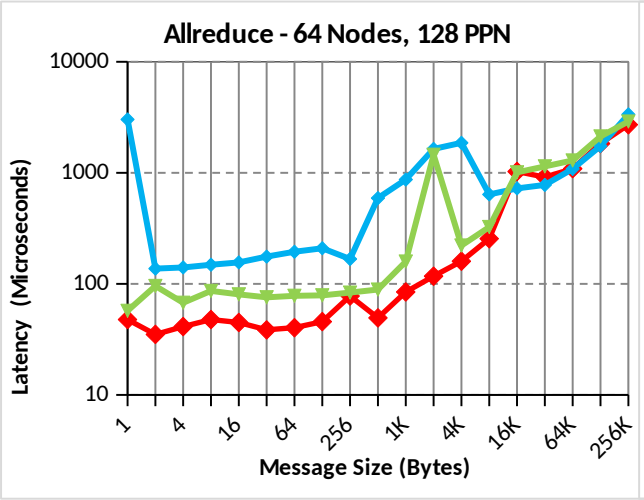
Performance Evaluation – Micro-benchmarks

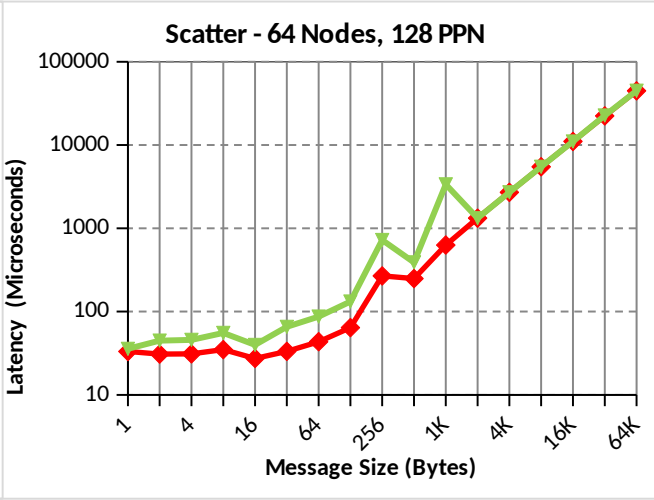
- Experimental results from Dell Bluebonnet
- Up to 20% reduction in small message point-to-point latency
- From 0.1x to 2x increase in bandwidth
- Up to 12.4x lower MPI_Allreduce latency
- Up to 5x lower MPI_Scatter latency



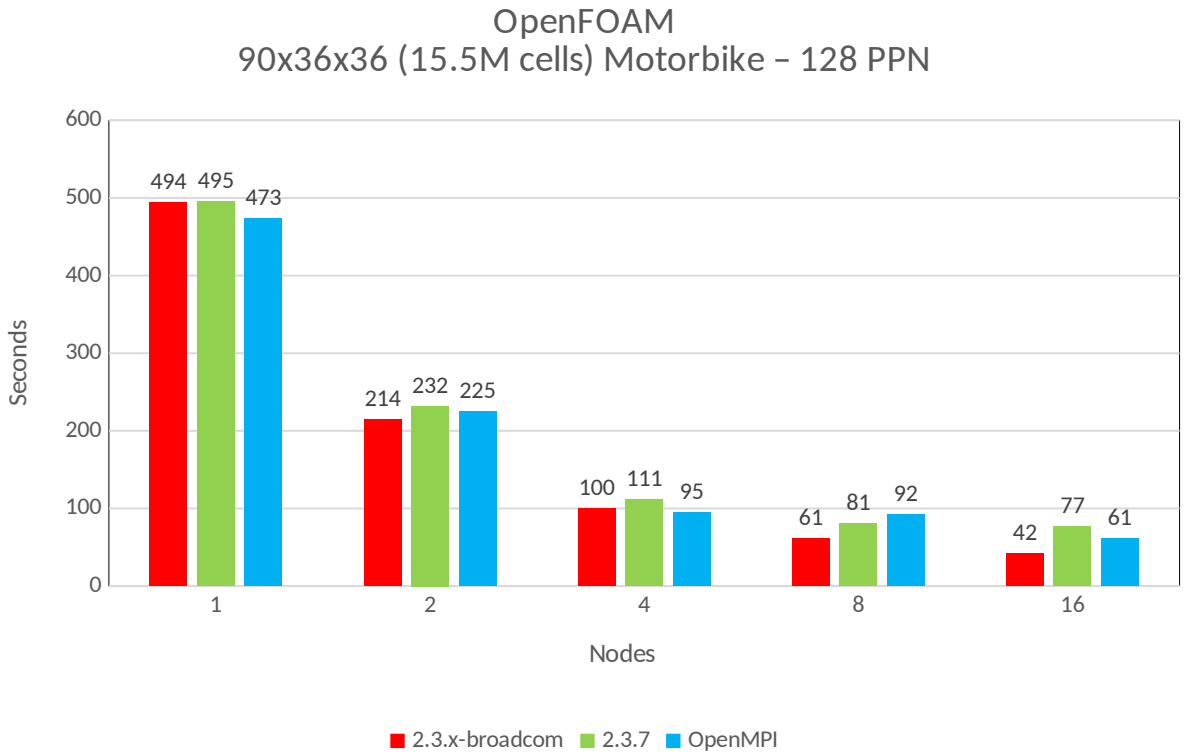


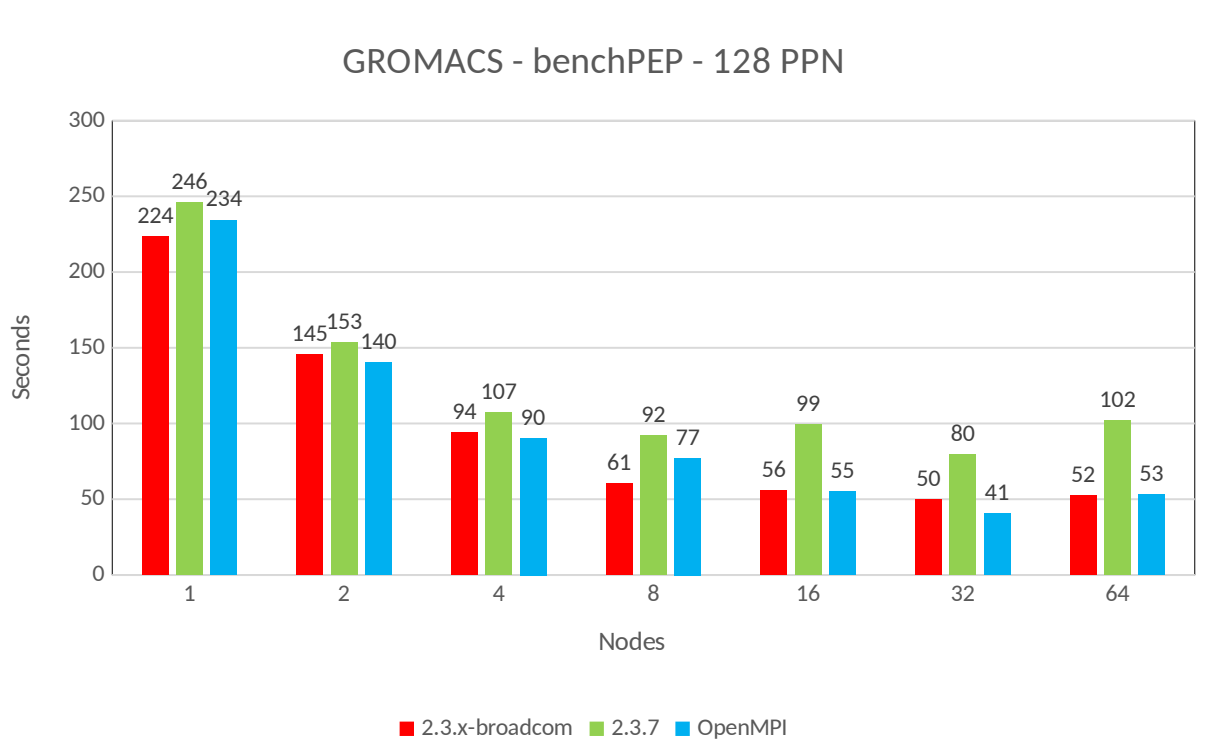






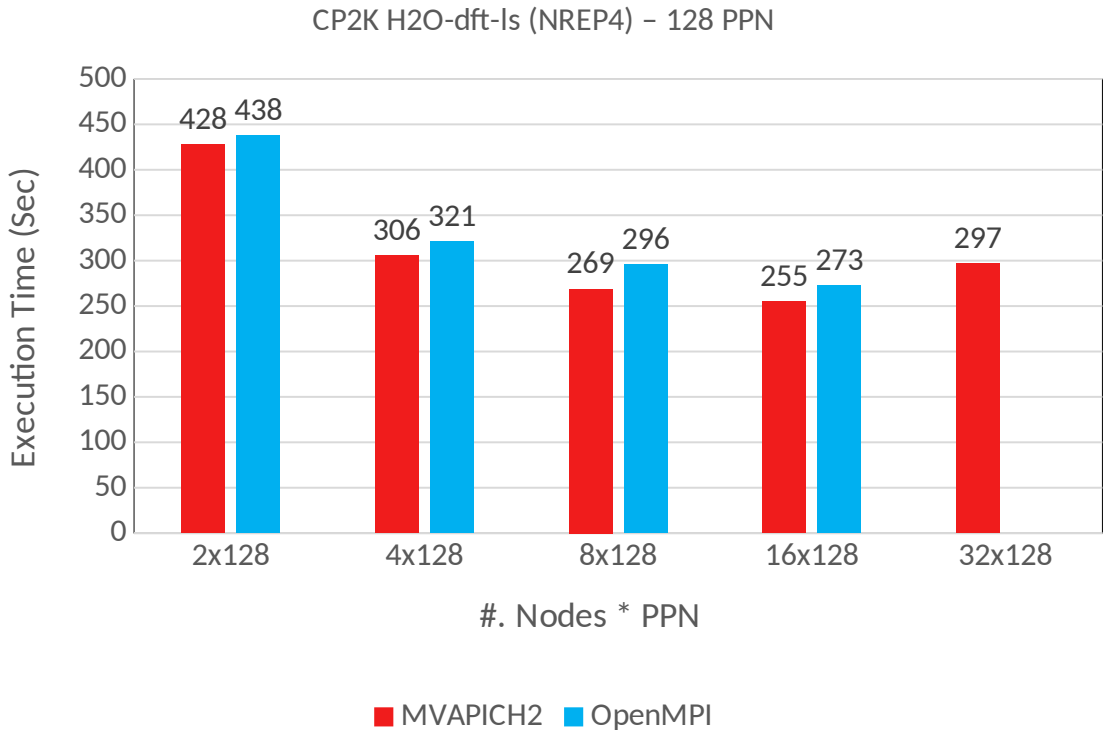
Performance Evaluation – Applications

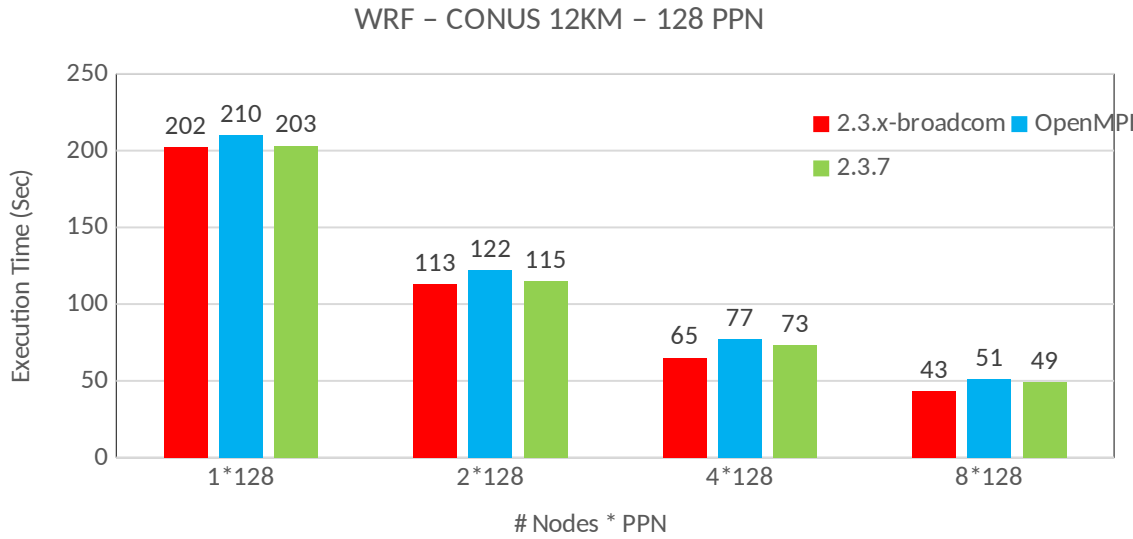


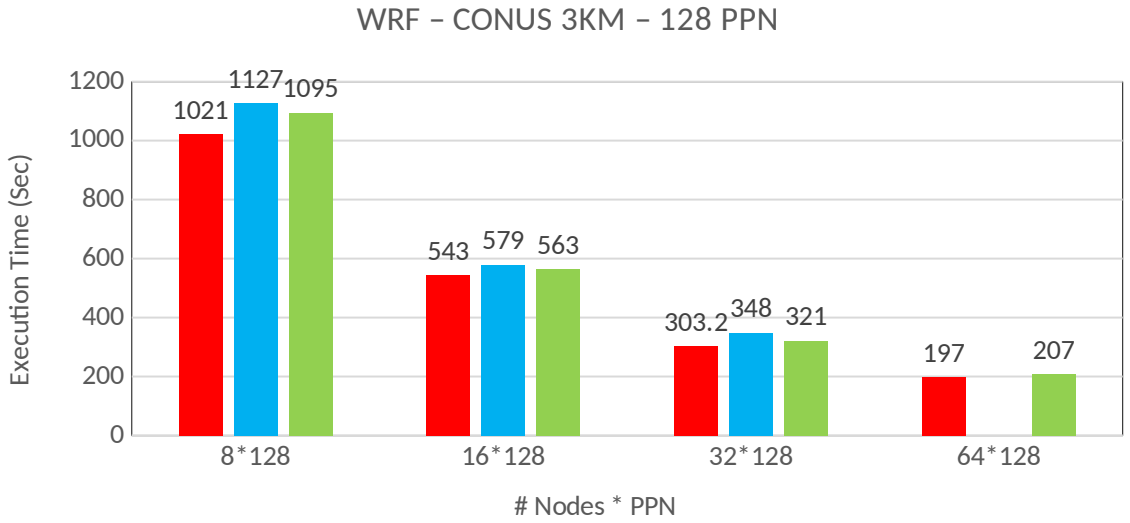


- Reduce up to 45% execution time of OpenFOAM Motorbike on 16 nodes 128 PPN scale
- Reduce up to 51% execution time of GROMACS benchPEP on 64 nodes 128 PPN scale

Performance Evaluation – Applications





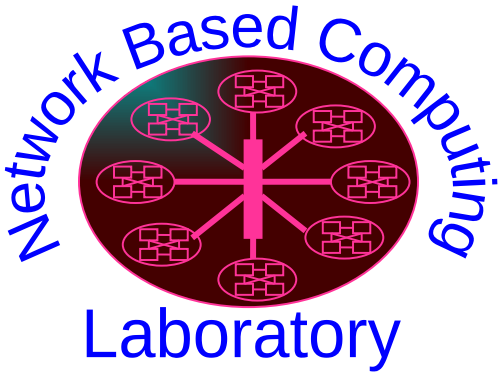


- Reduce up to 45% execution time of CP2K H2O-dft-ls (NREP4)
- Reduce up to 7% execution time of WRF CONUS 3KM

Conclusion & Future Work

- Conclusion:
  - Optimize MVAPICH2 MPI library performance on the latest Broadcom ROCE adapter
  - Evaluate performance improvements
    - Up to 12x lower MPI_Allreduce latency on micro-benchmark level
    - Up to 51% reduction in execution time of GROMACS HPC application
- Future Work:
  - Add larger scale (up to 64 nodes) collective optimizations for both CPU & GPU nodes with Broadcom adapter
  - Optimize additional applications
  - Integrate existing optimizations with MVAPICH-3.0 on Broadcom systems

THANK YOU!



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

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

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

HiBD
High-Performance Big Data

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

HiDL
High-Performance Deep Learning

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