MVAPICH2 on Thor: High Performance MPI Meets Mainstream Ethernet Controller

Hemal Shah, Moshe Voloshin, and Devesh Sharma

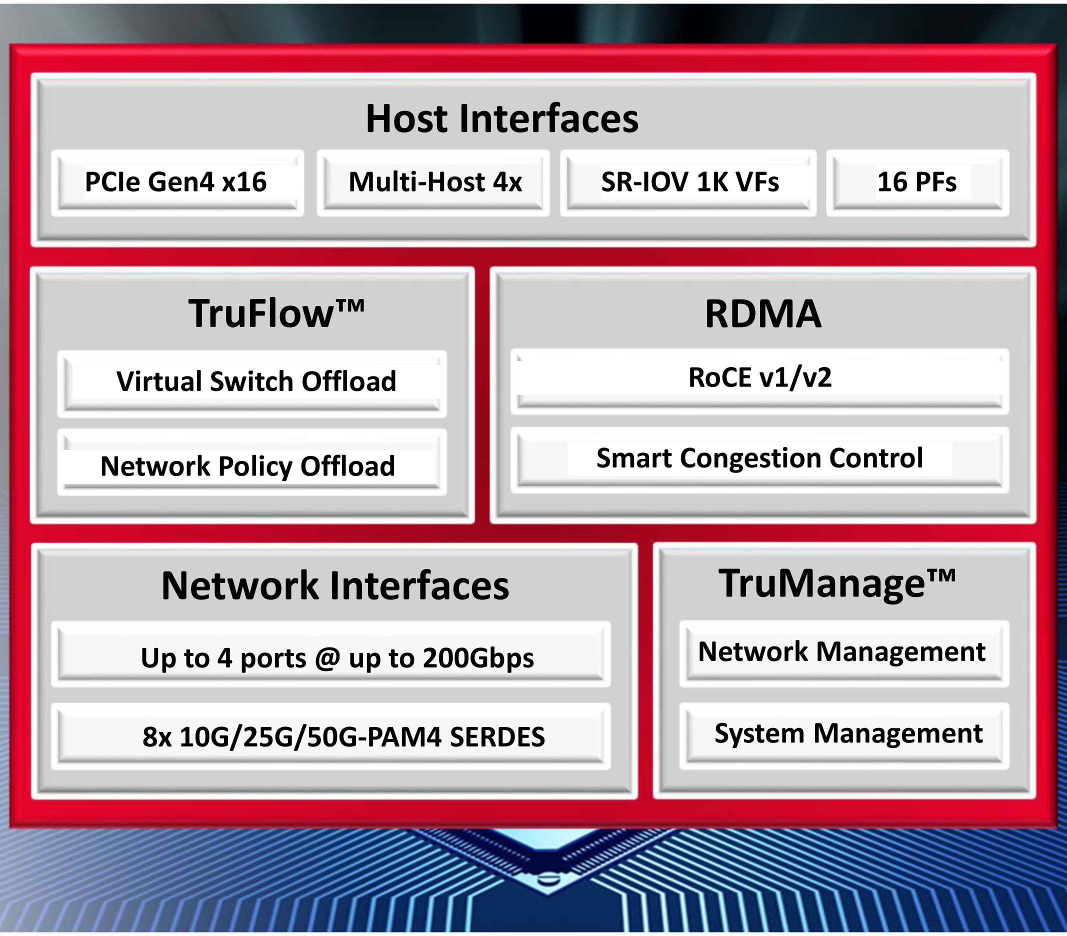
August 25, 2020

Agenda

- Thor Overview
- Thor RoCE Features
- Thor RoCE Firmware/Software Architecture
- MVAPICH2 on Thor
- MPI Test Results

Thor: High Performance Ethernet Controller

- **Performance**
  - 200Gbps Throughput & 100Mpps
  - E2E Latency (TX + RX) < 1 usec
- **Host Interface**
  - PCIe Gen4 x16
  - Multi-Host up to 4 End Points
  - 16 PFs and 1K VFs
- **Network Interface**
  - Octal 50Gbps PAM4 SERDES
  - Quad Ports, High Availability
- **RDMA**
  - RoCEv2
  - Smart Congestion Control
  - GPUdirect



Host Interfaces

PCIe Gen4 x16

Multi-Host 4x

SR-IOV 1K VFs

16 PFs

TruFlow™

Virtual Switch Offload

Network Policy Offload

RDMA

RoCE v1/v2

Smart Congestion Control

Network Interfaces

Up to 4 ports @ up to 200Gbps

8x 10G/25G/50G-PAM4 SERDES

TruManage™

Network Management

System Management

Thor RoCE Hardware Features

Network Performance	
Interface Bandwidths	10/25/40/50/100/200 Gbps
Throughput	200 Gbps
Latency	< 1 usec (chip TX + RX)

Network Aspects	
RoCE framing	Concurrent v1 and v2 support
DCB	PFC, ETS
Partitioning	Up to 16 PFs enabled with RoCE
SR-IOV support	Up to 1K VFs enabled with RoCE
QoS	Hierarchical TX scheduling
Congestion Control	ECN/CNP (RoCEv2)

Connection Types	
RC	Supported
UD	Supported, including QP1
XRC	Not supported
Raw Eth QP	Supported (kernel bypass for L2 traffic)
UC, RD	Not supported

Memory Management & Protection	
Regions & Windows	1 Million
Region/Window Size	1GB → 1TB
MW Types	1, 2A, 2B
Page Sizes	4KB, 8KB, 64KB, 256KB, 1MB, 2MB, 4MB, 1GB
Doorbell Page Sizes	4KB, 8KB, 64KB, 256KB, 1MB, 4MB
Protection Domains	1 Million
Fast Memory Register	Supported

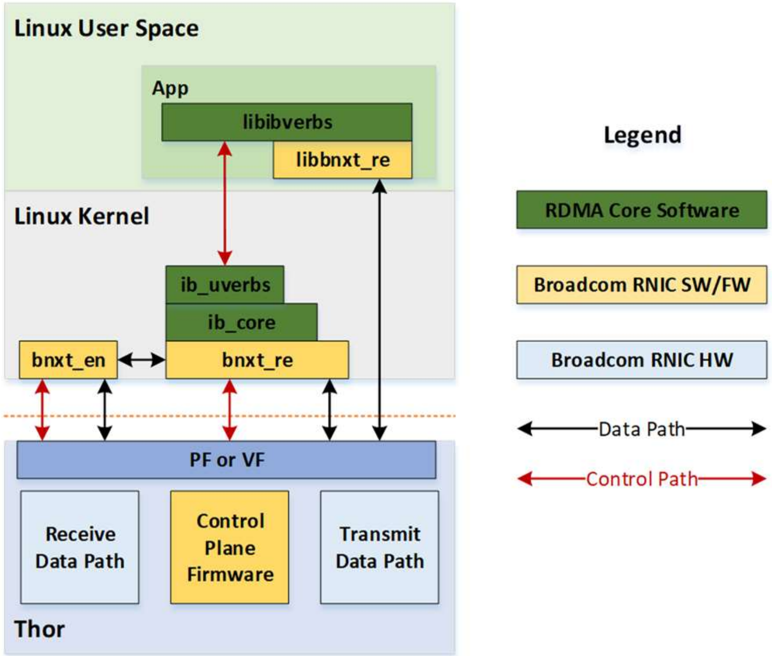
Scale	
QPs for RC	1 Million (up to 32K WQEs per SQ/RQ)
Doorbell Pages or DPIs	64K
SRQs	64K
CQs	1 Million
Scatter Gather List per WQE	Up to 30 SGEs

Miscellaneous	
CQ Resize	Supported
PCIe Bandwidth	Gen4 x16
Retransmission & Duplicate handling	Supported
SR-IOV & RoCE	Supported - Both PF/VF RDMA simultaneously
Embedded RoCE processor	Manage RoCE resources/state/exceptions

Thor – RoCE Advanced Features

Features	Notes
Deterministic Marking	ECN based marking
Probabilistic Marking	ECN based marking
Congestion Control	Scales with the number of competing flows
WQE Caching • **SQ, RQ, and RDMA Read WQEs**	Improves latency and overall performance
Variable Size MTU	Flexibility
Number of DPIs	Higher App Scaling
Asymmetric PFC	PFC enhancement for switches
PFC Watchdog	Hardware support for PFC watchdog
Shared PD	New Feature for Application Scaling
Large Shared MR	New Feature for Application Scaling
CoS & RoCE counters	New Counters for monitoring RoCE traffic and CoSQs

Thor RoCE Software for Linux



Supports Host OS, VM, or Container

Same Driver for PF or VF

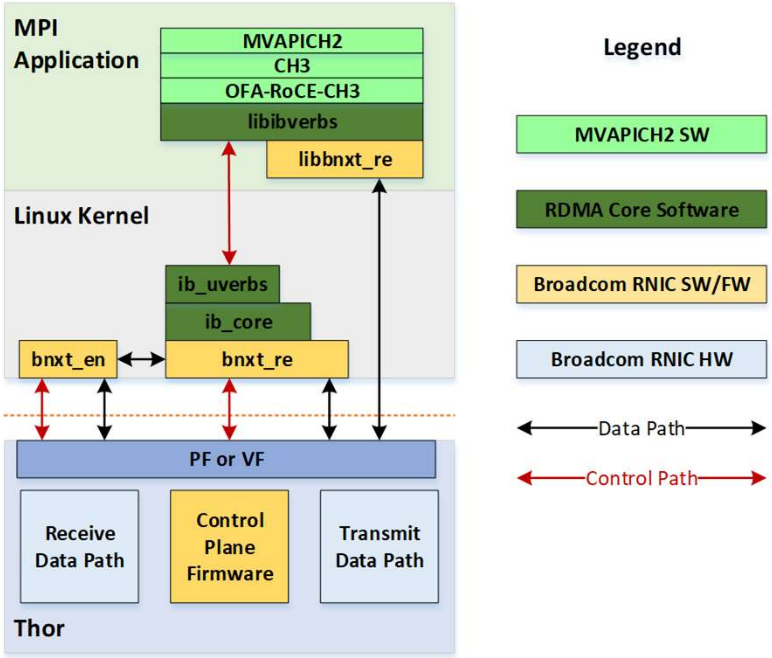
Enables open source RDMA stack

- Aligned with Linux kernel
- Kernel modules are upstream to kernel
- User libraries are submitted to OFED/linux-rdma

Broadcom Linux RDMA Components

- RoCE User Library (libbnxt_re)
- RoCE driver (bnxt_re)
- NIC driver (bnxt_en)
- RDMA Control Plane Firmware

MVAPICH2 over Thor



Builds on Standards Verbs stack

Enables MVAPICH2 over Verbs Provider

Unmodified MPI applications

No proprietary extensions

Thor Software/Firmware Components

- RoCE User Library (libbnxt_re)
- RoCE driver (bnxt_re)
- NIC driver (bnxt_en)
- RDMA Control Plane Firmware

Initial MPI Testing on Thor

- Focus is on a set of HPC applications
- Not attempting to compare two MPI implementations
- Small scale testing this time due to cluster scale availability
- Plan to have detailed MPI performance results in the future

Verbs Level Performance - Baseline

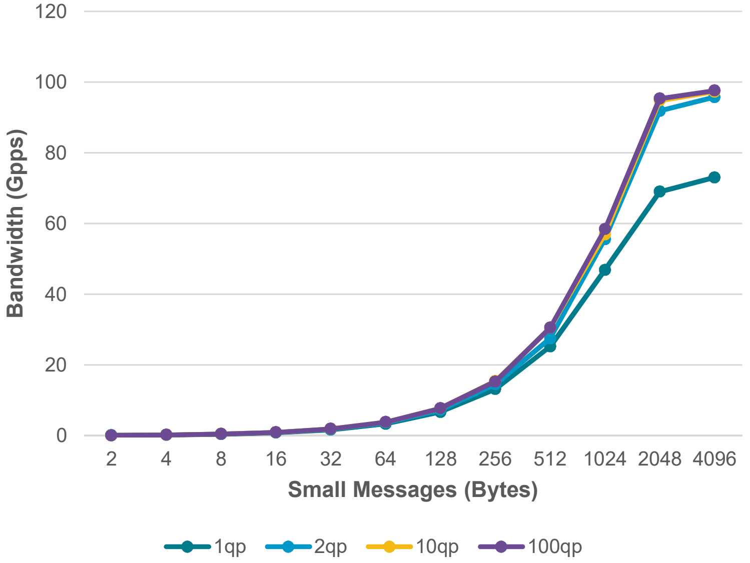
System setup

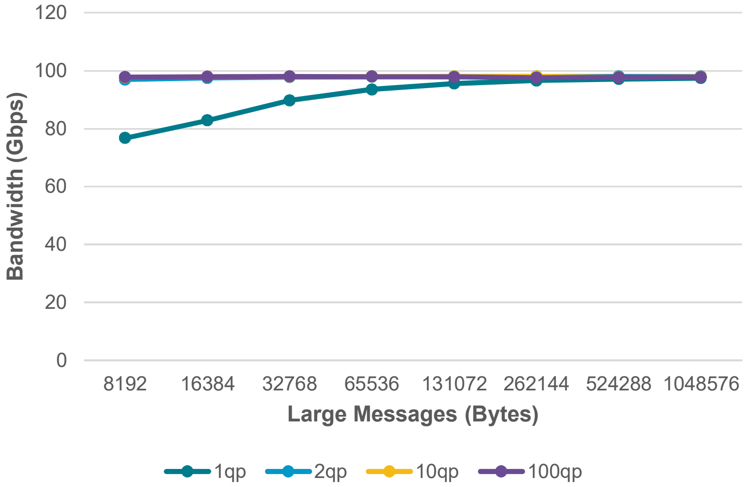
- CPU:Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz
- Cores: 2 Sockets, 32 cores
- 2 hosts connected back to back
- MTU set to 4K and 1K (Latency only)

Line rate BW for large messages for 2+ QPs

Small Message (2B) ½ RT latency is agnostic of MTU

RDMA Operation	1K MTU (usec)	4K MTU (usec)
RDMA-Write	1.93	1.93
Send/Recv	2.12	2.12
RDMA-Read	4.42	4.42





OSU Microbenchmark Point to Point

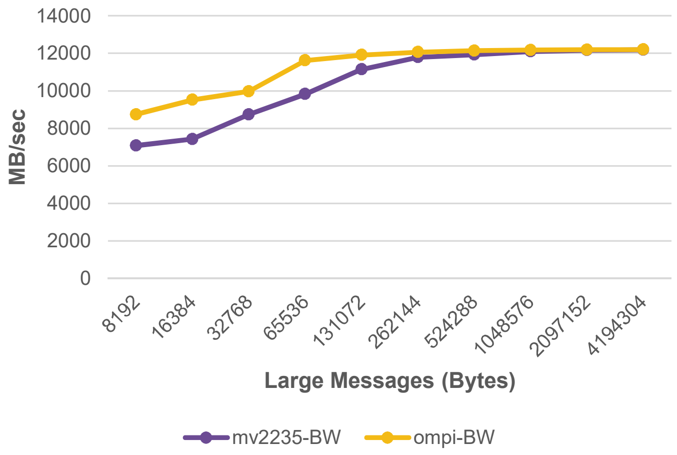
System setup

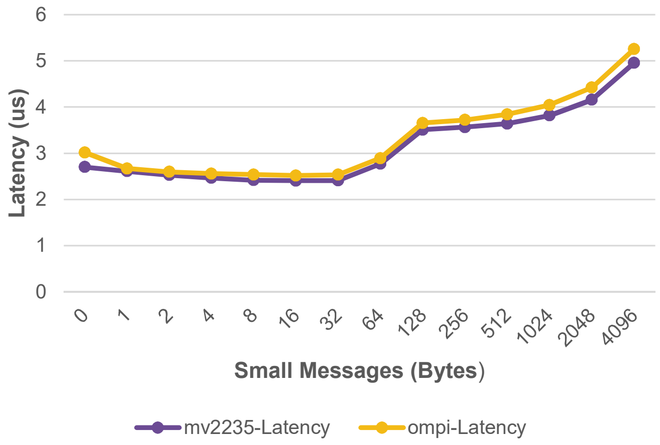
- 2 nodes connected to switch and configured for PFC
- 1 Thor 2x100G adapter per host
- CPU:Intel(R) Xeon(R) Gold 5218 @ 2.30GHz
- Cores: 2 Socket, 32 cores
- Mvapich2-2.3.5-pre, Openmpi-4.0.3 with ucx-1.8.1
- Osu-Benchmarks-5.6.3

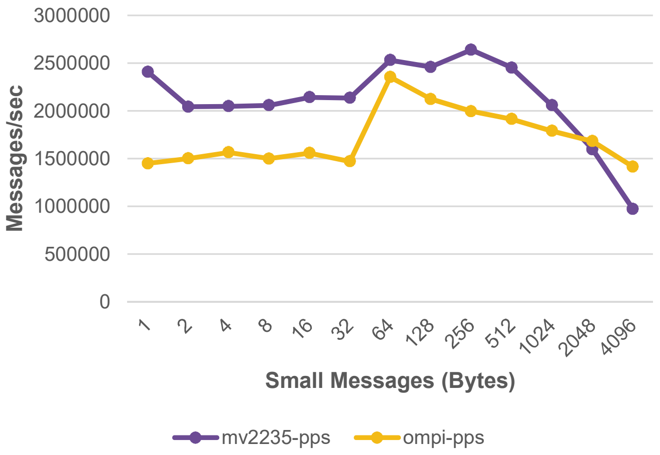
Tests

- Unidirectional bandwidth
- Latency and Message rate (1 pair)
- Bidirectional bandwidth

OSU Microbenchmark pt2pt: Uni-directional







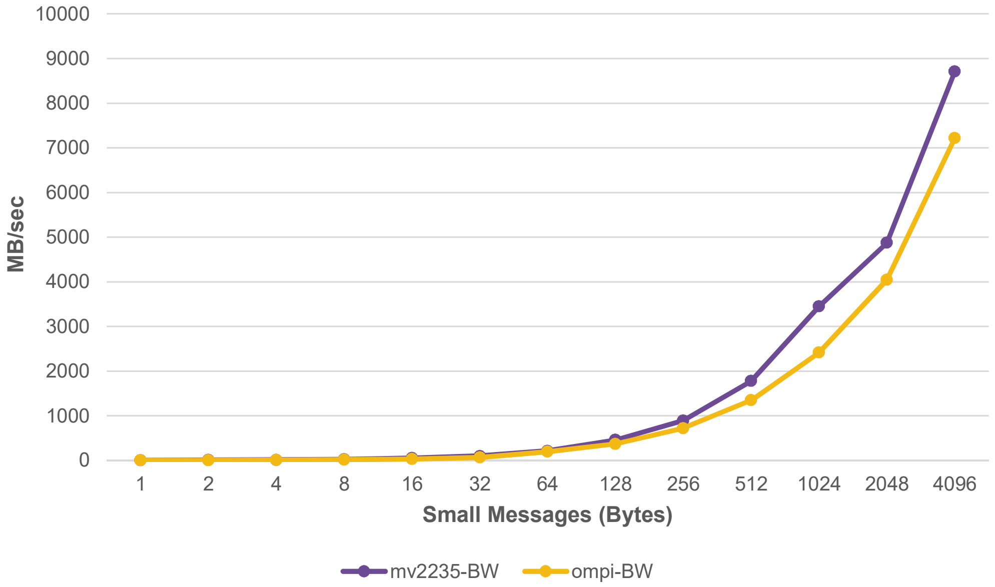
Small message latency with MVAPICH2 is slightly better than openmpi

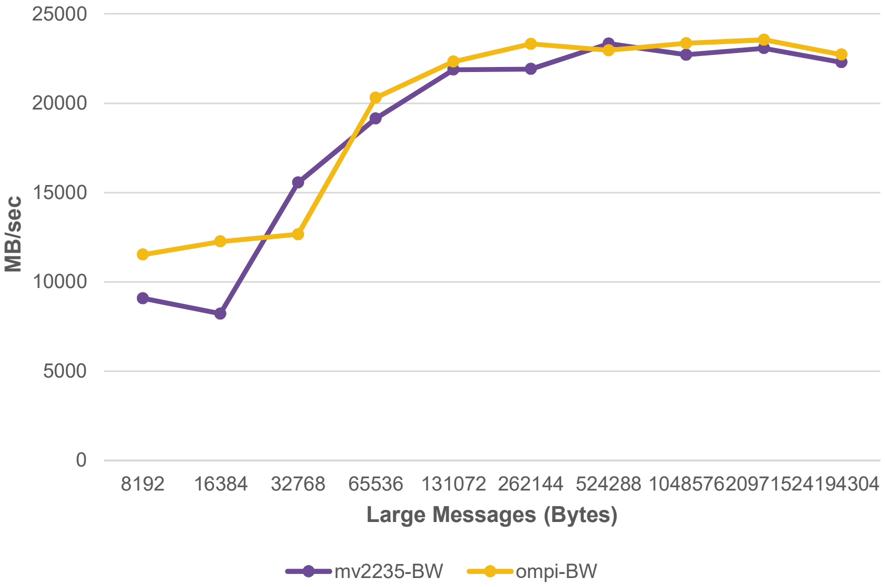
Small Message (1B to 1KB) rate with MVAPICH2 is ~40% more than openmpi

MVAPICH2 and openmpi both achieve line rate with Thor for large message sizes

Thor performs equally well for both MVAPICH2 and openmpi

OSU Microbenchmark pt2pt: Bi-directional





Large message Bandwidth is ~wire speed with both MPI implementations

OSU Microbenchmark Collectives

System setup:

- 8 nodes connected to switch and configured for PFC
- 1 Thor 2x100G adapter per host
- CPU:Intel(R) Xeon(R) Gold 5218 @ 2.30GHz
- Cores: 2 Socket, 32 cores
- Mvapich2-2.3.5-pre
- Osu-Benchmarks-5.6.3

Non-Reduce Operations (Scatter and Gather)

Reduce Operation (All_reduce)

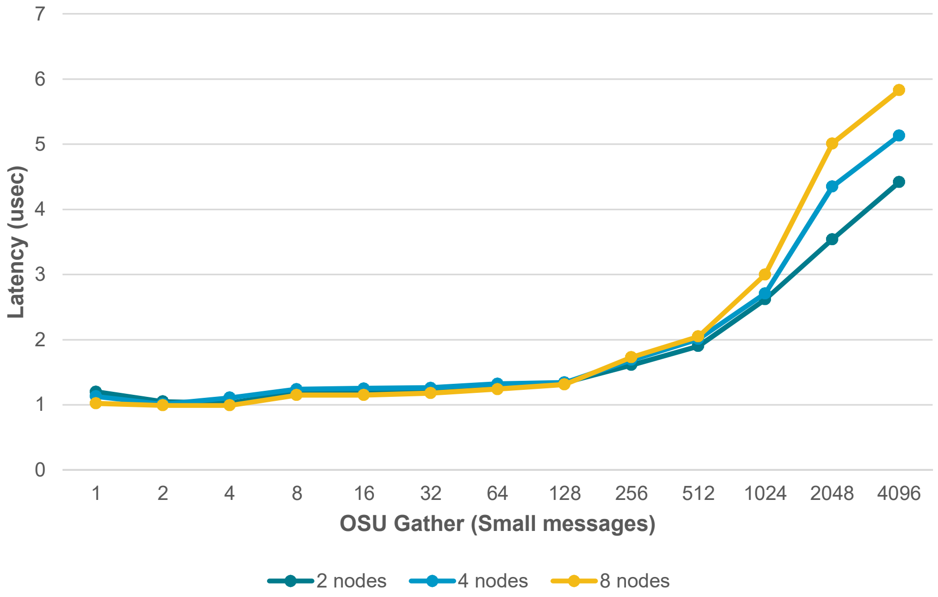
2,4,8 nodes

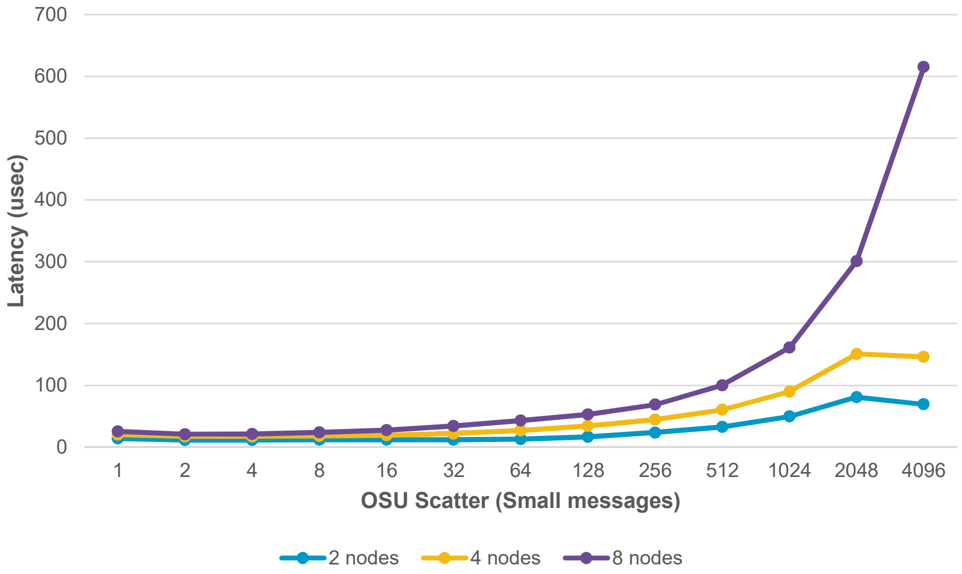
- 128, 256, 512 total processes

MVAPICH2 optimizations used in testing

- MV2_CPU_BINDING_POLICY = hybrid
- MV2_HYBRID_BINDING_POLICY = linear

OSU Microbenchmark Collectives: Gather and Scatter



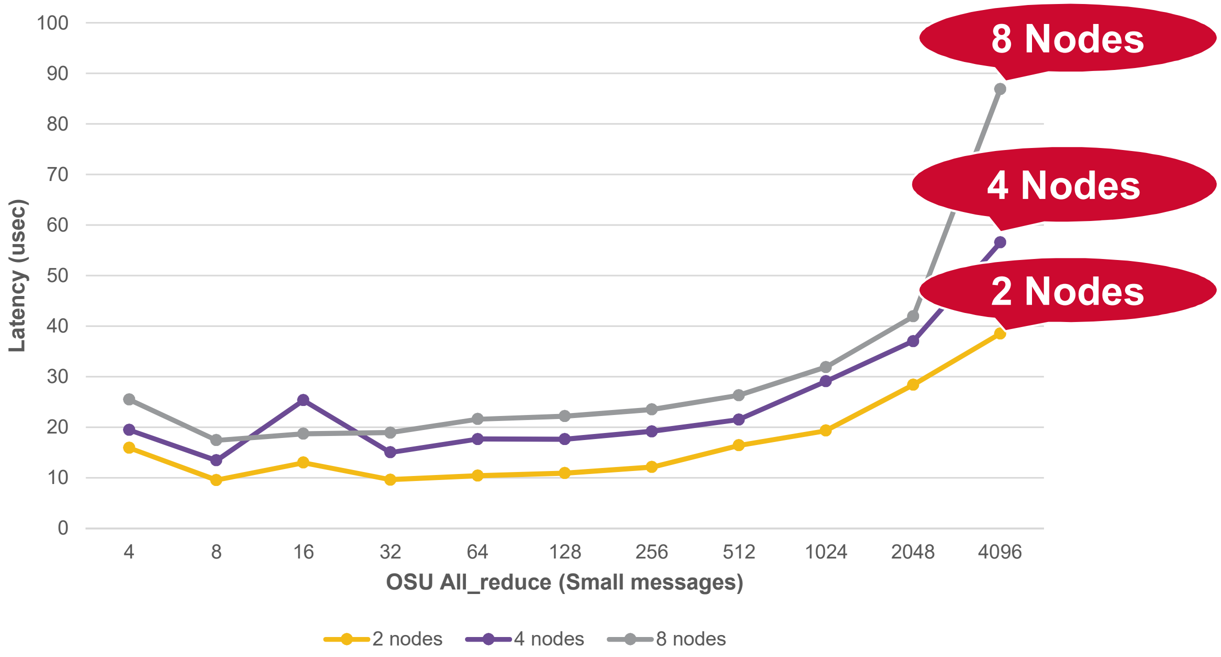


One to many communication with Thor (Scatter) scales well

- Scales really well for small message sizes

Many to one communications with Thor (Gather) equally scales well

OSU Microbenchmark Collectives: osu_allreduce



Thor performs well for Reduce operations

Reduce operation scales well for small messages

- latency slightly up for 8 nodes 4KB, 512 process

HPC Applications

System setup

- 8 nodes connected to switch and configured for PFC, 1 Thor 2x100G adaptor per host
- CPU:Intel(R) Xeon(R) Gold 5218 @ 2.30GHz
- Cores: 2 Socket, 32 cores
- MPI: Mvapich2-2.3.5-pre, Openmpi-4.0.3 with ucx-1.8.1
- NPB-3.3.1, miniGhost and CloverLeaf tip of github

NAS benchmarks

- ClassD, LU,BT,SP,CG
- 8-nodes: LU,BT,SP, 484 process, CG 512 process
- 4-nodes: LU,BT,SP,CG 256 process
- 2-nodes: LU,BT,SP 121 process, CG 128 process

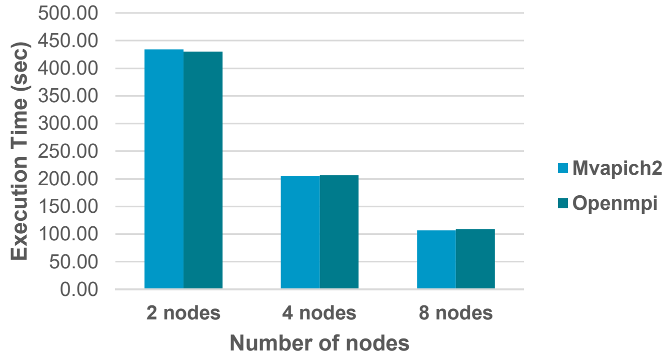
miniGhost

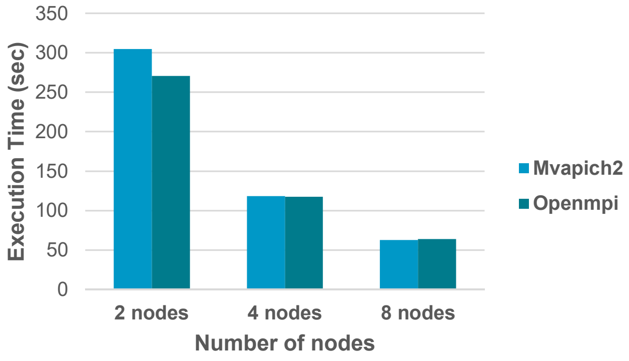
- 8-nodes, 512 process, 4-nodes 256 process, 2 nodes 128 process

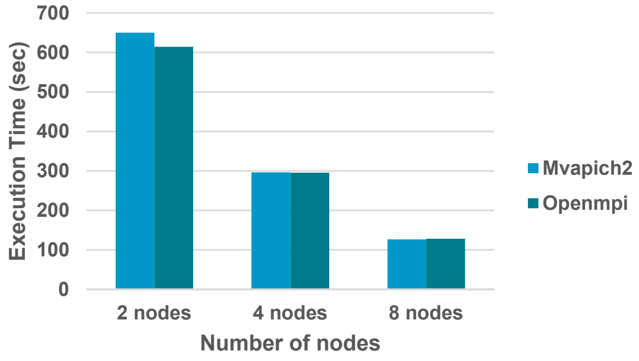
CloverLeaf

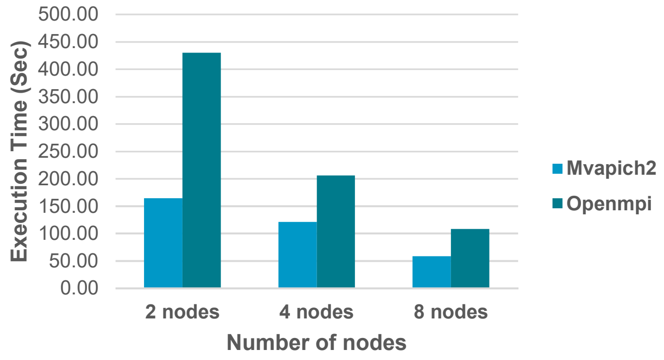
- 8-nodes, 512 process, 4-nodes 256 process, 2 nodes 128 process
- bm512_short

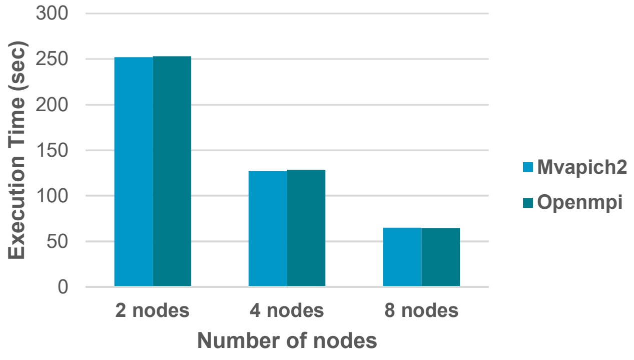
HPC Applications - Preliminary Results on Thor

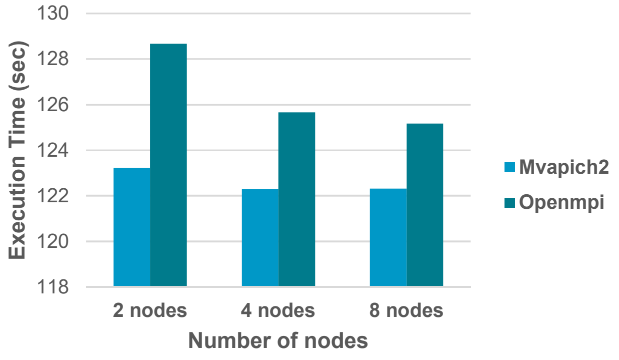

NAS-BT


NAS-LU


NAS-SP


NAS-CG


CloverLeaf


miniGhost

Thor Scales Well for HPC Applications at 100G Speed