MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

High Performance MPI over Slingshot

8/23/2022

Network-based Computing Laboratory
Department of Computer Science and Engineering
The Ohio State University

MUG 2022

Kawthar Shafie Khorassani
shafiekhorassani.1@osu.edu

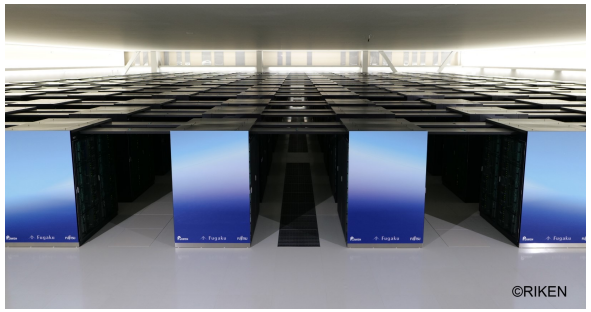
Introduction

- Frontier at OLCF (#1 Supercomputer on Top500 System) - deployed with Slingshot-11 networking across nodes
- MPI-level communication and performance on upcoming networking for exascale systems (i.e. Frontier & El-Capitan)

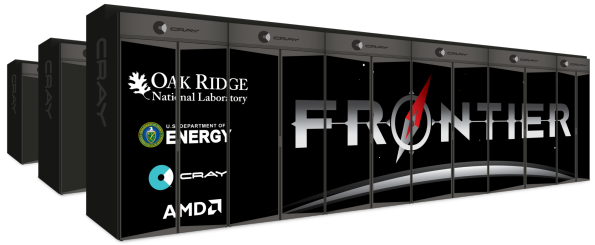
#1 Supercomputers Top500 Over Time …

Sunway TaihuLight (‘16-’17)

Summit (‘18-’19)



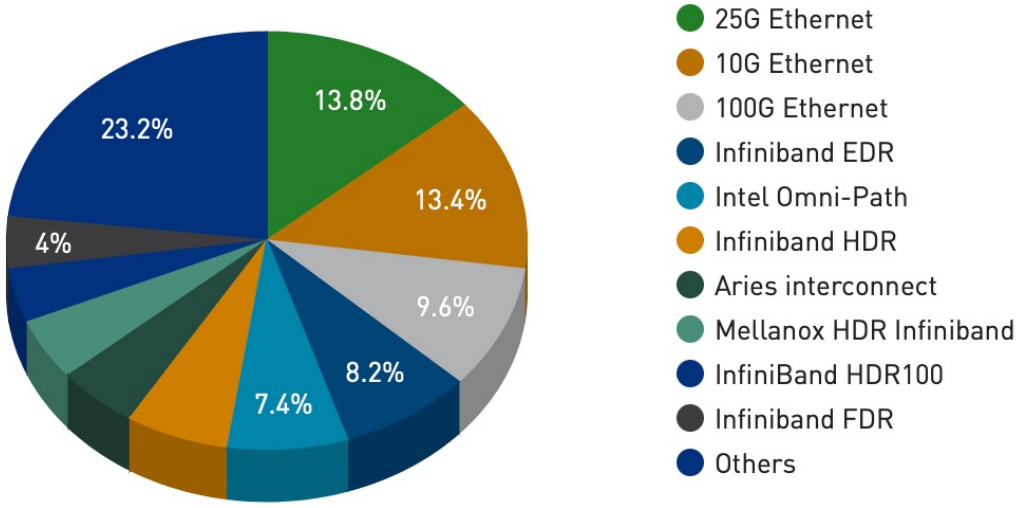
Fugaku (‘20-’21)



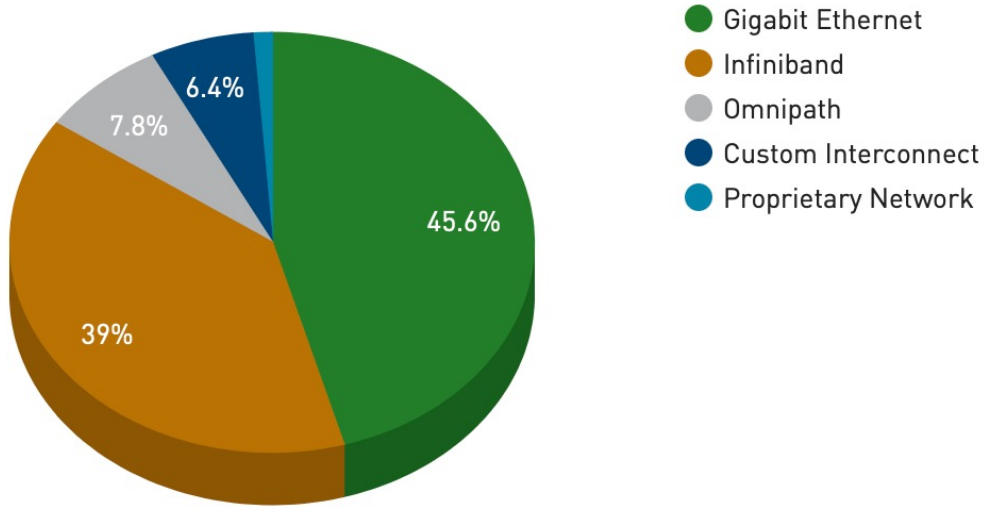
Frontier (‘22)

Top500 Supercomputers Interconnect Statistics

Interconnect System Share



Interconnect Family System Share



Reference: https://www.top500.org

Background

- Many Supercomputers deployed with Mellanox Infiniband Interconnect technology
- MPI Libraries have been optimized over the years to expand on Mellanox Infiniband features and support
- Underlying interconnect technology critical for achieving low latency and high throughput at scale on next-generation exascale systems

Drive future research and innovations to provide scalable and competitive options in this Slingshot ecosystem.

Slingshot Interconnect

High-performance network designed by HPE Cray for upcoming exascale-era systems

- Based on Ethernet
- Adaptive Routing
- Congestion Control
- Isolated Workloads

Empowering the #1 Supercomputer --- Frontier

- Deployed as the interconnect for inter-node communication
- Expected to be deployed on upcoming supercomputers --> El-Capitan at LLNL, Aurora at Argonne

Limitations of State-of-the-art Approaches for Communication

Current accessibility and deployment on early access Slingshot systems:

- Ecosystem with Slingshot-10 interconnection amongst nodes
- Slingshot-10 running over a Slingshot Network with a Mellanox Infiniband adapter

Future accessibility and deployment on upcoming Slingshot systems (i.e. El-Capitan and Frontier):

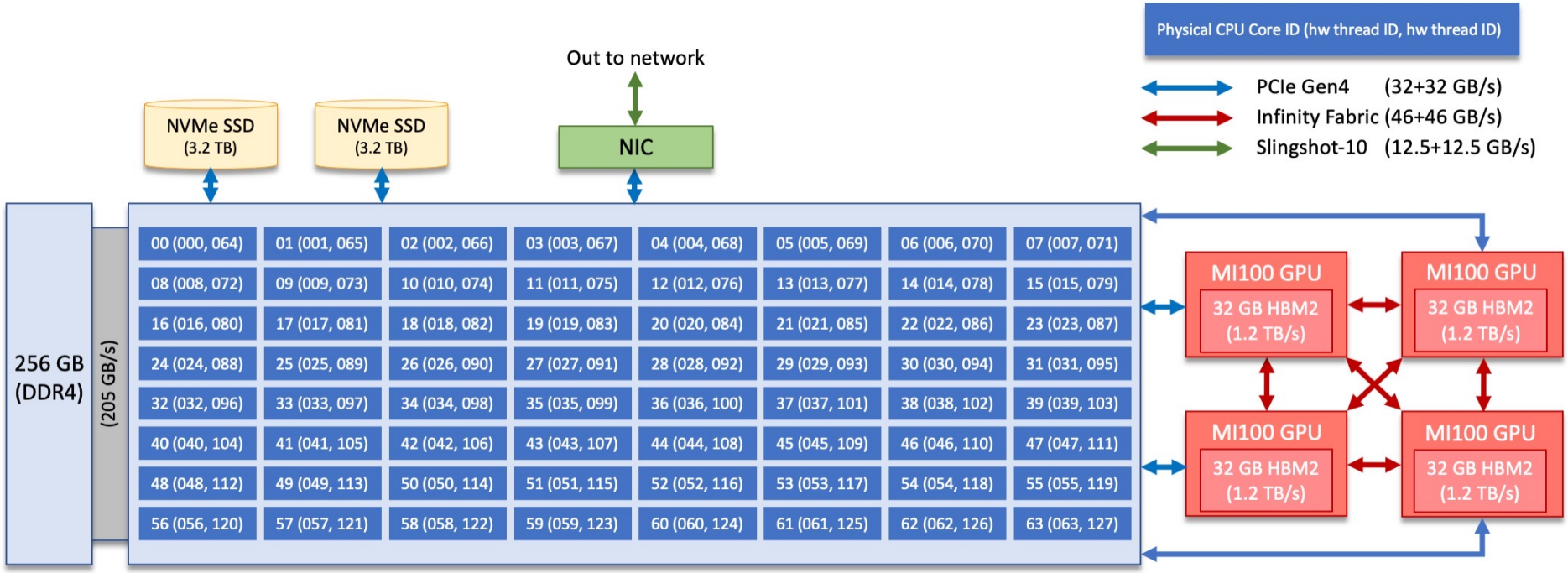
- Slingshot-11
- Deployed over a slingshot fabric and adapter

This second-generation deployment introduces additional challenges for communication libraries to develop functionality over the underlying adapter and fabrics.

Experimental Setup

System & Software Details

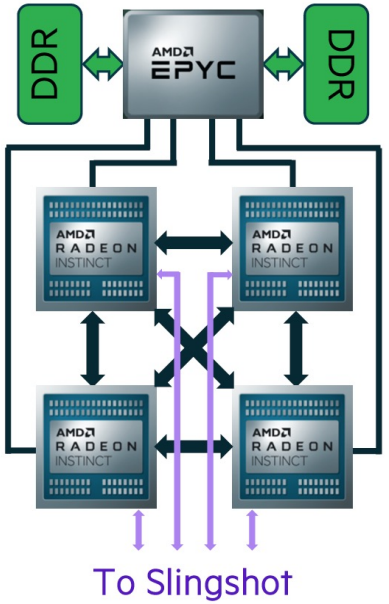
Spock Compute Node



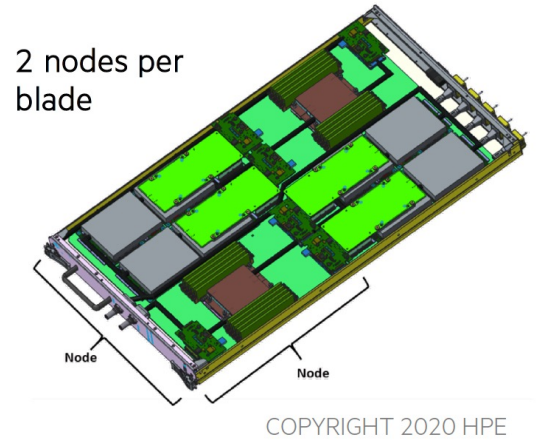
Reference: https://docs.olcf.ornl.gov/systems/spock_quick_start_guide.html

Frontier Compute Node

- 1 HPC and AI Optimized 3rd Gen AMD EPYC CPU
- 4 Purpose Built AMD Instinct 250X GPUs
- **CPU-GPU Interconnect:** AMD Infinity Fabric
- **System Interconnect:** Multiple Slingshot NICs providing 100 GB/s network bandwidth.
  - Slingshot network which provides adaptive routing, congestion management and quality of service.



AMD GPU (ORNL)



Reference: https://www.olcf.ornl.gov/frontier/

Software Details

MPI & Communication Libraries

- CrayMPICH 8.1.14
  - https://docs.nersc.gov/development/programming-models/mpi/cray-mpich/
- MVAPICH2-GDR 2.3.7 & MVAPICH2-X 2.3 & MVAPICH2-3.0a
  - https://mvapich.cse.ohio-state.edu
- OpenMPI 4.1.4 + UCX 1.12.1
  - https://www.open-mpi.org
- RCCL 5.0.2
  - https://github.com/ROCmSoftwarePlatform/rccl

OSU Microbenchmarks 5.9

- https://mvapich.cse.ohio-state.edu/benchmarks/

ROCm version 5.0.2

Experiment Details

CrayMPICH 8.1.14

- Module load cray-mpich/8.1.14
- Module load craype-accel-amd-gfx908
- Run: MPICH_GPU_SUPPORT_ENABLED=1

MVAPICH2-3.0a

- Configure: --with-device=ch4:ofi --with-libfabric=<path-to-libfabric>

MVAPICH2-GDR 2.3.7

- Run: MV2_USE_ROCM=1

OpenMPI 4.1.4 + UCX 1.12.1

- Compile UCX: --with-rocm=<path-to-rocm> --without-knem –without-cuda --enable-optimizations
- Compile OpenMPI: --with-ucx=<path-to-ucx> --without-verbs
- Run: -x UCX_RNDV_THRESH=128

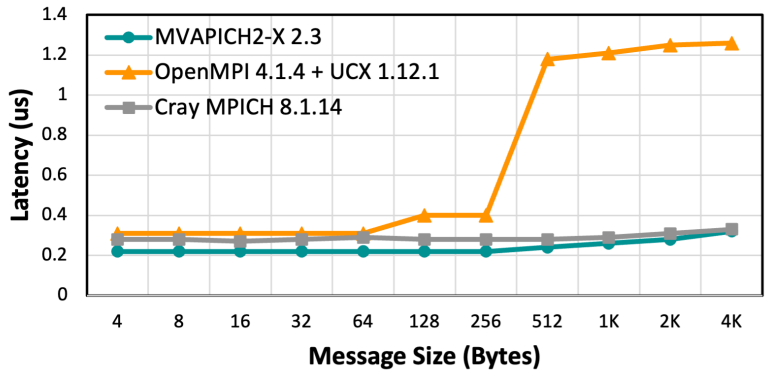
RCCL 5.0.2

- Compile: CXX=<path-to-rocm>/bin/hipcc

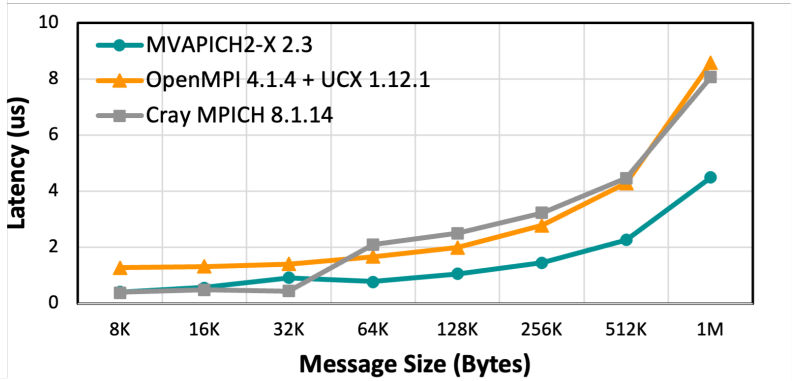
Performance Evaluation

CPU

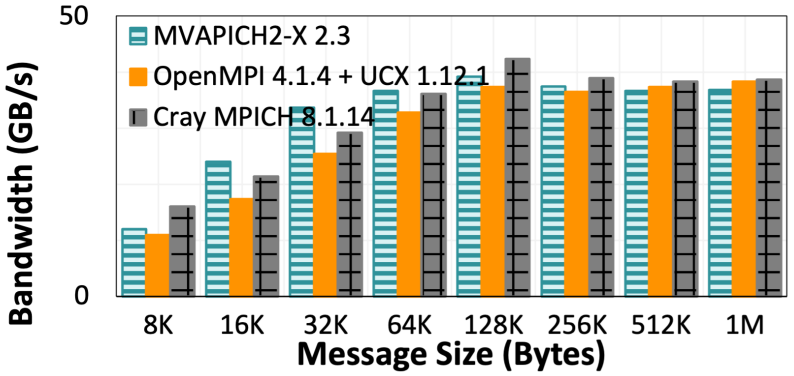
Point-to-Point Performance - Intra-Node CPU



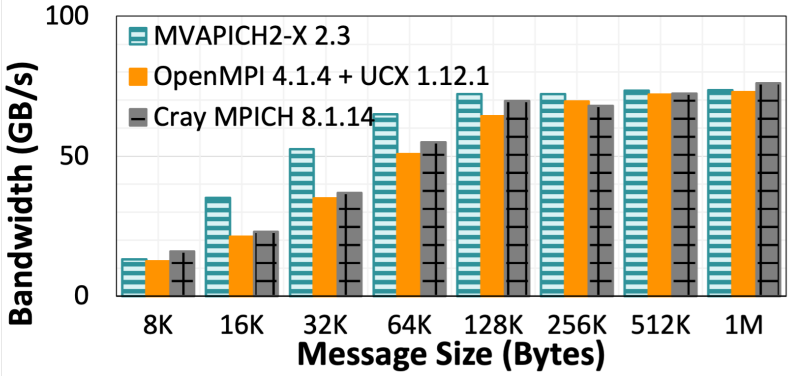
Latency (small messages)



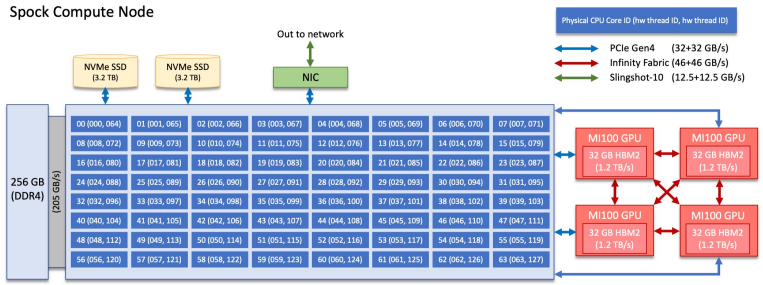
Latency (large messages)



Bandwidth



Bi-Directional Bandwidth



Peak Bandwidth:

- MVAPICH2-X **39.2 GB/s**
- OpenMPI+UCX **38.2GB/s**
- CrayMPICH **42 GB/s**

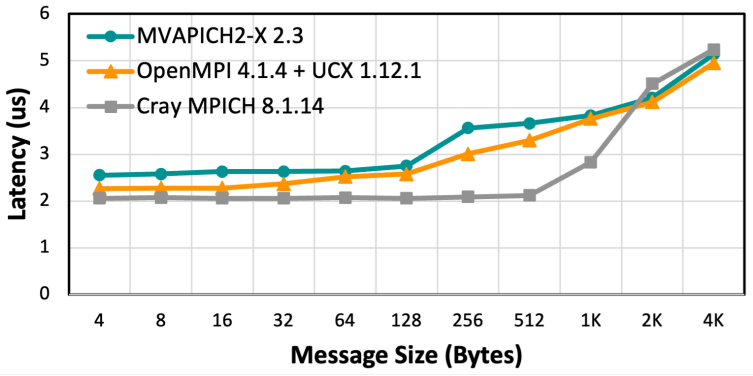
Latency at 4 Bytes:

- MVAPICH2-X **0.22 us**
- OpenMPI+UCX **0.31 us**
- CrayMPICH **0.27 us**

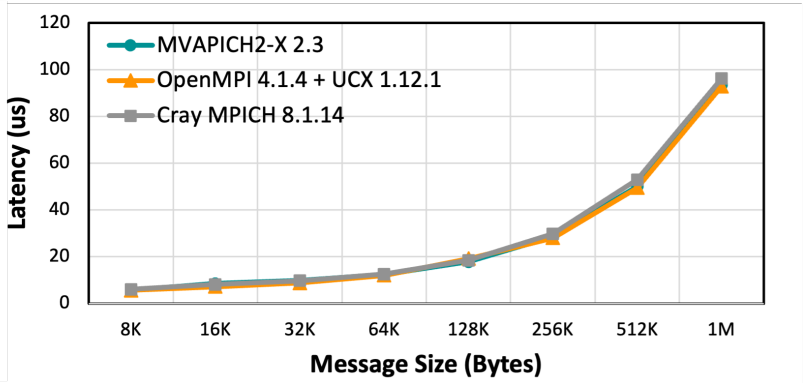
AMD Epyc Rome CPUs on Spock System

Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

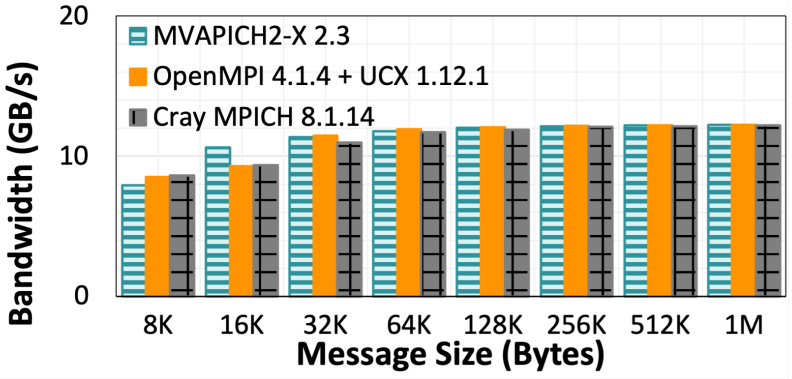
Point-to-Point Performance - Inter-Node CPU



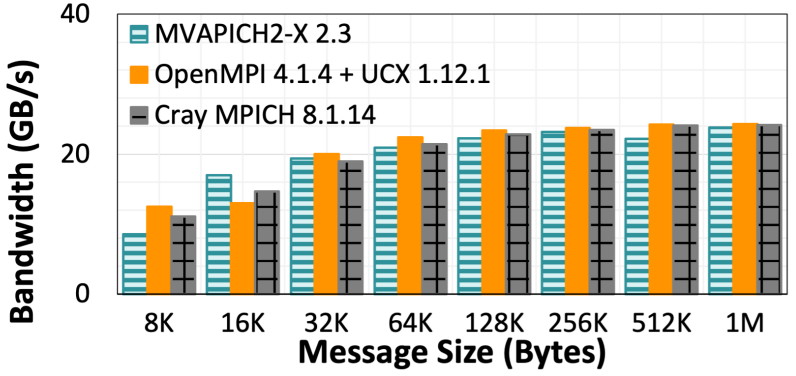
Latency (small messages)



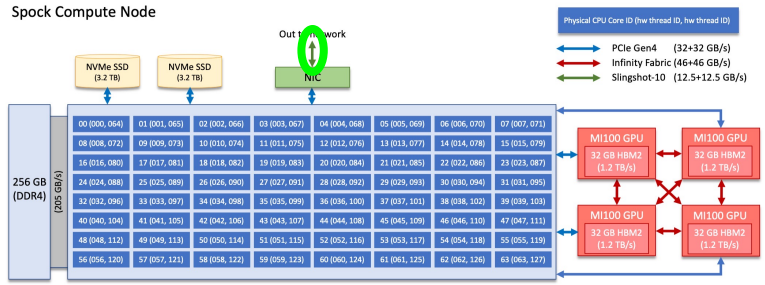
Latency (large messages)



Bandwidth



Bi-Directional Bandwidth



Slingshot-10 Interconnect for over network communication (12.5+12.5 GB/s)

Peak Bandwidth:

- MVAPICH2-X **122.4 MB/s**
- OpenMPI+UCX **122.4 MB/s**
- CrayMPICH **122.4 MB/s**

Latency at 4 Bytes:

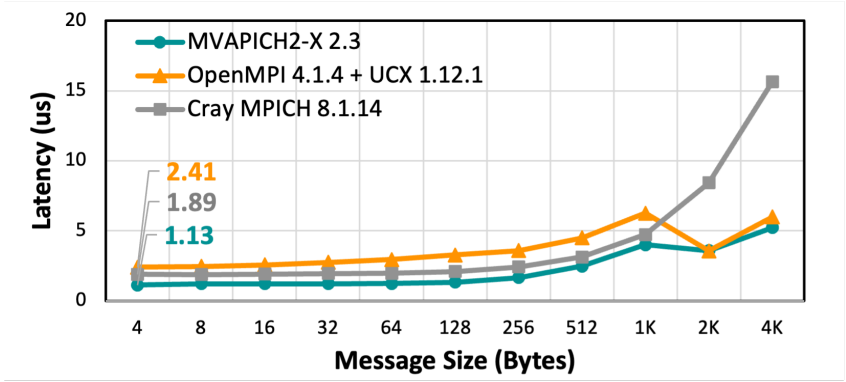
- MVAPICH2-X **2.55 us**
- OpenMPI+UCX **2.27 us**
- CrayMPICH **2.07 us**

AMD Epyc Rome CPUs on Spock System

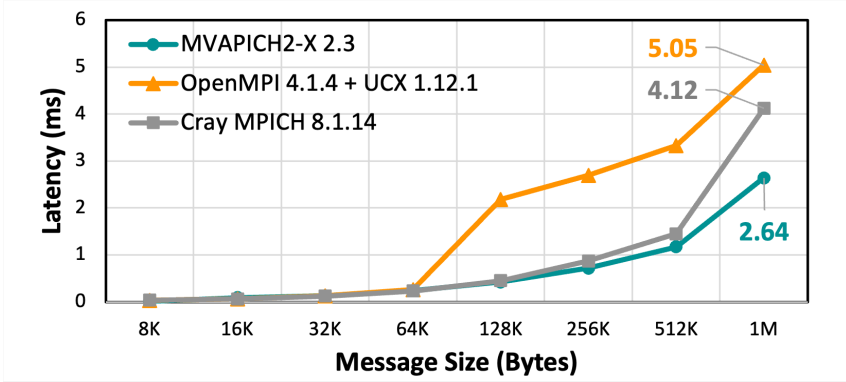
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Collectives Performance - CPU

GATHER

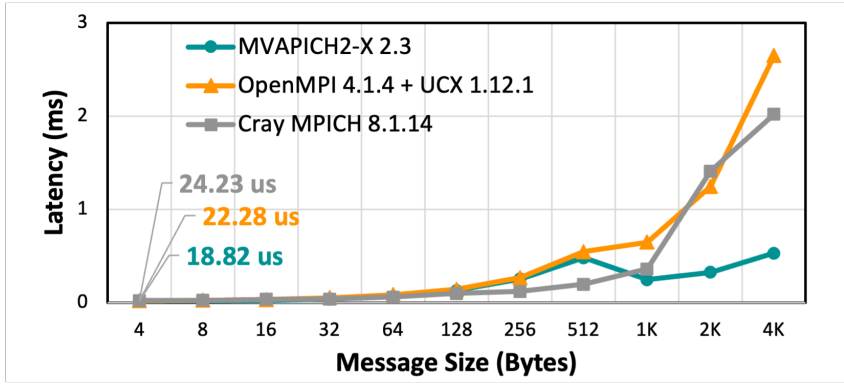


Gather (small messages)

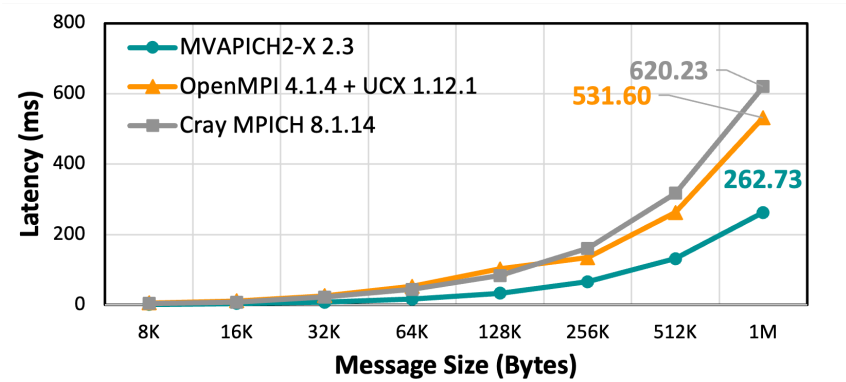


Gather (large messages)

ALLGATHER



Allgather (small messages)



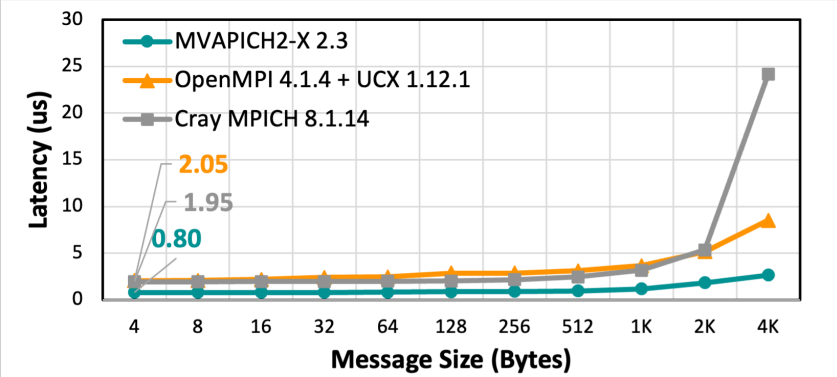
Allgather (large messages)

256 CPUs – 4 Nodes & 64 PPN on Spock System

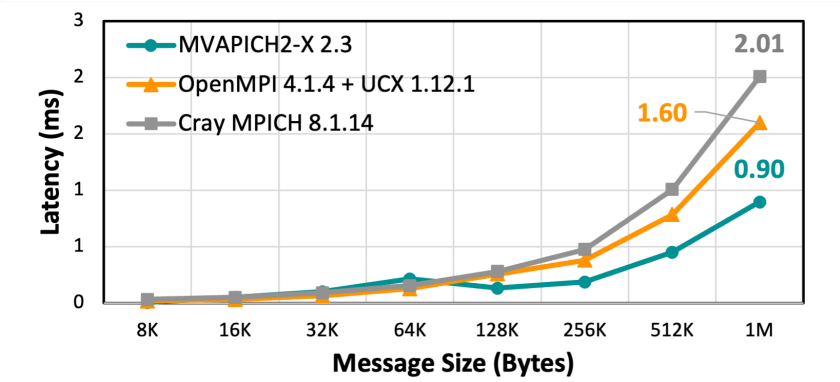
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Collectives Performance - CPU

REDUCE

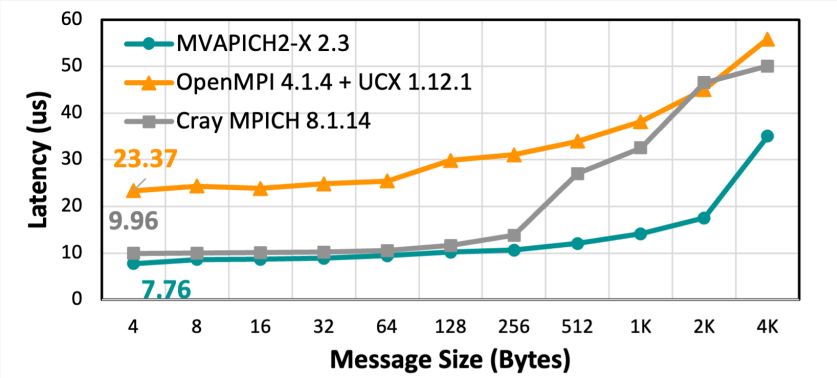


Reduce (small messages)

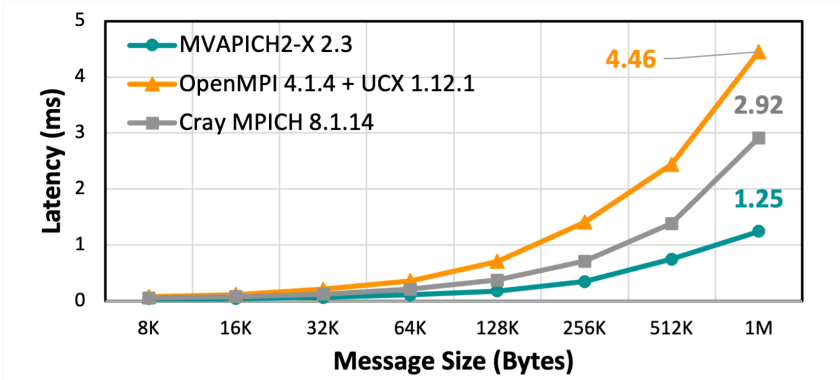


Reduce (large messages)

ALLREDUCE



Allreduce (small messages)



Allreduce (large messages)

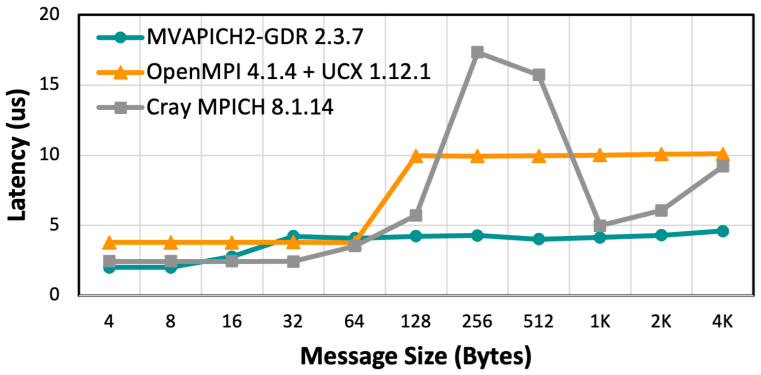
256 CPUs – 4 Nodes & 64 PPN on Spock System

Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

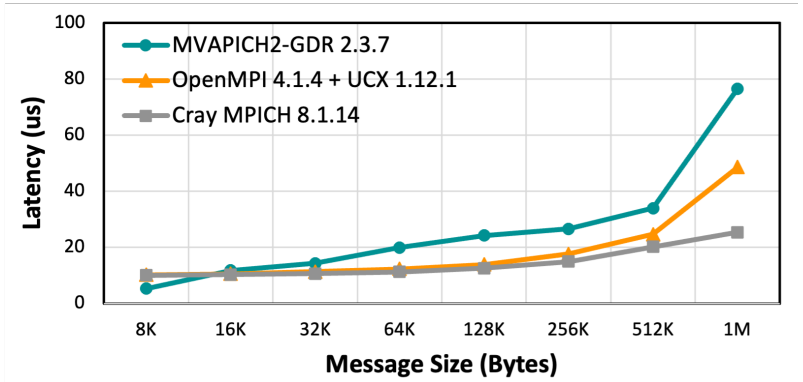
Performance Evaluation

GPU

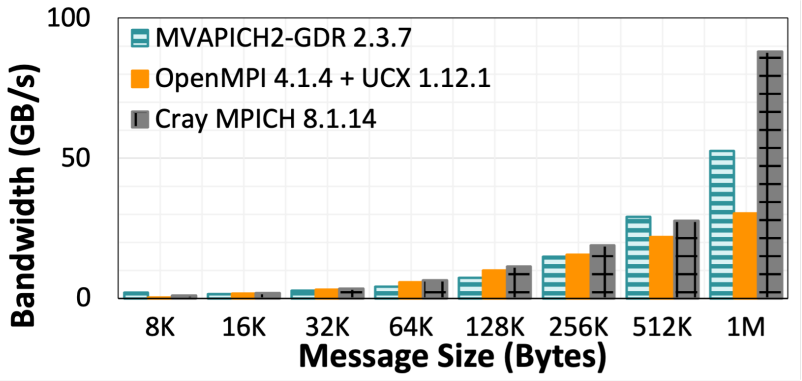
Point-to-Point Performance - Intra-Node GPU



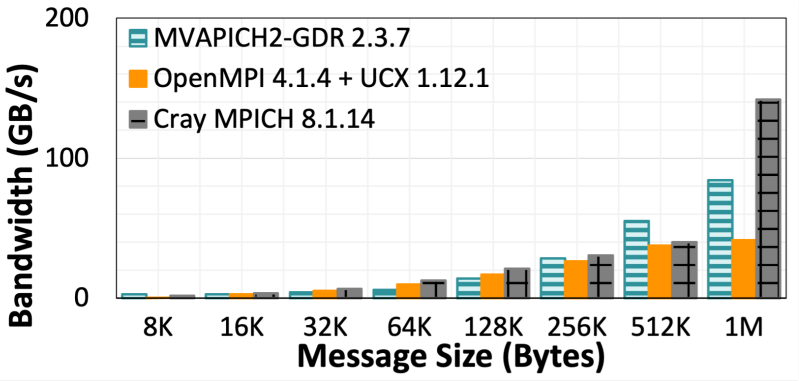
Latency (small messages)



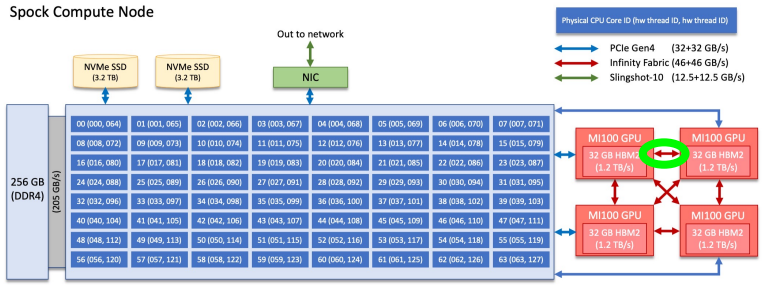
Latency (large messages)



Bandwidth



Bi-Directional Bandwidth



All GPUs connected by Infinity Fabric (46+46GB/s)

- PCI Bar Mapped Memory for small message sizes.
- ROCm Inter-Process Communication (IPC) used in med-large message range.

Peak Bandwidth:

- MVAPICH2-GDR **52.5 GB/s**
- OpenMPI+UCX **30.2 GB/s**
- CrayMPICH **88 GB/s**

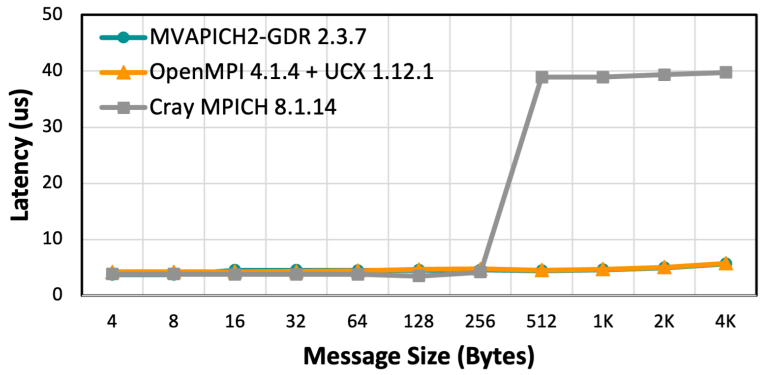
Latency at 4 Bytes:

- MVAPICH2-GDR **2.01 us**
- OpenMPI+UCX **3.79 us**
- CrayMPICH **2.44 us**

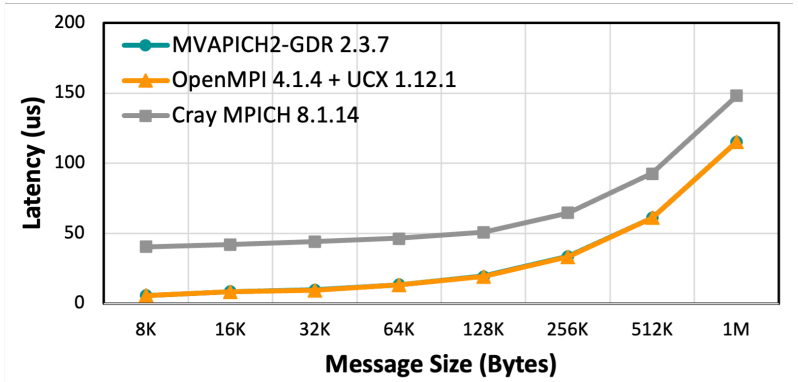
MI100 GPUs on Spock System

Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

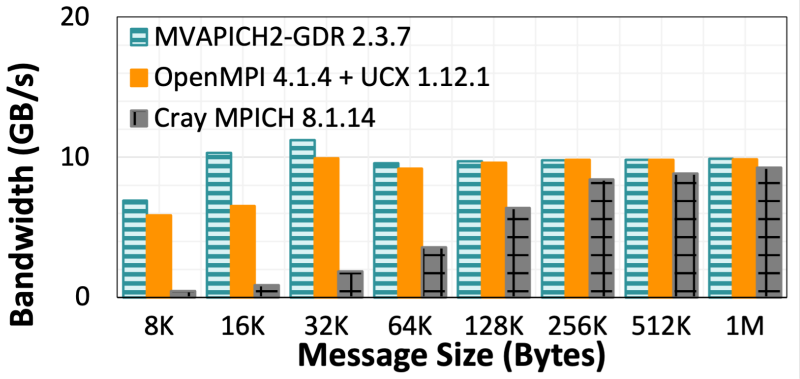
Point-to-Point Performance - Inter-Node GPU



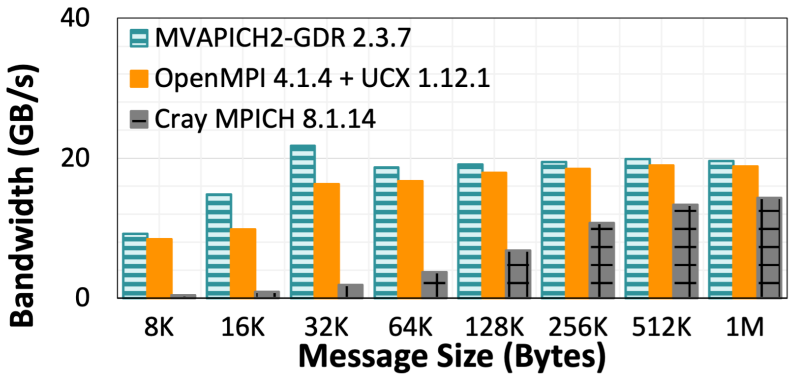
Latency (small messages)



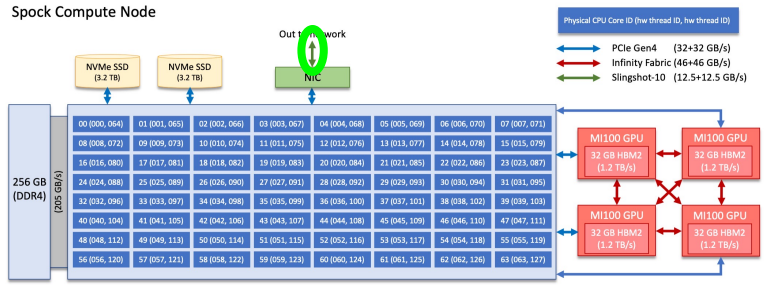
Latency (large messages)



Bandwidth



Bi-Directional Bandwidth



Slingshot-10 Interconnect for over network communication (12.5+12.5 GB/s)

Peak Bandwidth:

- MVAPICH2-GDR **9.9 GB/s**
- OpenMPI+UCX **9.8 GB/s**
- CrayMPICH **9.2 GB/s**

Latency at 4 Bytes:

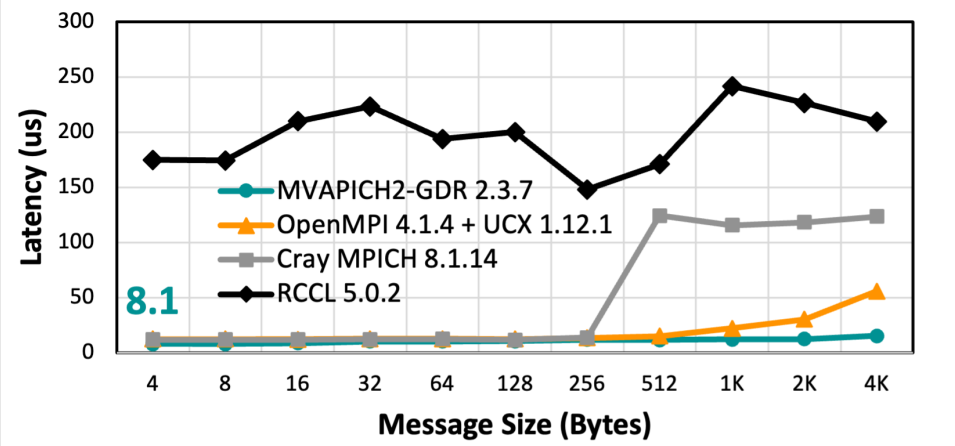
- MVAPICH2-GDR **3.73 us**
- OpenMPI+UCX **4.23 us**
- CrayMPICH **3.8 us**

MI100 GPUs on Spock System

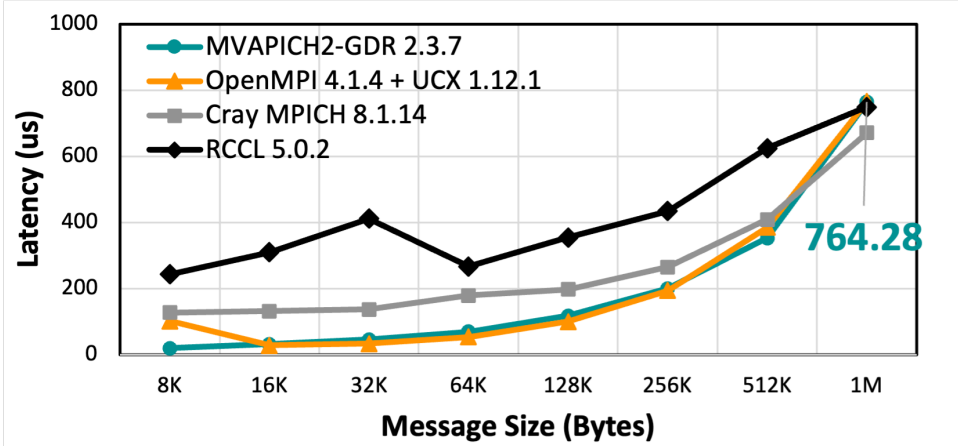
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Collectives Performance - GPU

BROADCAST

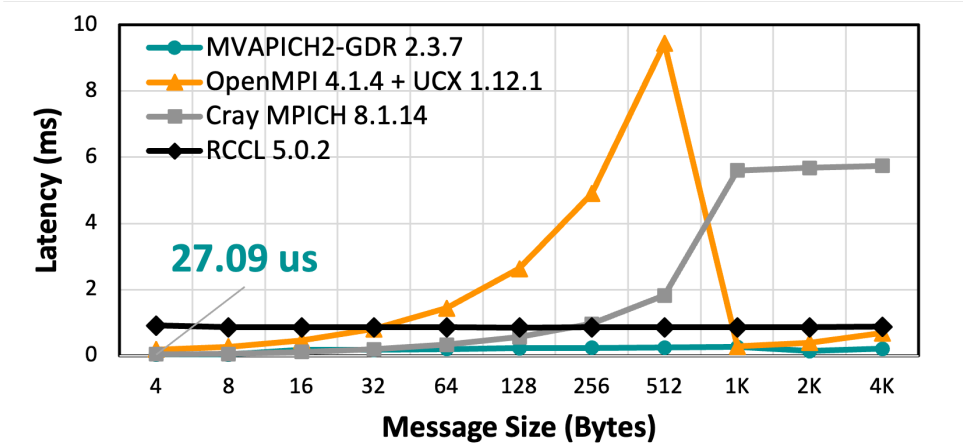


Broadcast (small messages)

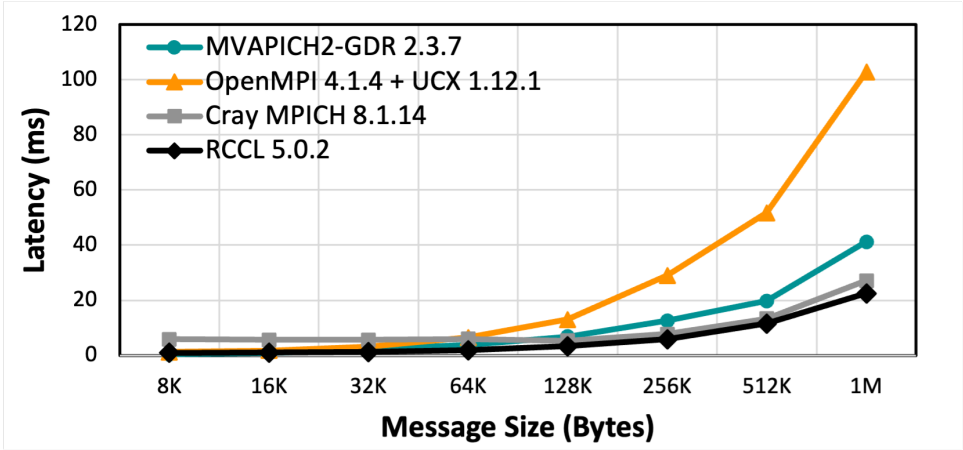


Broadcast (large messages)

ALLTOALL



Alltoall (small messages)



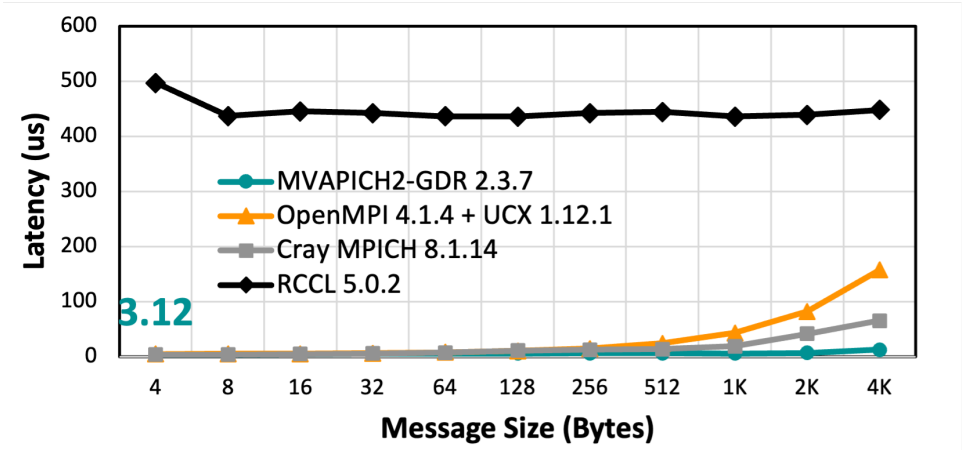
Alltoall (large messages)

64 GPUs – 16 Nodes & 4 GPUs Per Node on Spock System

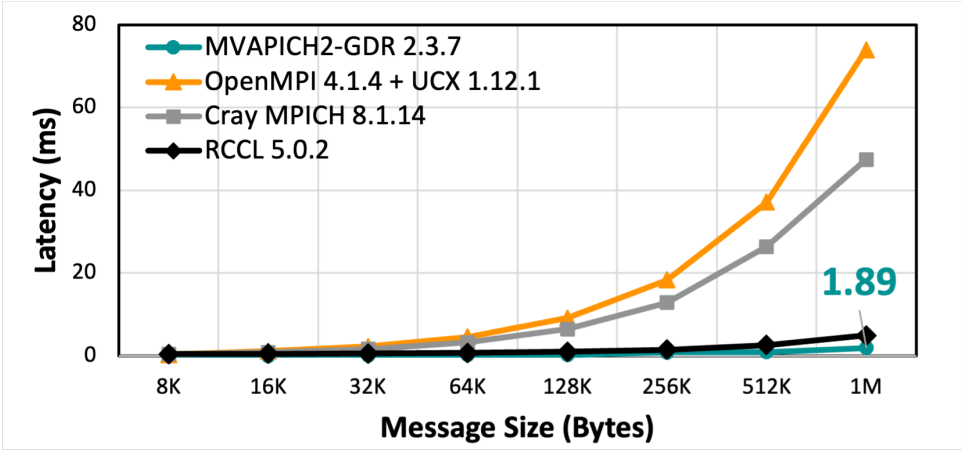
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Collectives Performance - GPU

GATHER

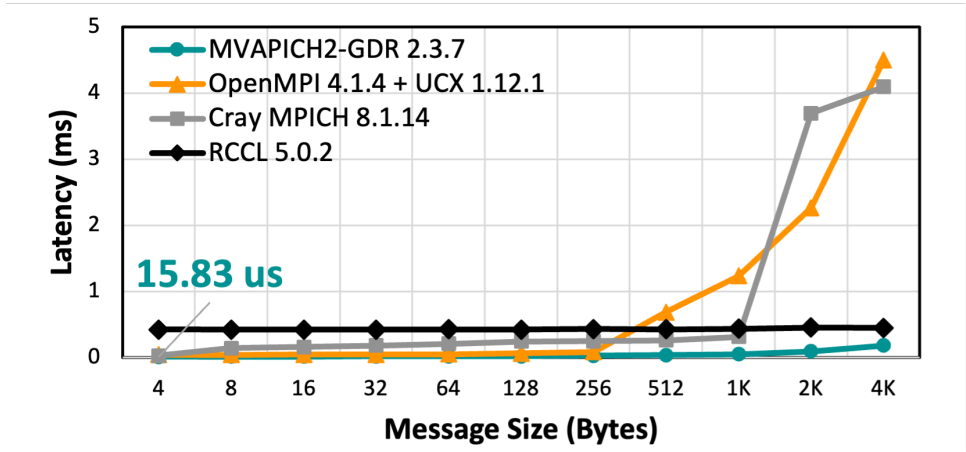


Gather (small messages)

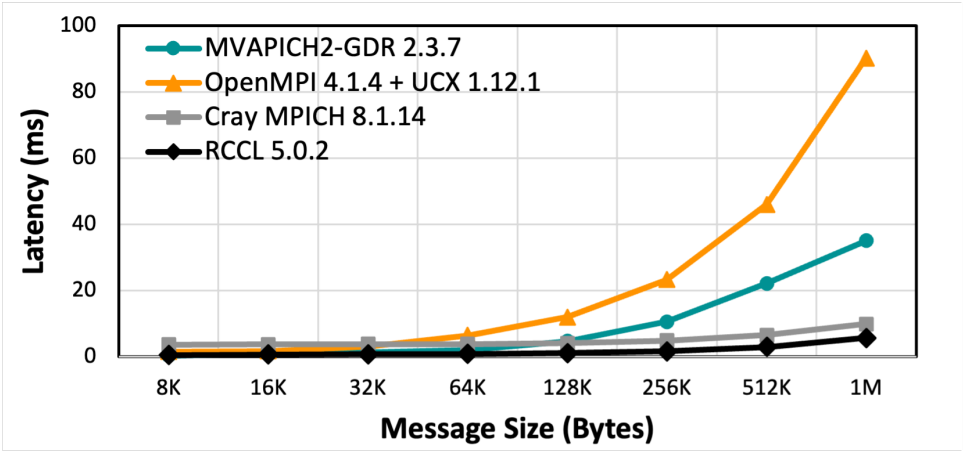


Gather (large messages)

ALLGATHER



Allgather (small messages)



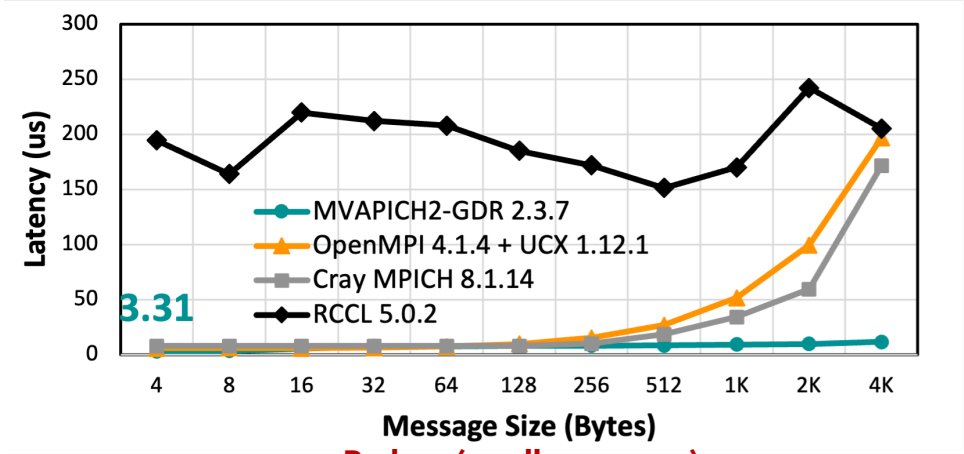
Allgather (large messages)

64 GPUs - 16 Nodes & 4 GPUs Per Node on Spock System

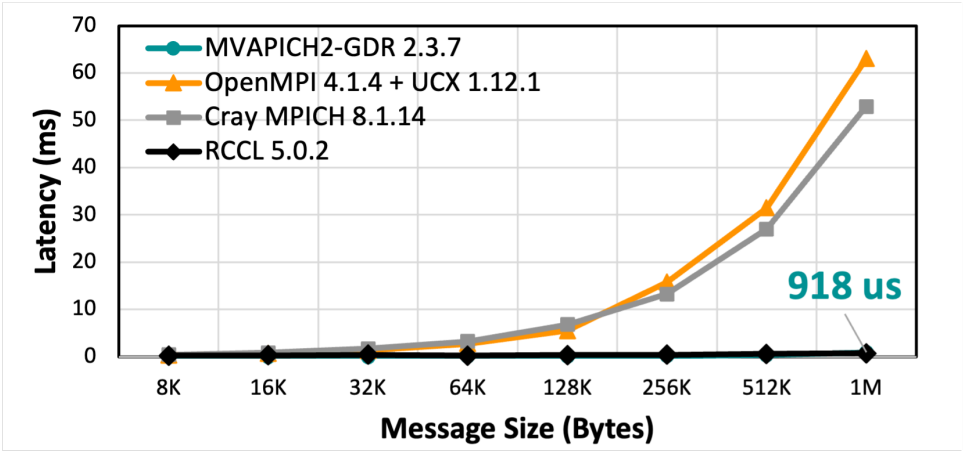
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Collectives Performance - GPU

REDUCE

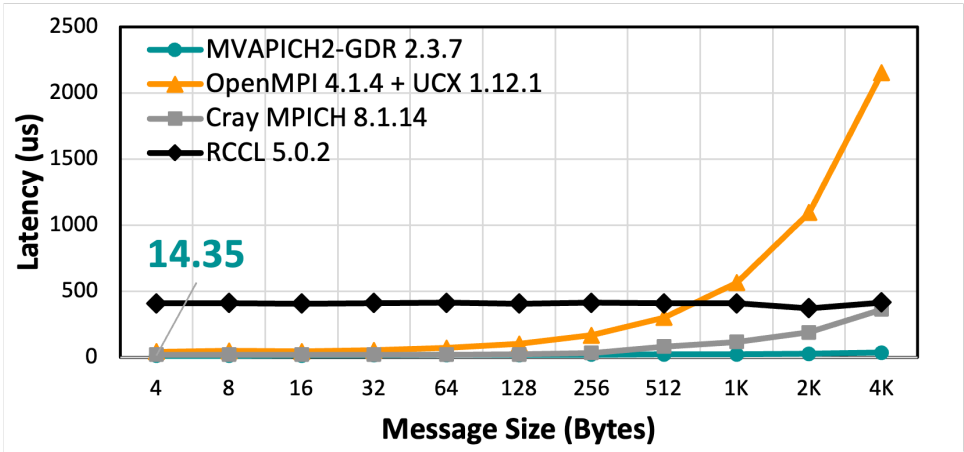


Reduce (small messages)

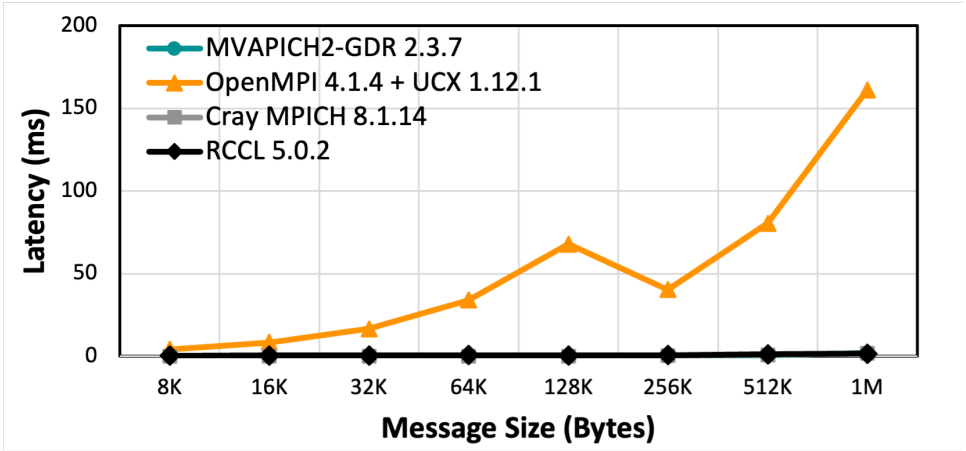


Reduce (large messages)

ALLREDUCE



Allreduce (small messages)



Allreduce (large messages)

64 GPUs – 16 Nodes & 4 GPUs Per Node on Spock System

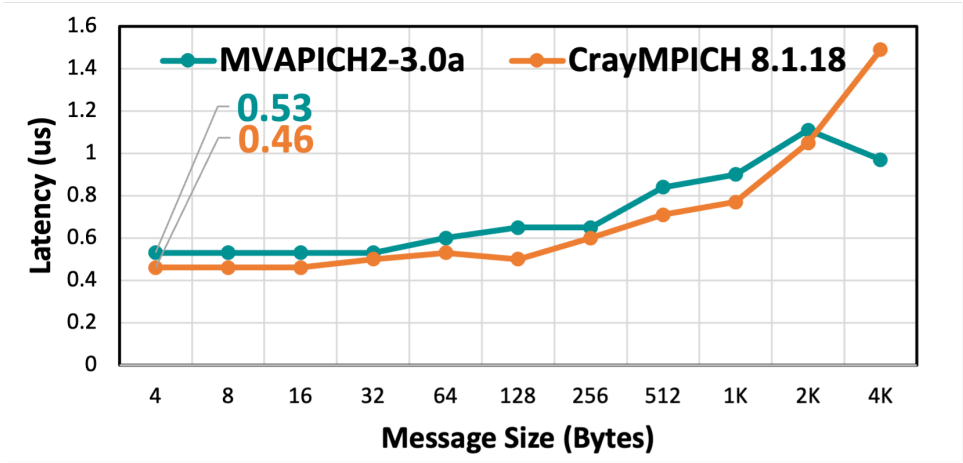
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

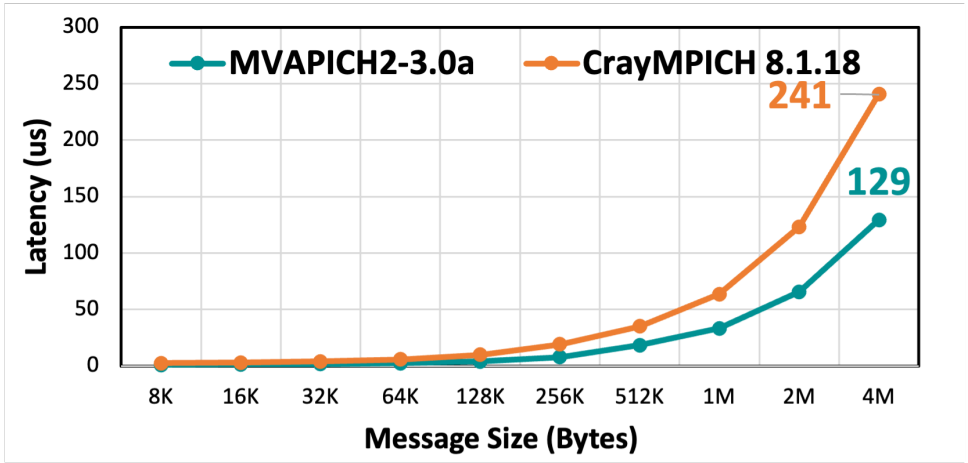
Performance Evaluation Slingshot-11

CPU

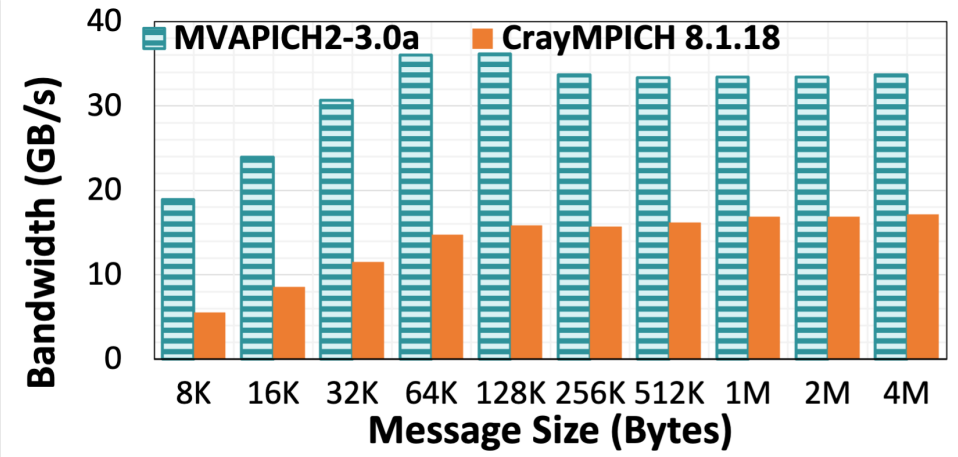
Point-to-Point Performance - Intra-Node CPU

LATENCY

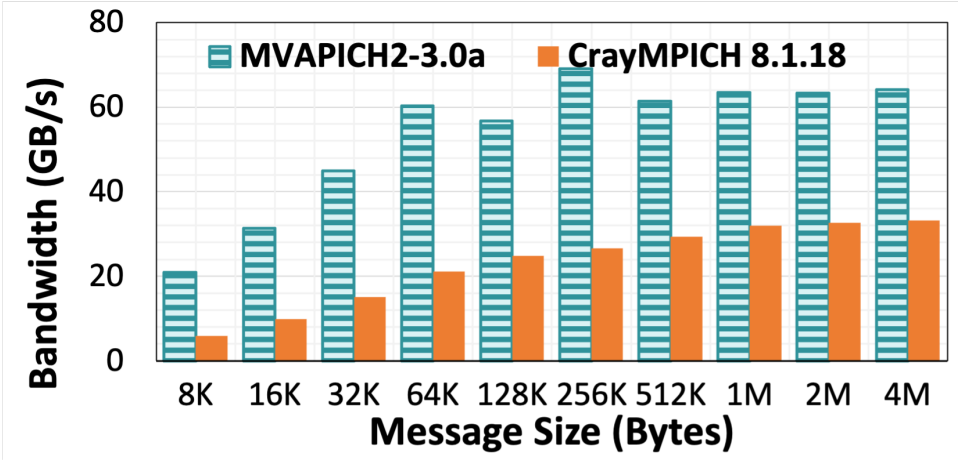




BANDWIDTH



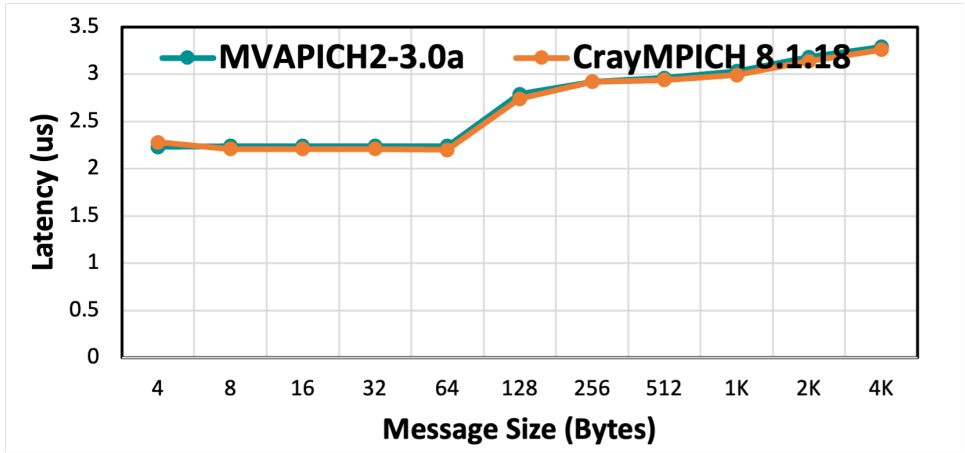
BI-BANDWIDTH

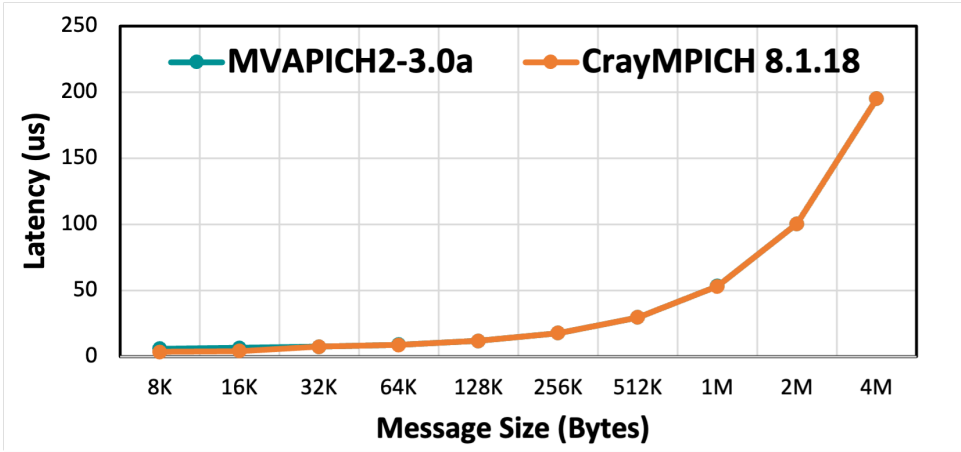


System with Slingshot-11 Networking

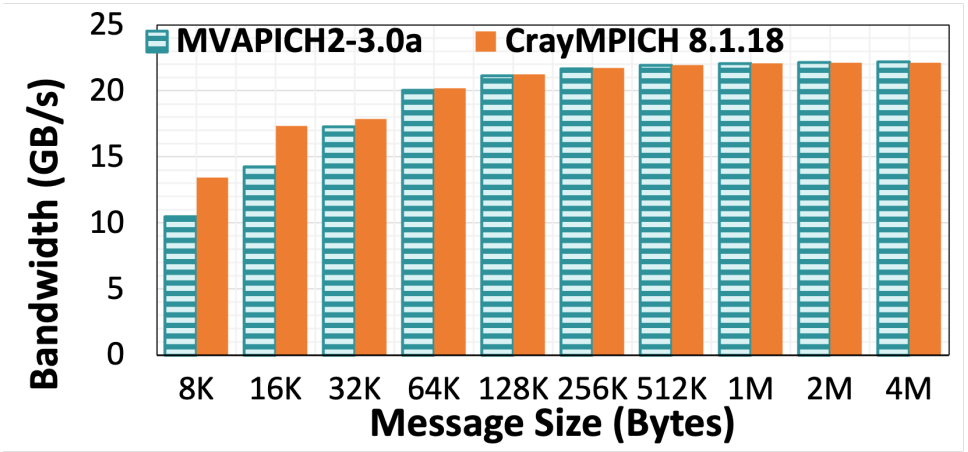
Point-to-Point Performance - Inter-Node CPU

LATENCY

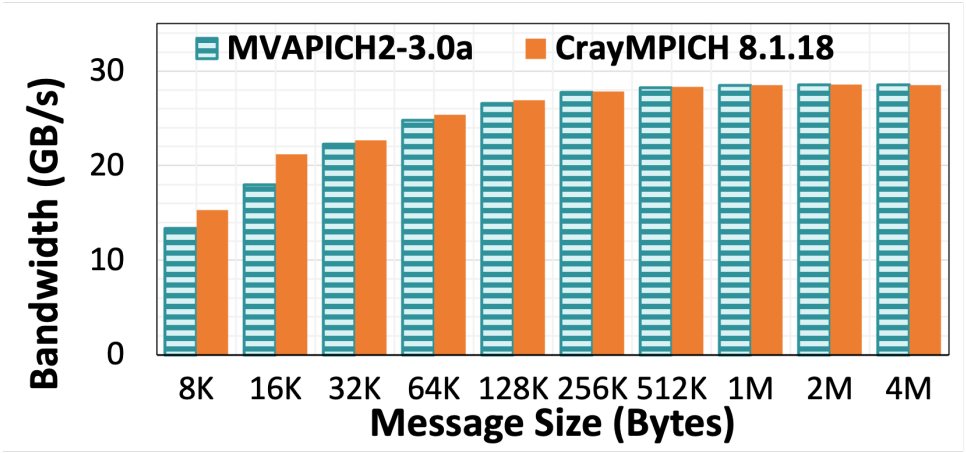




BANDWIDTH

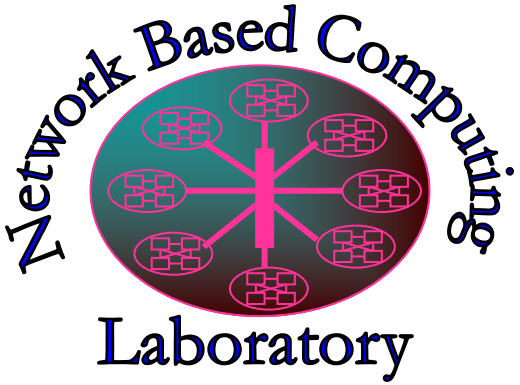


BI-BANDWIDTH



System with Slingshot-11 Networking

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/