2022 OFA Virtual Workshop

Performance Evaluation of MPI on the Slingshot Interconnect

Kawthar Shafie Khorassani, Hari Subramoni, and Dhabaleswar K. (DK) Panda

The Ohio State University

panda@cse.ohio-state.edu

OUTLINE

- **Introduction**
- MPI + Slingshot
- CPU-Level Performance
- GPU-Level Performance
- Conclusion

INTRODUCTION

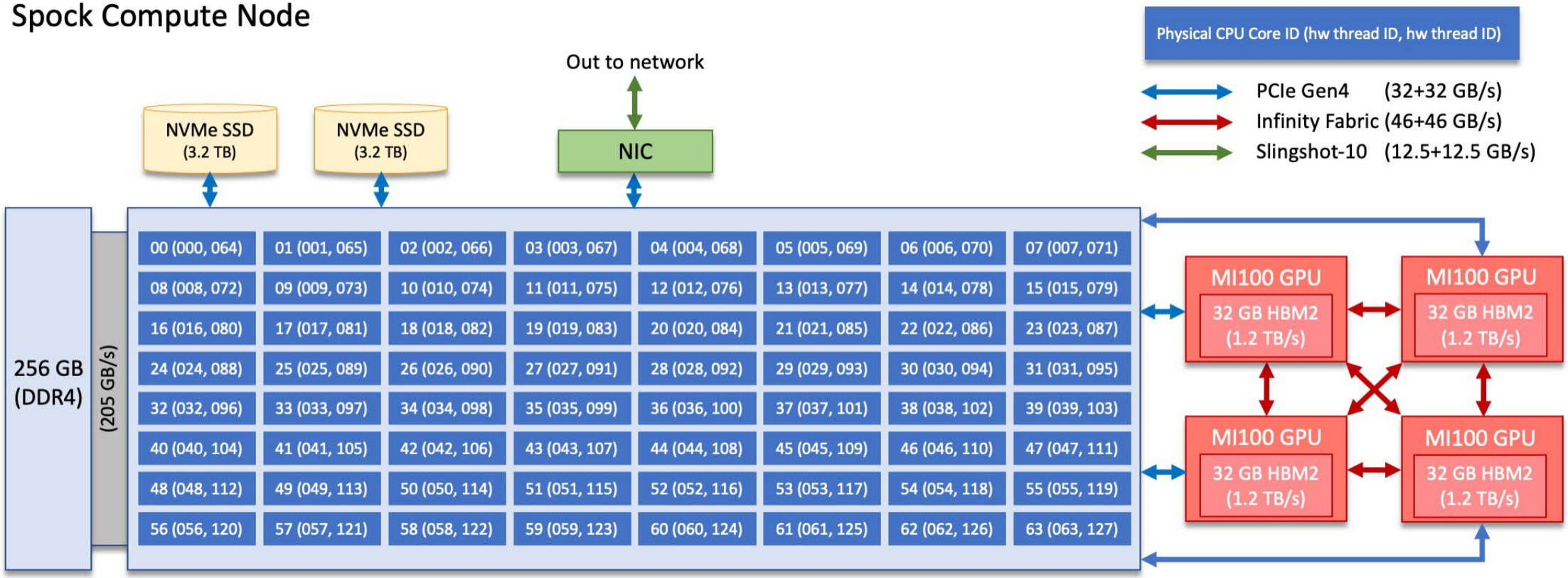
The HPE Cray Slingshot Interconnect technology will be deployed on upcoming exascale systems such as Frontier@OLCF and El-Capitan@LLNL.

Evaluation of the Slingshot Networking Ecosystem and MPI on CPUs and GPUs.

Slingshot Features:

- Adaptive Routing
- Congestion Control
- QoS Features

SPOCK COMPUTE NODE

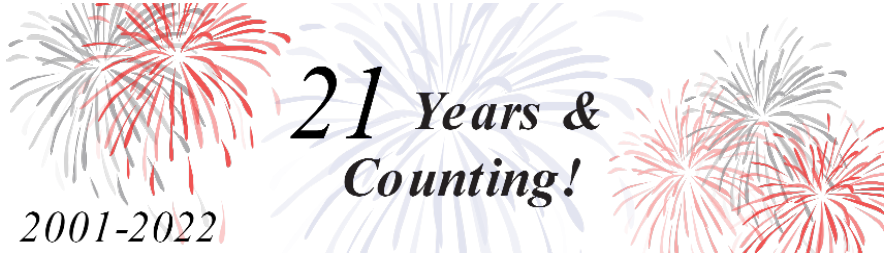


OUTLINE

- Introduction
- **MPI + Slingshot**
- CPU-Level Performance
- GPU-Level Performance
- Conclusion

OVERVIEW OF THE MVAPICH2 PROJECT

- **High Performance open-source MPI Library**
- **Support for multiple interconnects**
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, Rockport Networks, and Slingshot
- **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 3,200 organizations in 89 countries
- More than 1.57 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '21 ranking)
  - 4th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 13th, 448, 448 cores (Frontera) at TACC
  - 26th, 288,288 cores (Lassen) at LLNL
  - 38th, 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 13th ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

OUTLINE

- Introduction
- MPI + Slingshot
- **CPU-Level Performance**
- GPU-Level Performance
- Conclusion

EXPERIMENTAL SETUP

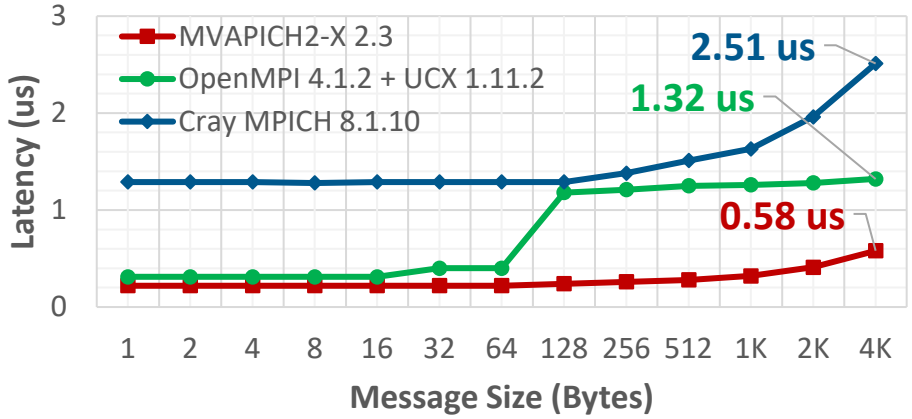
- **MPI and Communication Libraries:**
  - MVAPICH2-GDR 2.3.6 (on GPUs) & MVAPICH2-X 2.3 (on CPUs)
    - http://mvapich.cse.ohio-state.edu/downloads/
  - OpenMPI 4.1.2 + UCX 1.11.2
    - https://www.open-mpi.org
  - CrayMPICH 8.1.10
    - https://docs.nersc.gov/development/programming-models/mpi/cray-mpich/
  - RCCL 4.5.0
    - https://github.com/ROCmSoftwarePlatform/rccl

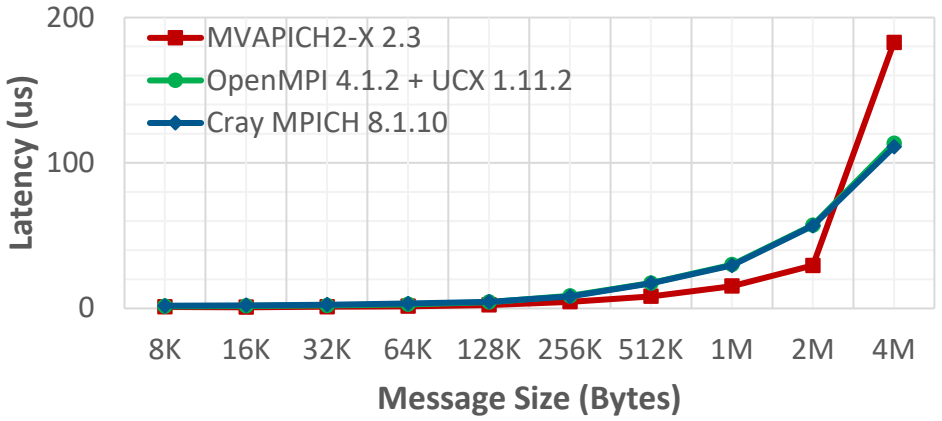
- **Spock Cluster at OLCF**
  - ROCm Version 5.0.0
  - AMD Epyc Rome CPUs
  - MI100 GPUs

- **Performance evaluation done using point-to-point and collective benchmarks from the OSU-Microbenchmarks 5.9 suite.**
  - http://mvapich.cse.ohio-state.edu/benchmarks/

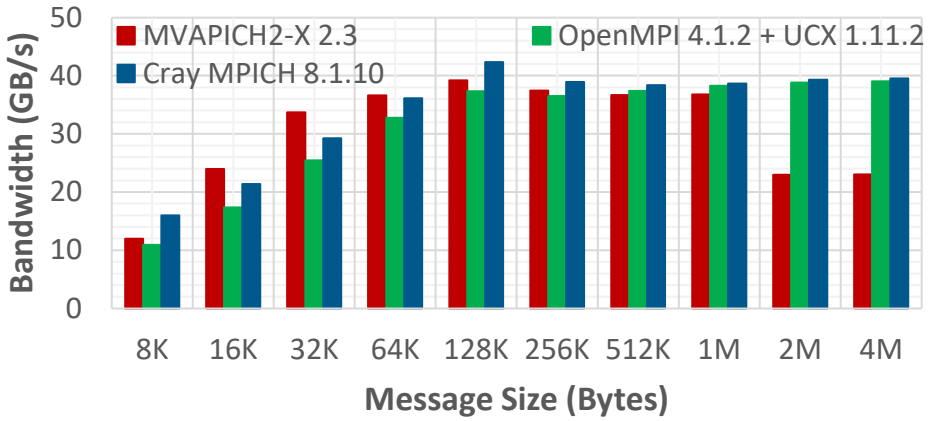
POINT-TO-POINT INTRA-NODE PERFORMANCE (CPU)

LATENCY:

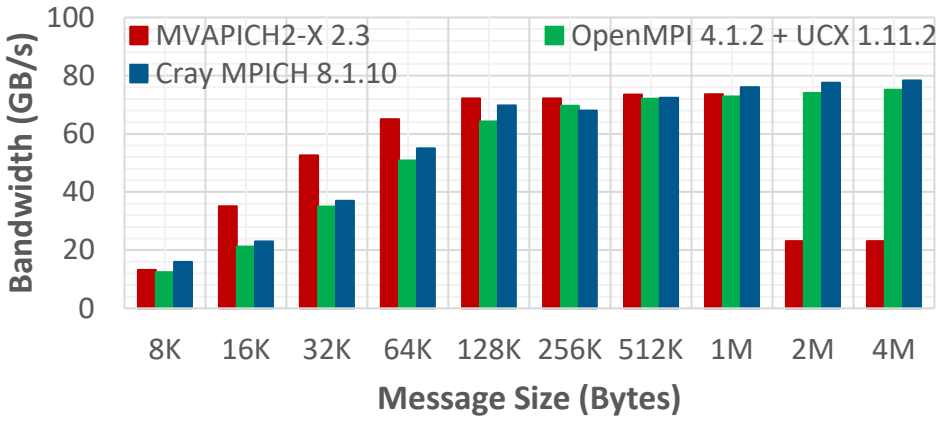




BANDWIDTH:



BI-DIRECTIONAL BANDWIDTH:



OLCF Spock Cluster – AMD Epyc Rome CPUs

MVAPICH2-X XPMEM for intra-node large message range.

All libraries Peak Bandwidth at 128KB:

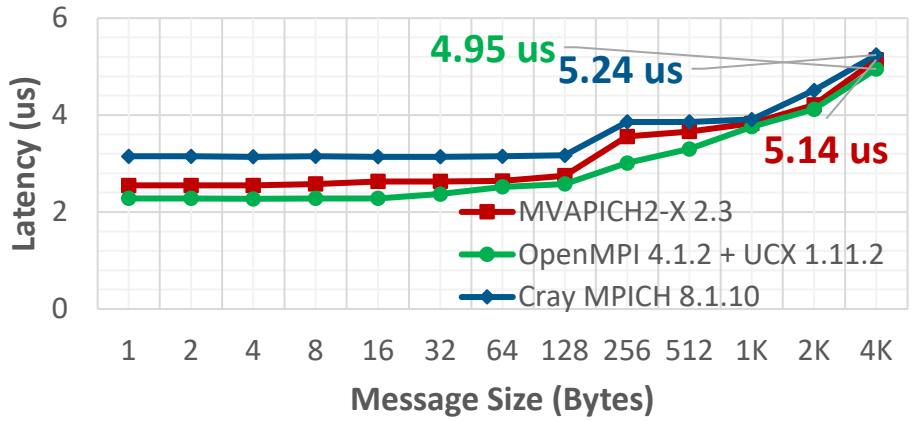
- MVAPICH2-X **39 GB/s**
- OpenMPI+UCX **37 GB/s**
- CrayMPICH **42 GB/s**

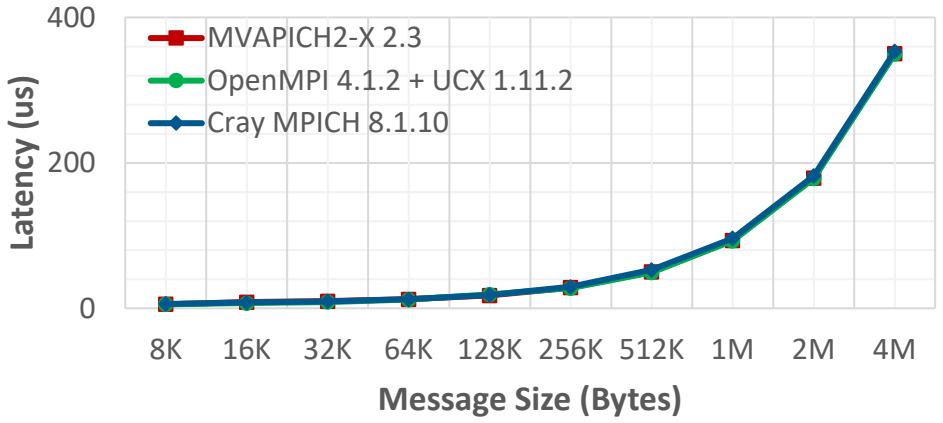
Minimum Latency at 1 Byte

- MVAPICH2-X **0.22 us**
- OpenMPI+UCX **0.31 us**
- CrayMPICH **1.29 us**

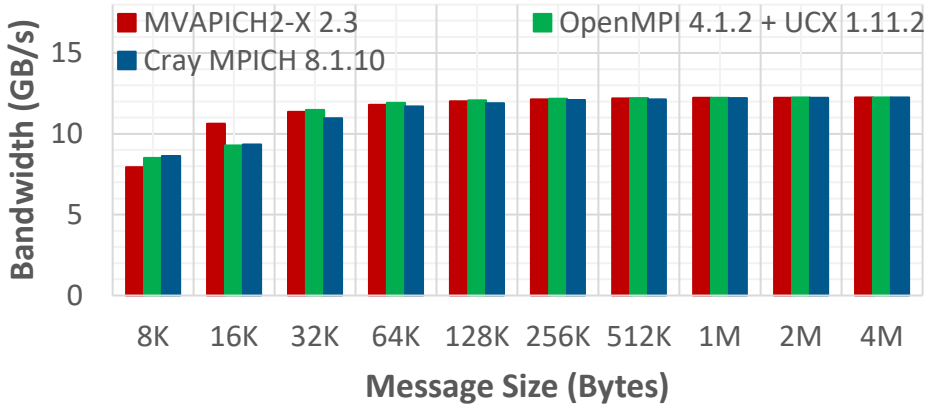
POINT-TO-POINT INTER-NODE PERFORMANCE (CPU)

LATENCY:

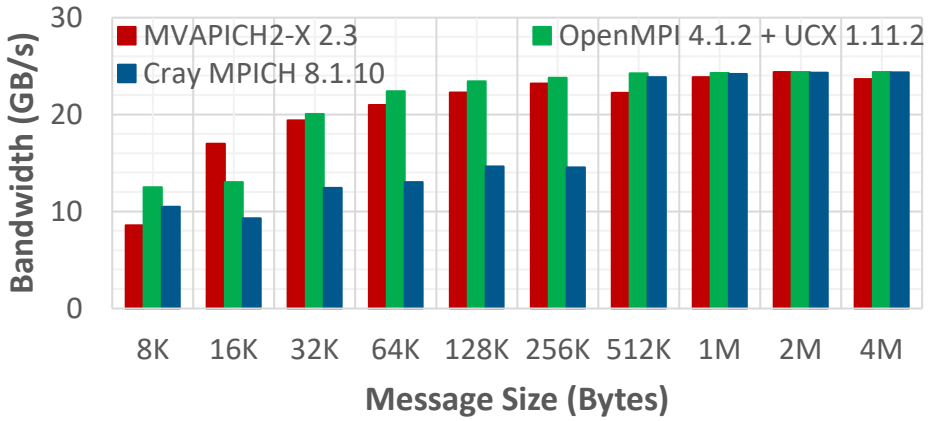




BANDWIDTH:



BI-DIRECTIONAL BANDWIDTH:



OLCF Spock Cluster – AMD Epyc Rome CPUs

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

All libraries Peak Bandwidth at 4MB similar range:

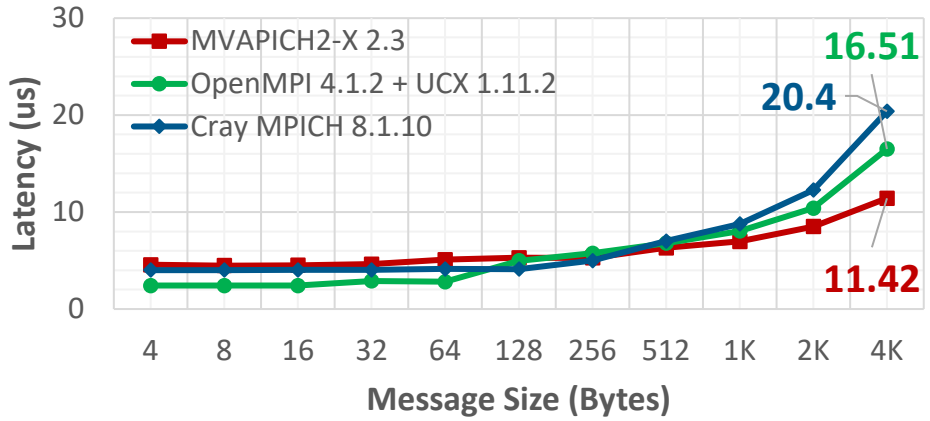
- MVAPICH2-X **12.2 GB/s**
- OpenMPI+UCX **12.2 GB/s**
- CrayMPICH **12.2 GB/s**

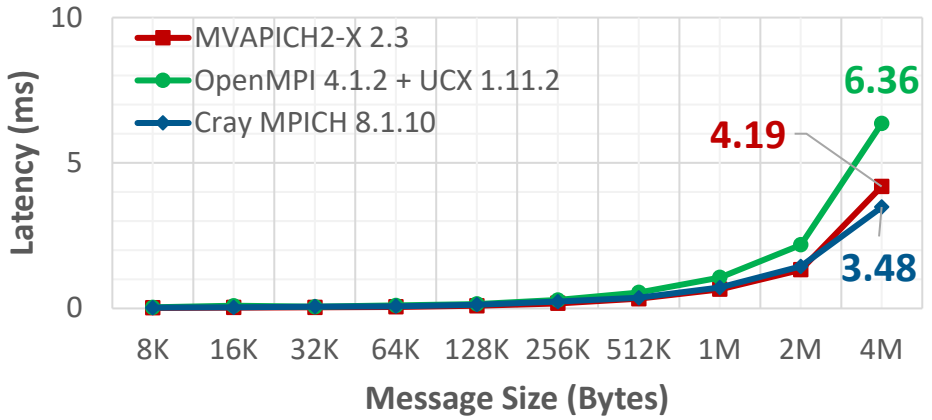
Minimum Latency at 1 Byte

- MVAPICH2-X **2.55 us**
- OpenMPI+UCX **2.28 us**
- CrayMPICH **3.15 us**

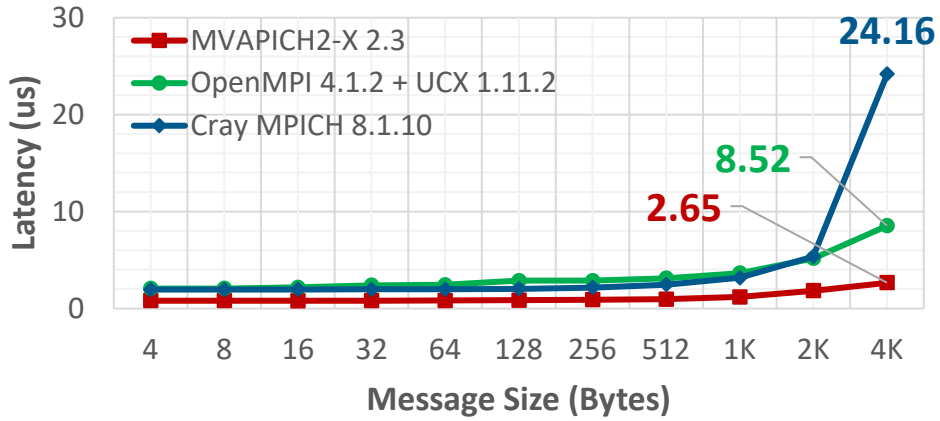
COLLECTIVES PERFORMANCE (CPU)

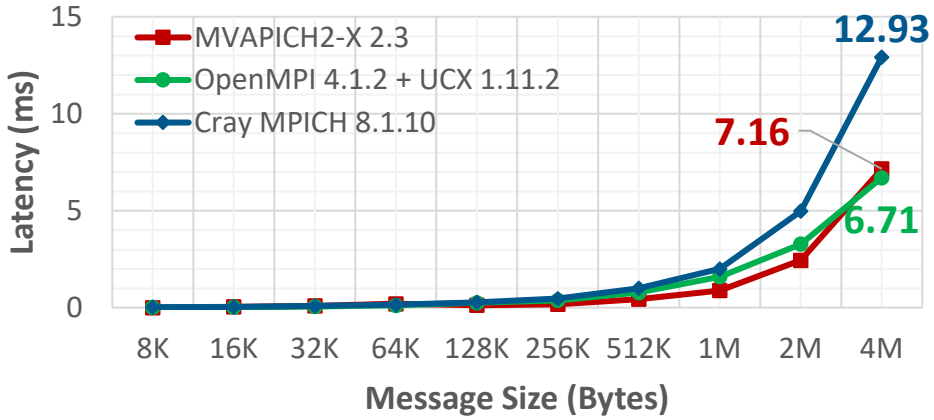
BROADCAST:





REDUCE:

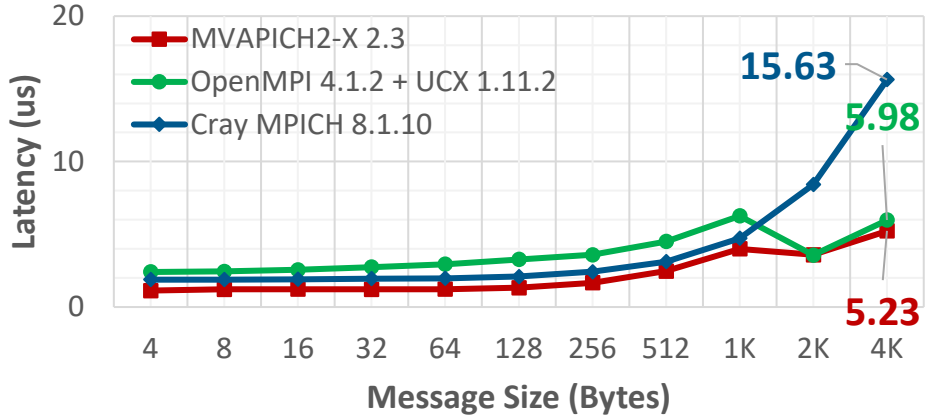


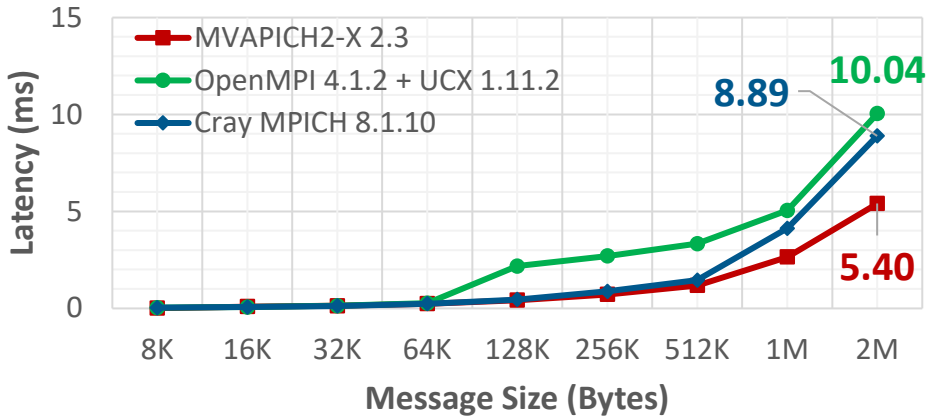


OLCF Spock Cluster – AMD Epyc Rome CPUs (4 Nodes 64 PPN – 256 Processes)

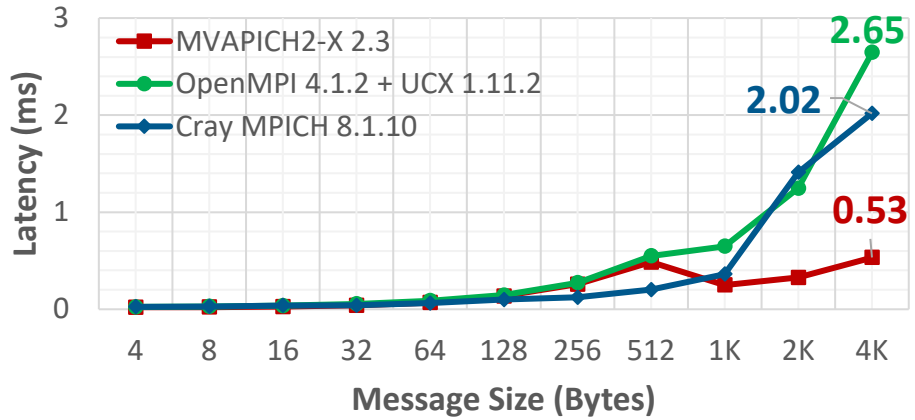
COLLECTIVES PERFORMANCE (CPU)

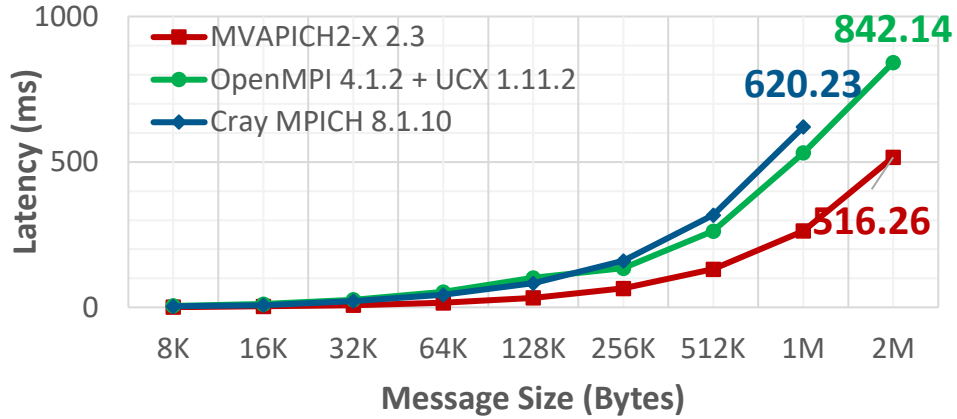
GATHER:





ALLGATHER:

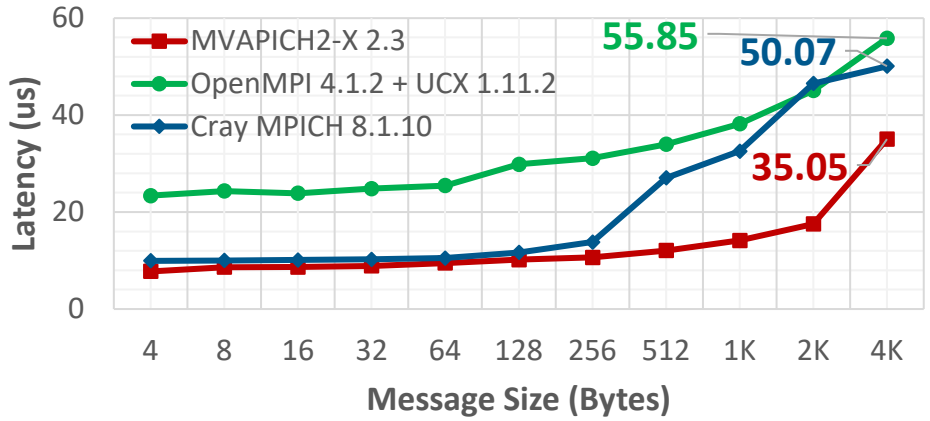


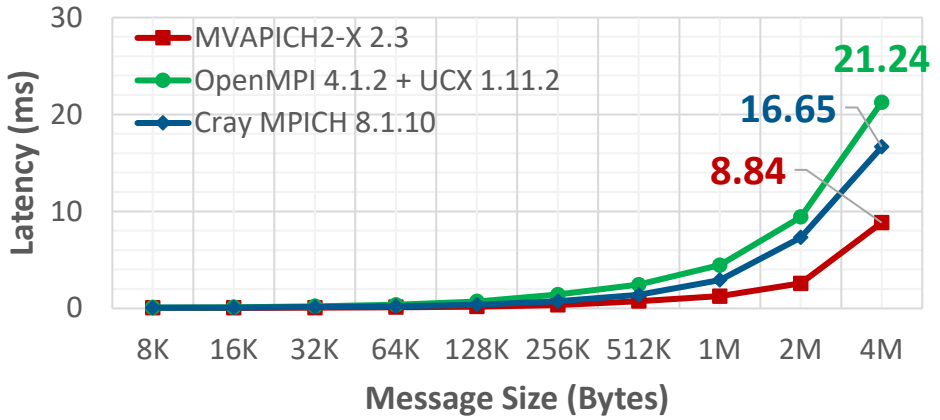


OLCF Spock Cluster – AMD Epyc Rome CPUs (4 Nodes 64 PPN – 256 Processes)

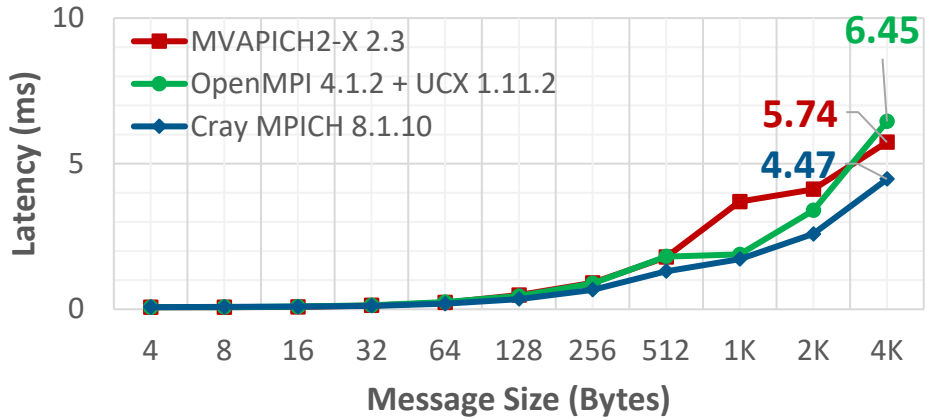
COLLECTIVES PERFORMANCE (CPU)

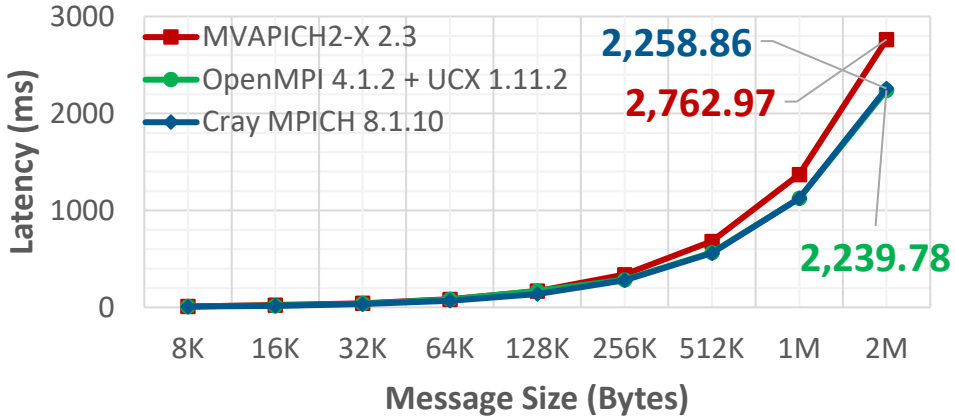
ALLREDUCE:





ALLTOALL:





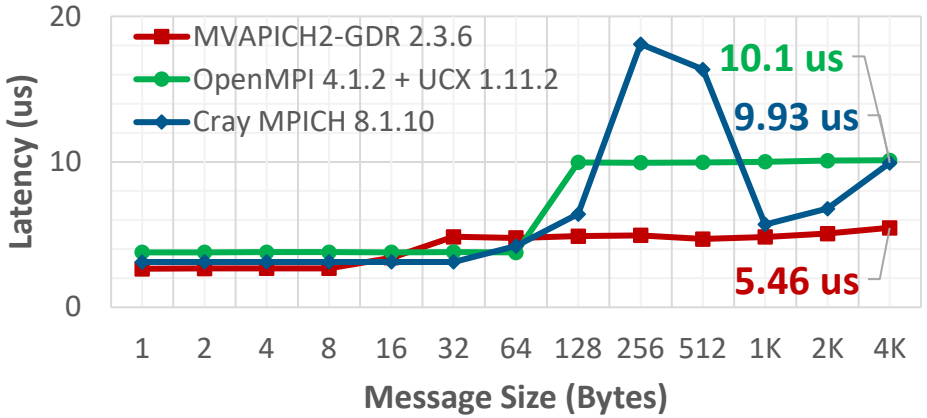
OLCF Spock Cluster – AMD Epyc Rome CPUs (4 Nodes 64 PPN – 256 Processes)

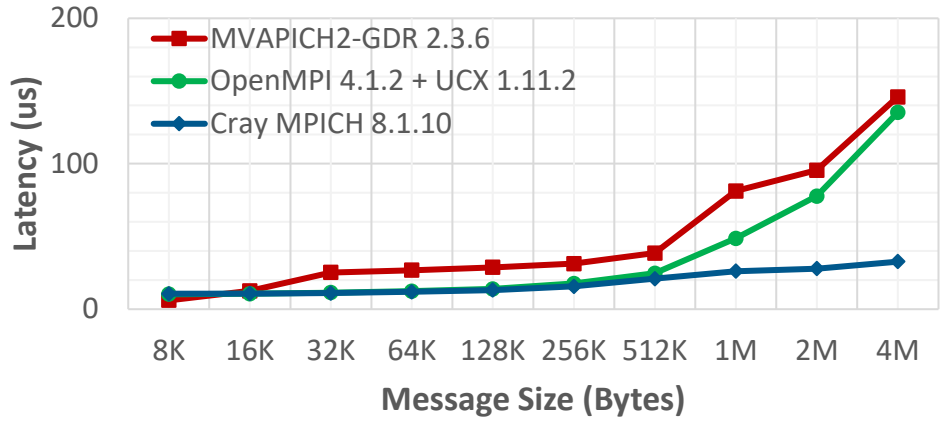
OUTLINE

- Introduction
- MPI + Slingshot
- CPU-Level Performance
- **GPU-Level Performance**
- Conclusion

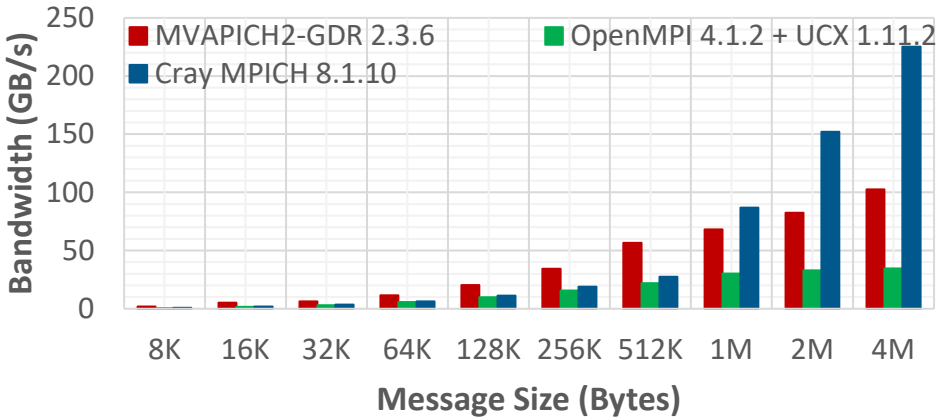
POINT-TO-POINT INTRA-NODE PERFORMANCE (GPU)

LATENCY:

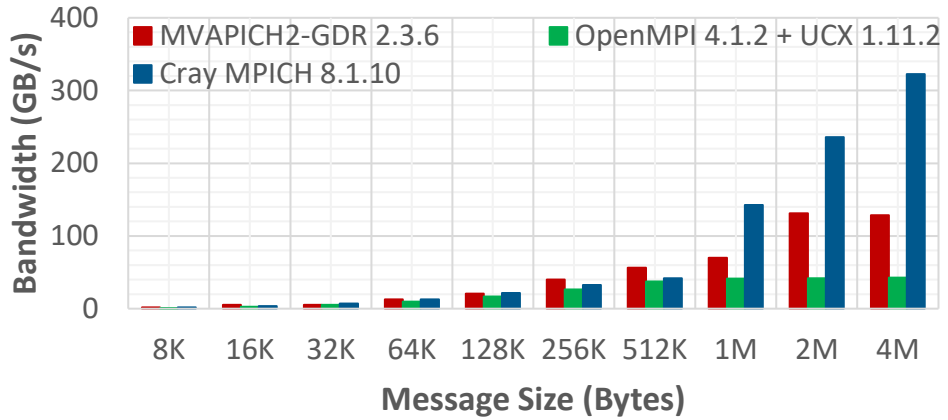




BANDWIDTH:



BI-DIRECTIONAL BANDWIDTH:



OLCF Spock Cluster – ROCm 5.0 + MI100 GPUS

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

MVAPICH2-GDR ROCm Inter-Process Communication (IPC) used in med-large message range.

PCI Bar Mapped Memory for small message sizes.

Peak Bandwidth at 4MB:

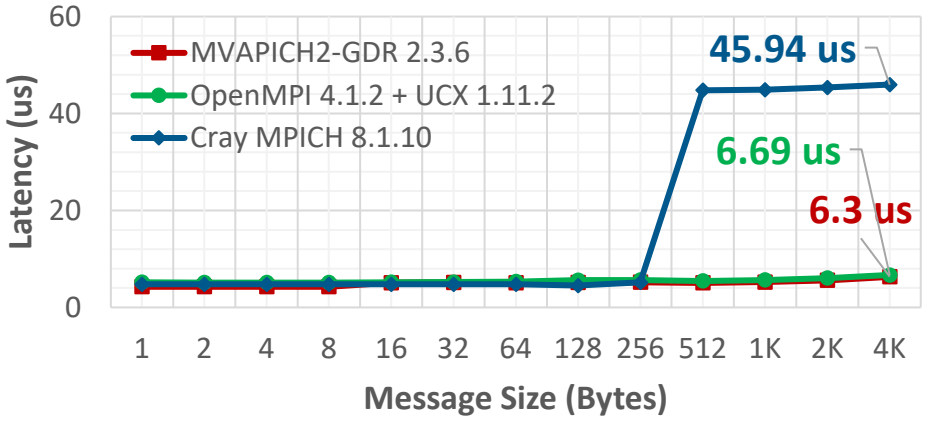
- MVAPICH2-GDR **102 GB/s**
- OpenMPI+UCX **34.5 GB/s**
- CrayMPICH **225 GB/s**

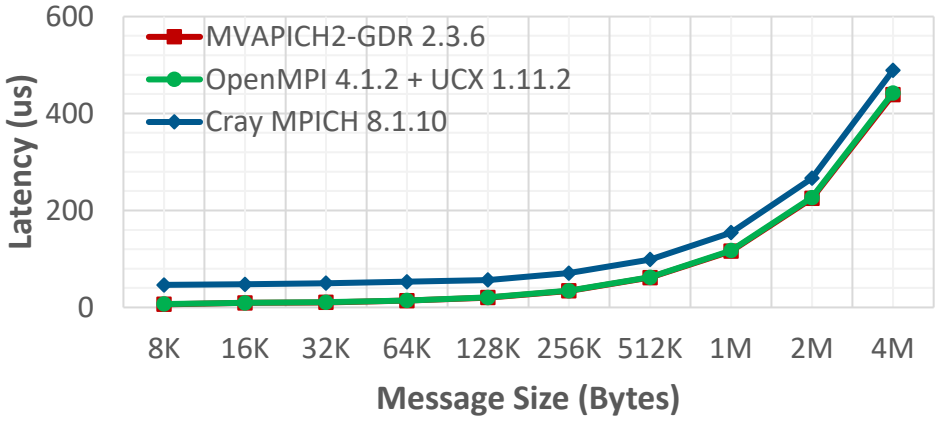
Minimum Latency at 1 Byte

- MVAPICH2-GDR **2.65 us**
- OpenMPI+UCX **3.78 us**
- CrayMPICH **3.11 us**

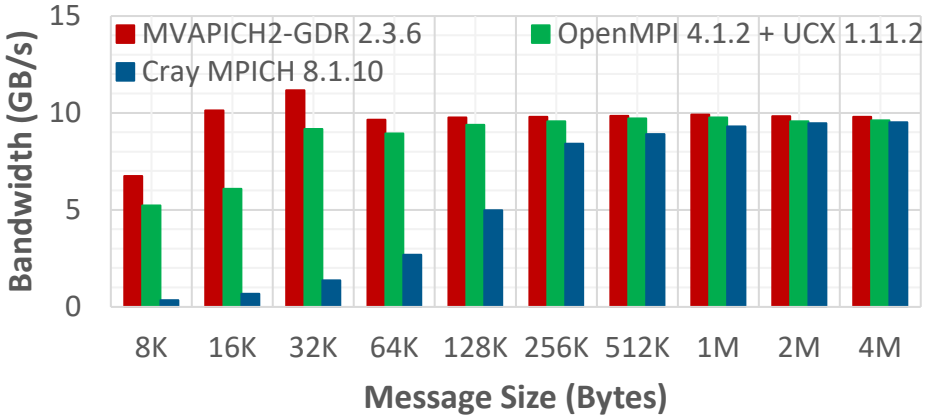
POINT-TO-POINT INTER-NODE PERFORMANCE (GPU)

LATENCY:

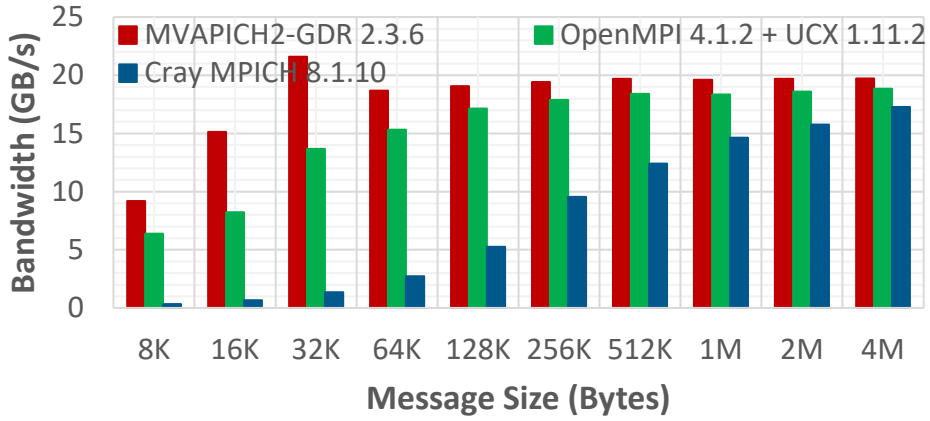




BANDWIDTH:



BI-DIRECTIONAL BANDWIDTH:



OLCF Spock Cluster – ROCm 5.0 + MI100 GPUS

MVAPICH2-GDR GPUDirect Loopback for medium message range and GPUDirect

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

Bandwidth at 4MB:

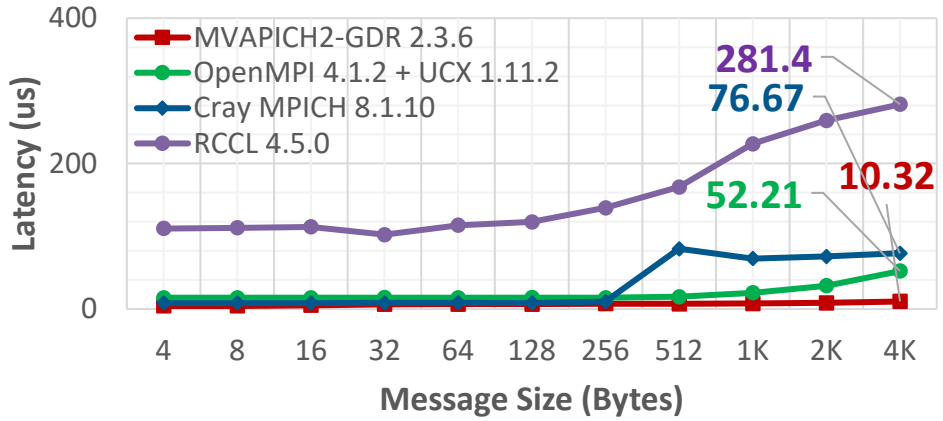
- MVAPICH2-GDR **9.8 GB/s**
- OpenMPI + UCX **9.6 GB/s**
- CrayMPICH **9.5 GB/s**

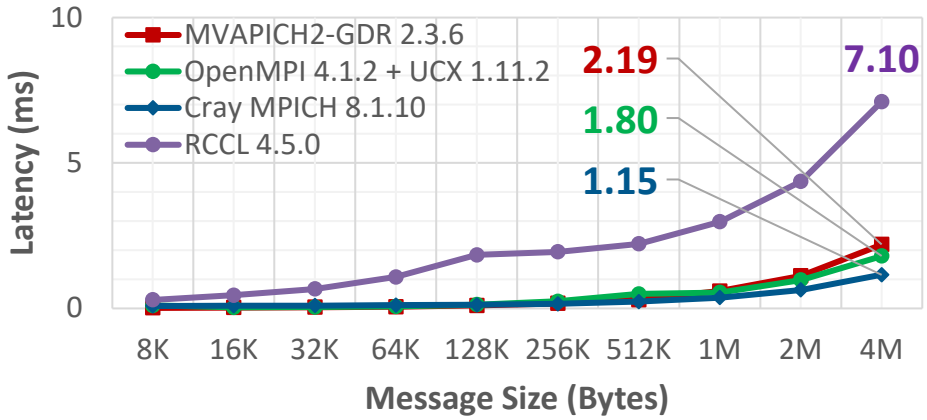
Minimum Latency at 1 Byte

- MVAPICH2-GDR **4.28 us**
- OpenMPI + UCX **5.18 us**
- CrayMPICH **4.76 us**

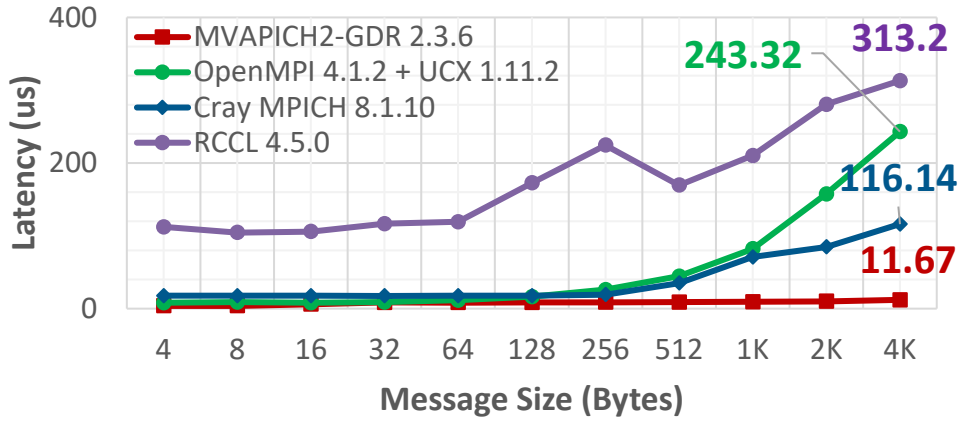
COLLECTIVES PERFORMANCE (GPU)

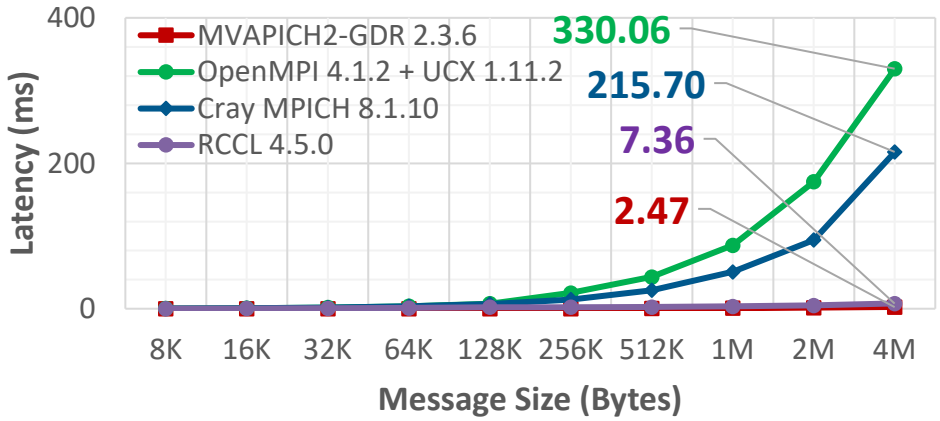
BROADCAST:





REDUCE:

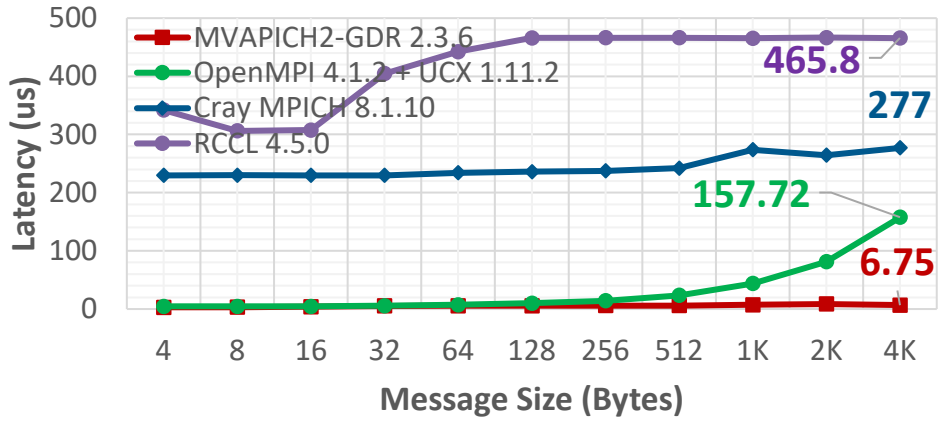


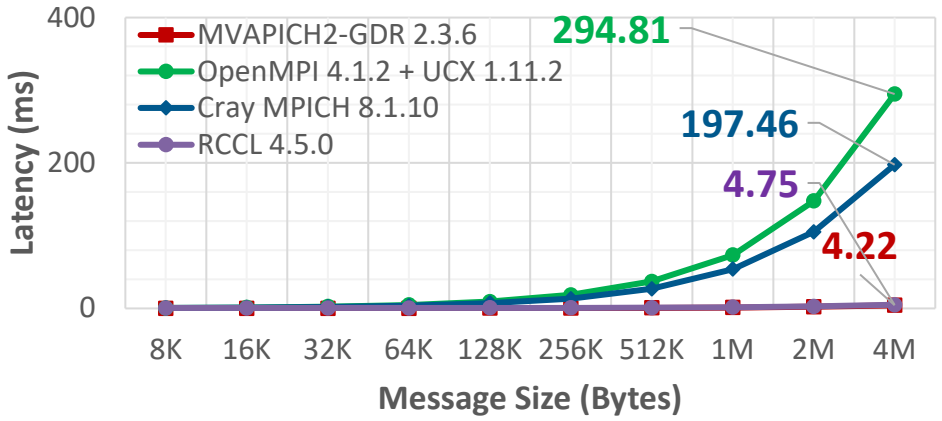


OLCF Spock Cluster – ROCm 5.0 + MI100 GPUS (4 Nodes 4 PPN – 16 GPUs)

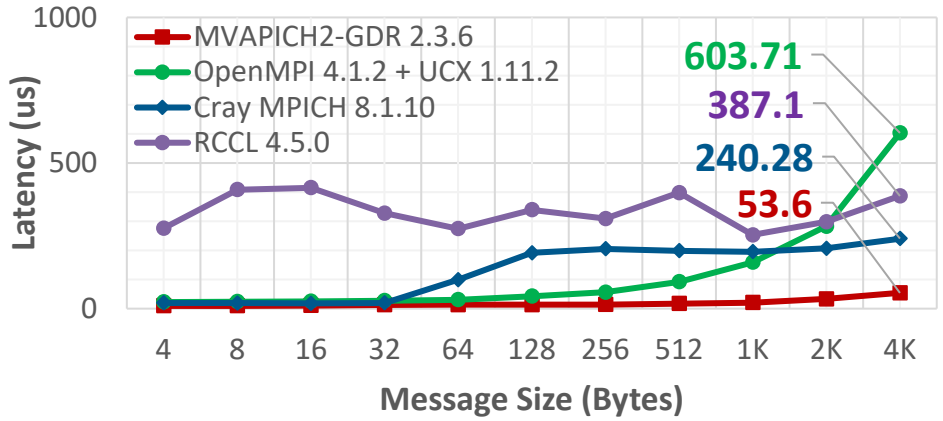
COLLECTIVES PERFORMANCE (GPU)

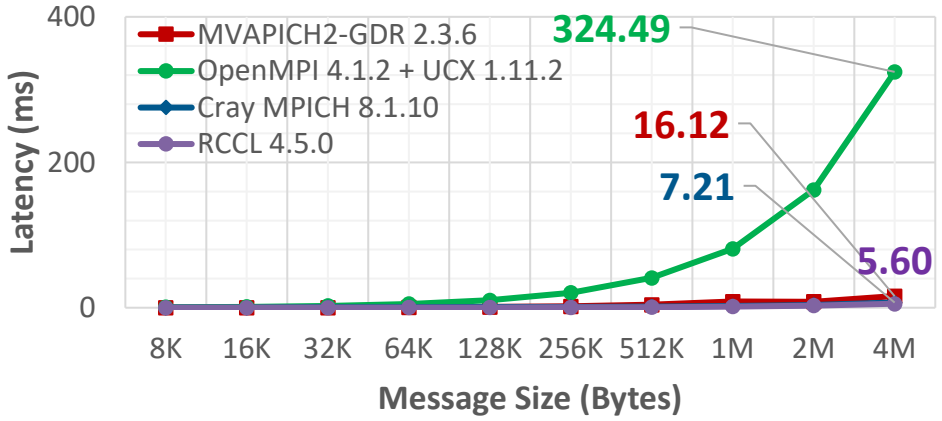
GATHER:





ALLGATHER:

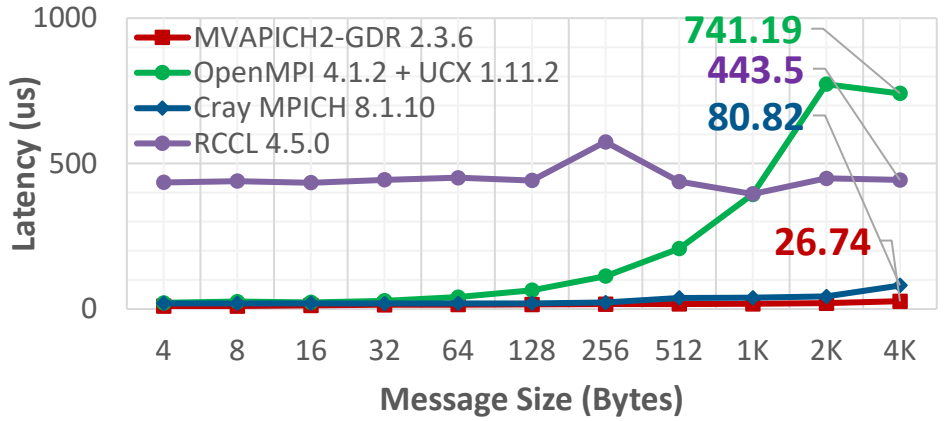


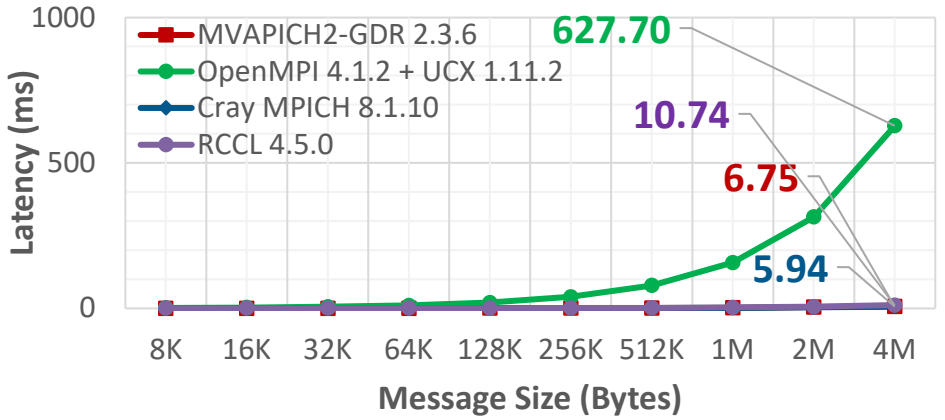


OLCF Spock Cluster – ROCm 5.0 + MI100 GPUS (4 Nodes 4 PPN – 16 GPUs)

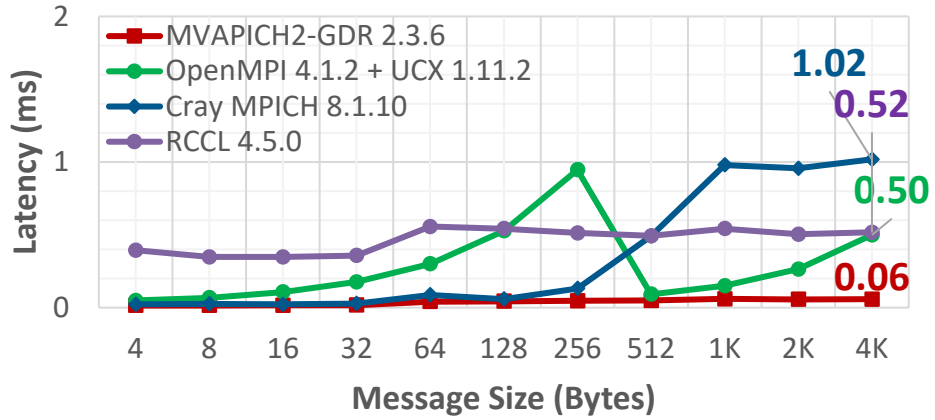
COLLECTIVES PERFORMANCE (GPU)

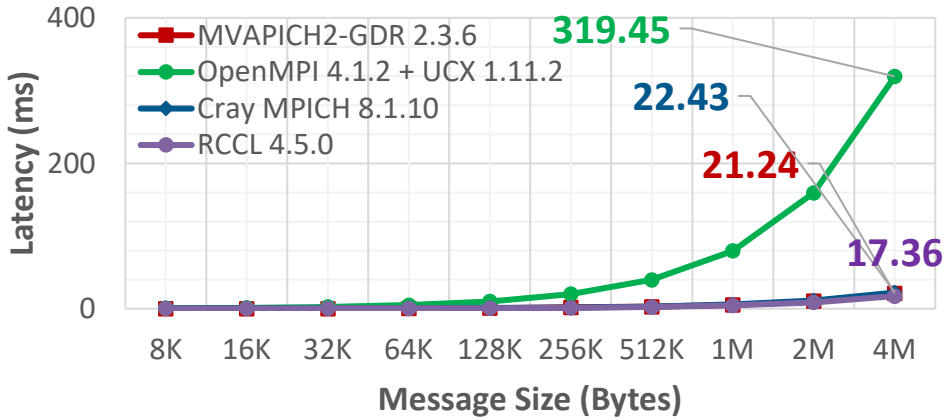
ALLREDUCE:





ALLTOALL:





OLCF Spock Cluster – ROCm 5.0 + MI100 GPUS (4 Nodes 4 PPN – 16 GPUs)

OUTLINE

- Introduction
- MPI + Slingshot
- CPU-Level Performance
- GPU-Level Performance
- **Conclusion**

RELATED PUBLICATION

Additional Information, Analysis, and Evaluation can be found in the paper below at PEARC'22:

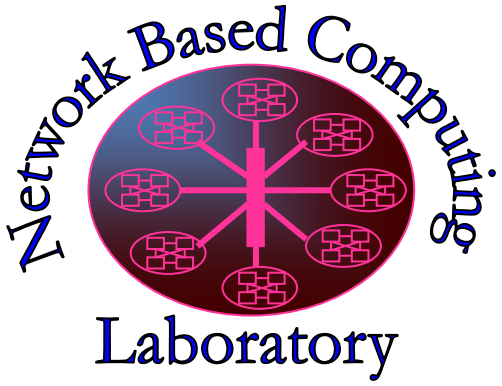
High Performance MPI over the Slingshot Interconnect: Early Experiences. K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda. Practice and Experience in Advanced Research Computing (PEARC 22), Jul 2022.

CONCLUSION AND FUTURE WORK

- The deployment of Slingshot networking on upcoming exascale systems creates a need for efficient and optimized MPI to utilize and exploit features of Slingshot.
- Performance Evaluation of MPI on Spock Early Access System with Slingshot 10 Networking.

- **Future Work:**
  - Evaluate and provide functionality of MPI on Slingshot 11
  - Optimizations and Performance can be found in upcoming releases of MVAPICH2-GDR for GPUs and MVAPICH2-X for CPUs

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/