MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

High Performance MPI Over Slingshot

Kawthar Shafie Khorassani
shafiekhorassani.1@osu.edu

SC 22 - OSU Booth

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

linkedin.com/in/kawthar-shafie

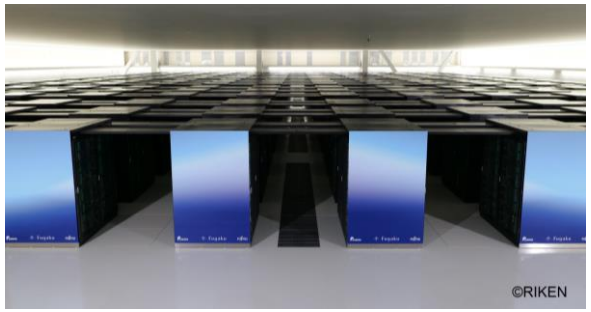
Introduction

- 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)

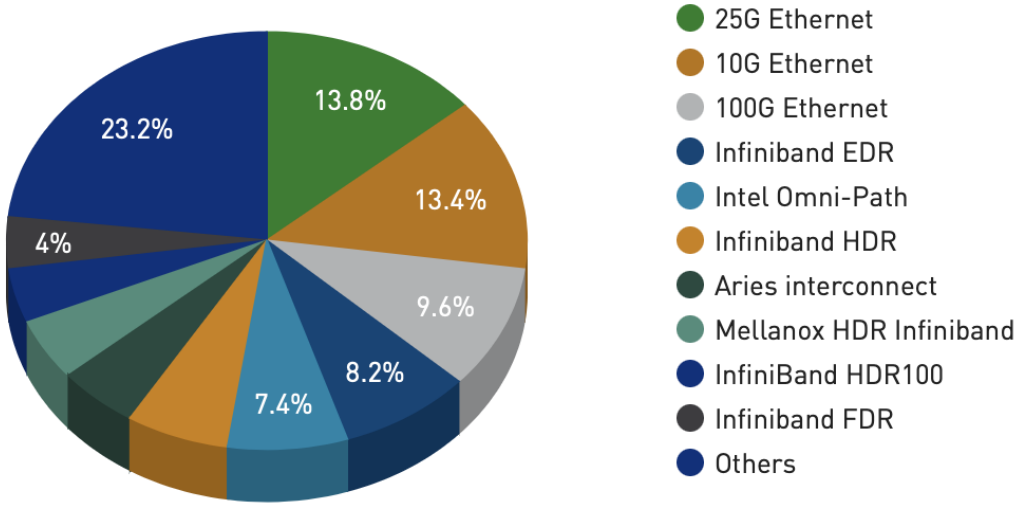


Fugaku (‘20-’21)

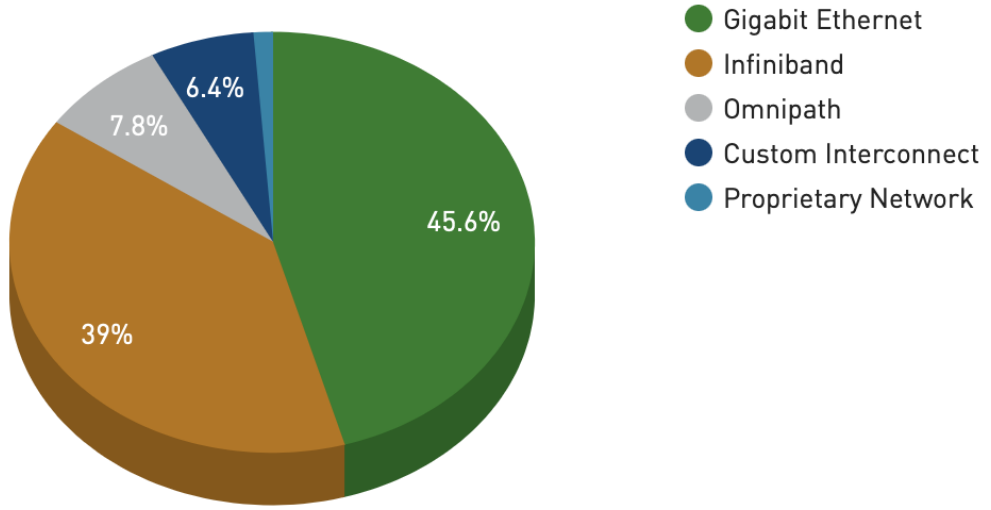
Frontier (‘22)

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

Accessibility and deployment on early access Slingshot systems:

- Ecosystem with Slingshot-10 interconnection amongst nodes

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

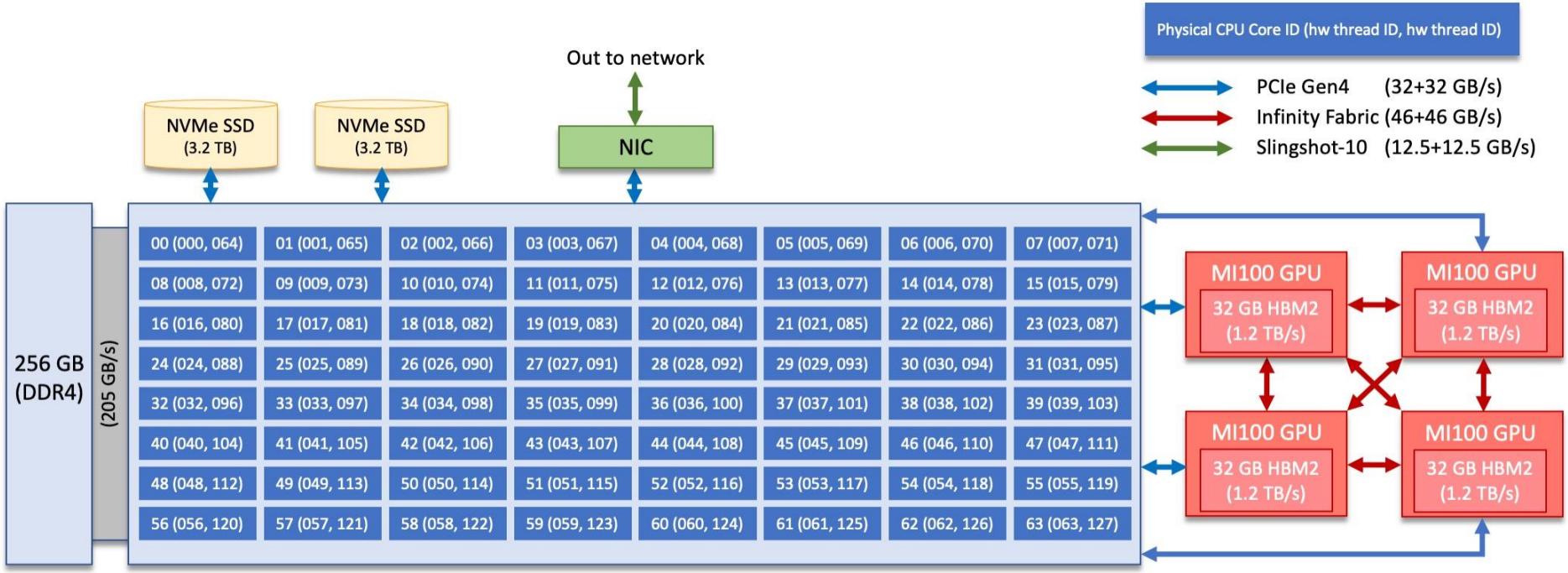
- Slingshot-11
- Deployed over a slingshot fabric and adapter

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

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

MVAPICH-PLUS

- Run: MV2_ENABLE_GPU=1

OpenMPI 4.1.4 + UCX 1.12.1

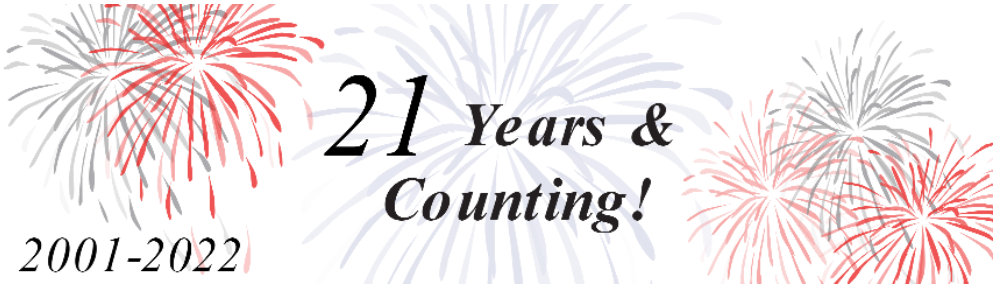
- Compile UCX: --with-rocm=<path-to-rocm> --without-knem –without-cuda --enable-optimizations
- Compile OpenMPI: --with-ucx=<path-to-ucx> --without-verbs
- Run: -x UCX_RNDV_THRESH=128

RCCL 5.0.2

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

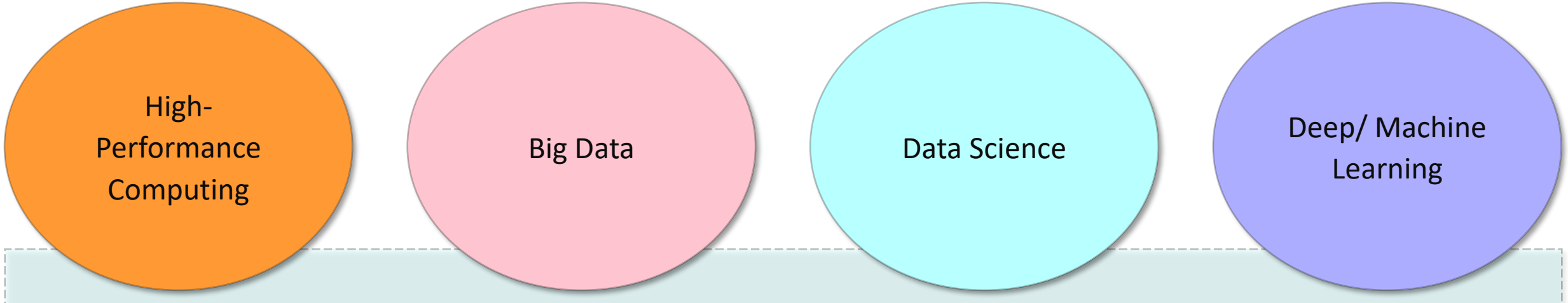
Overview of the MVAPICH Project

- 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 3,290 organizations in 90 countries
- More than 1.63 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '22 ranking)
  - 7th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 19th, 448, 448 cores (Frontera) at TACC
  - 34th, 288,288 cores (Lassen) at LLNL
  - 46th, 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 19th ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

One Runtime to Rule them all!



MVAPICH-Plus

(Support for all combinations of CPU, Interconnect, Accelerator, DPU)

Advanced HPC Hardware

Interconnect Technologies	Processor Technologies	Accelerator Technologies	Network Offload
InfiniBand, Omni-Path, Ethernet, Slingshot 10/11, OPX, Broadcom, Rockport	x86 (Intel/AMD), ARM, OpenPOWER	GPUs (NVIDIA/AMD), FPGAs	Datacenter Processing Units, Switch Offload, Network Adapter Offload

MVAPICH-Plus

- Released on 11/11/2022
- Based on MVAPICH 3.0
- Advanced MPI with unified MVAPICH2-GDR and MVAPICH2-X features
- Support for NVIDIA and AMD GPUs
- Optimized designs for HPC, DL, ML, Big Data and Data Science applications
- Added support for the ch4:ucx and ch4:ofi devices
- Added support for the Cray Slingshot 11 interconnect over OFI
  - Supports Cray Slingshot 11 network adapters
- Added support for the Cornelis OPX library over OFI
  - Supports Intel Omni-Path adapters
- Added support for the Intel PSM3 library over OFI
  - Supports Intel Columbiaville network adapters
- Added support for IB verbs over UCX
  - Supports IB and RoCE network adapters

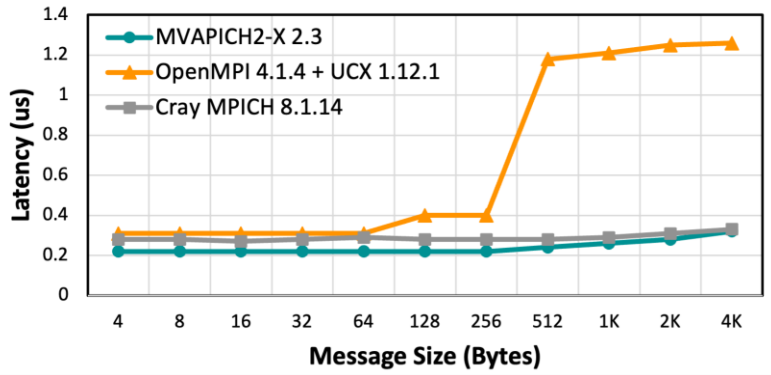
Features of OFI and UCX Support

- Support a broad range of interconnects with widely used libraries
  - Configure with `--with-device=ch4:ofi` or `--with-device=ch4:ucx`
- Runtime provider selection via CVARs
  - `MPIR_CVAR_OFI_USE_PROVIDER=<prov>`
- System default, embedded, or custom installation of OFI/UCX
  - Configure with `--with-libfabric=embedded` or `--with-libfabric=<path>`
  - Configure with `--with-ucx=embedded` or `--with-ucx=<path>`
- Enhanced MVAPICH2 collective designs

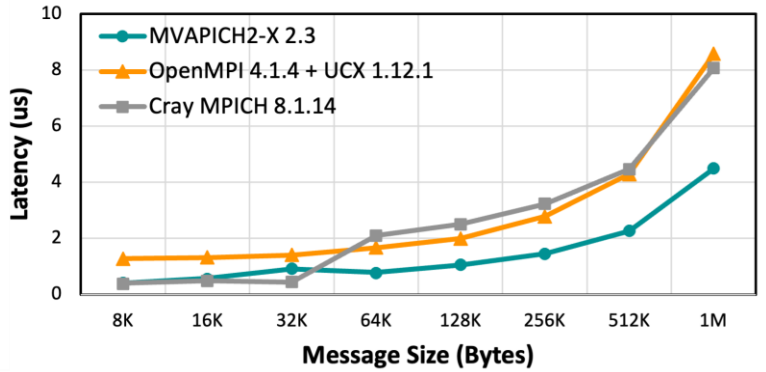
Performance Evaluation

CPU

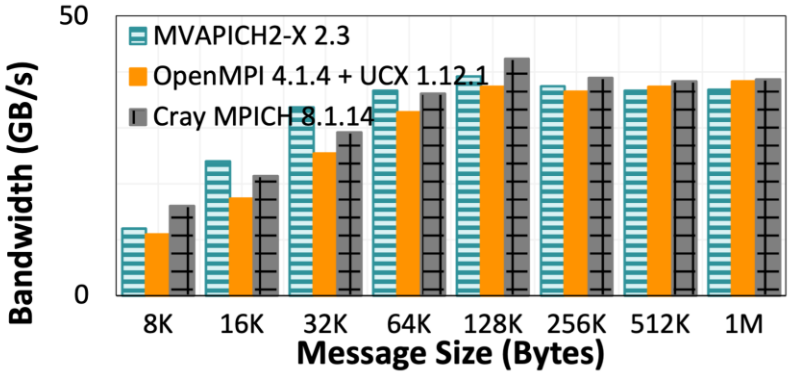
Point-to-Point Performance - Intra-Node CPU



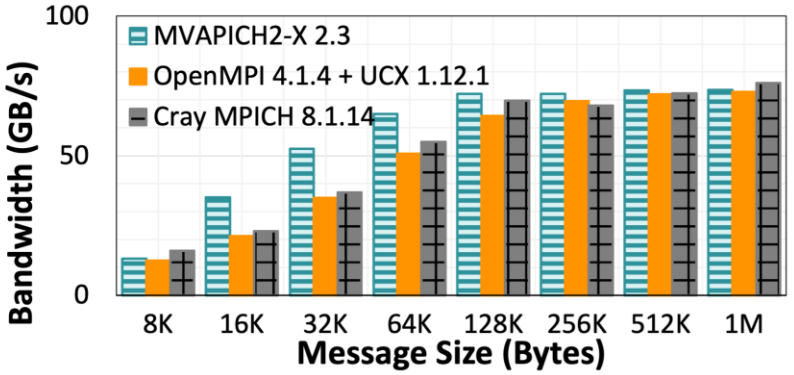
Latency (small messages)



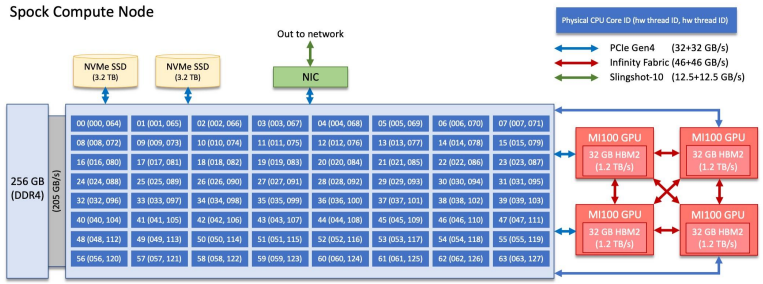
Latency (large messages)



Bandwidth



Bi-Directional Bandwidth



Peak Bandwidth:

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

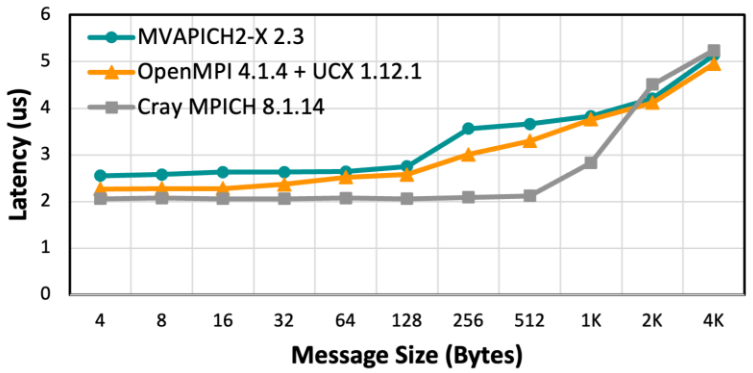
Latency at 4 Bytes:

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

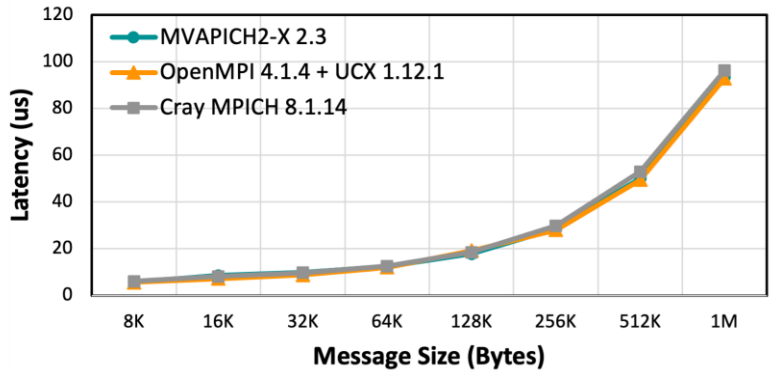
AMD Epyc Rome CPUs on Spock System

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.

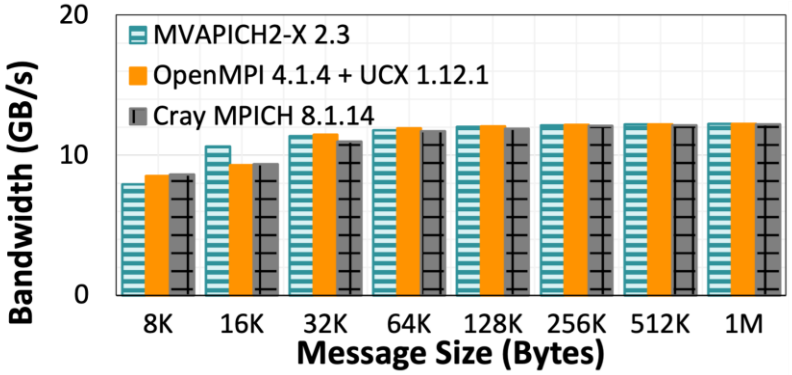
Point-to-Point Performance - Inter-Node CPU



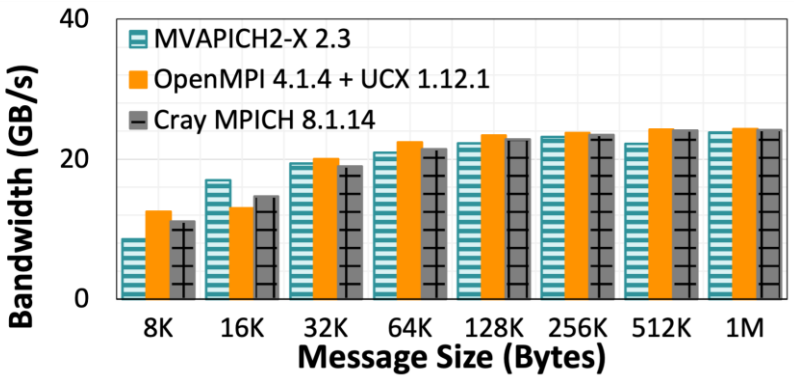
Latency (small messages)



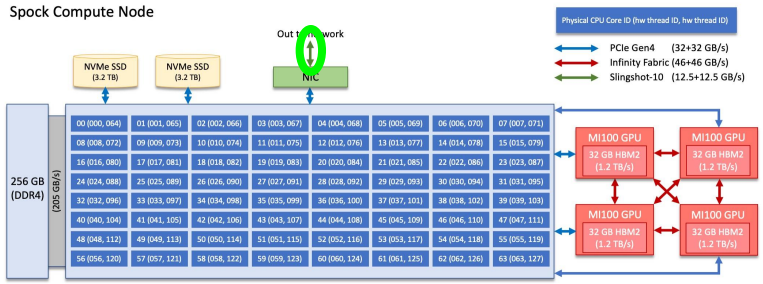
Latency (large messages)



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:

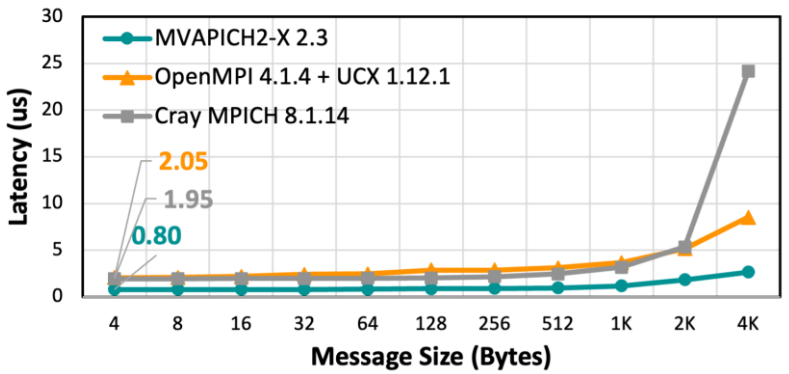
- MVAPICH2-X **2.55 us**
- OpenMPI+UCX **2.27 us**
- CrayMPICH **2.07 us**

AMD Epyc Rome CPUs on Spock System

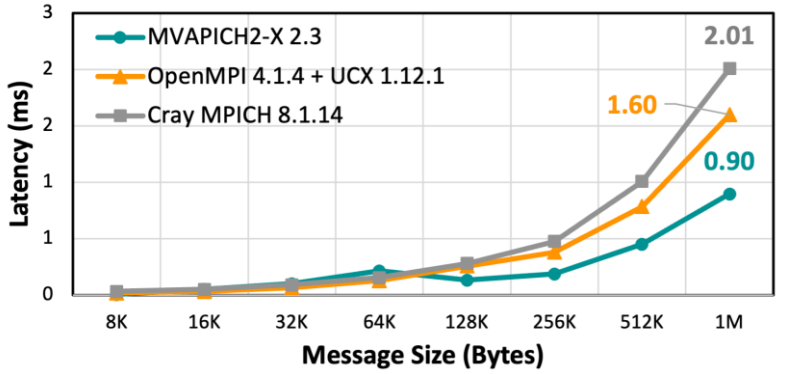
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Collectives Performance - CPU

REDUCE

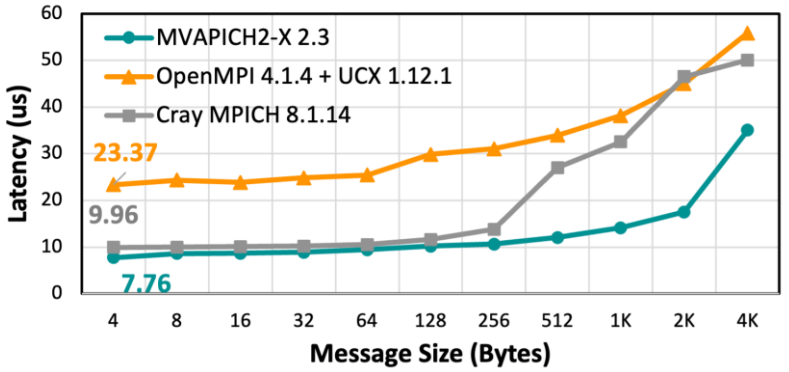


Reduce (small messages)

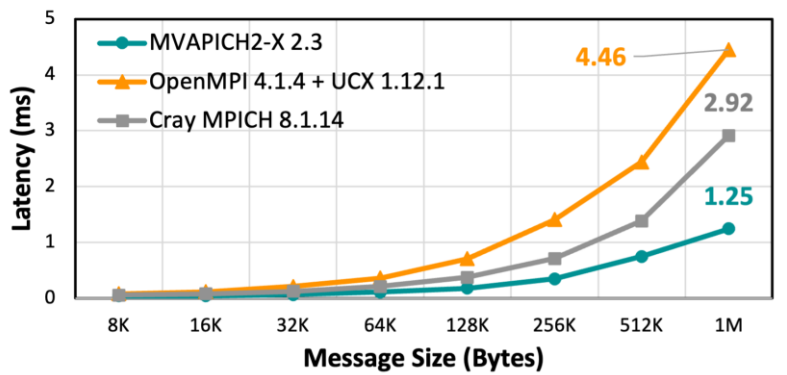


Reduce (large messages)

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

GPU-Aware (CUDA/ROCm) MPI Library: MVAPICH-GPU

- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

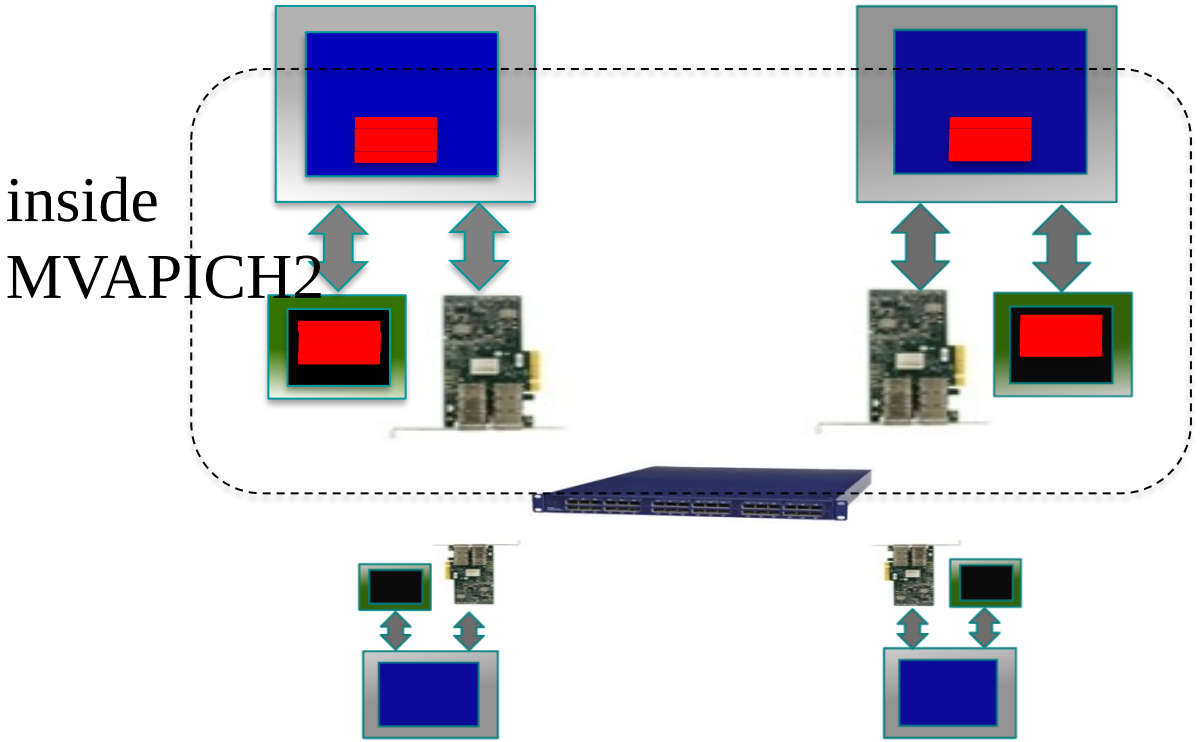
At Sender:

```
MPI_Send(s_devbuf, size, …);
```

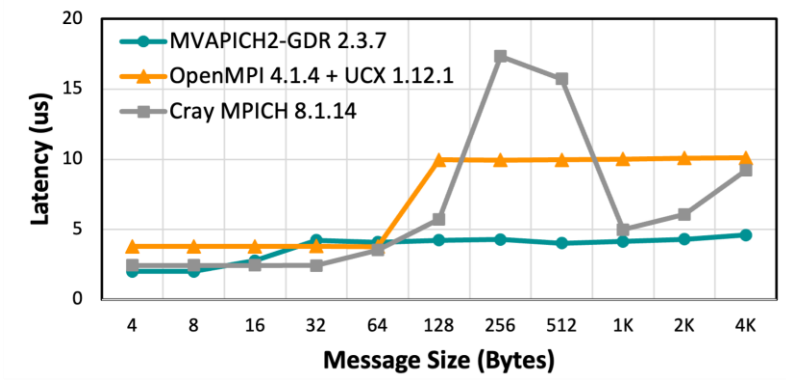
At Receiver:

```
MPI_Recv(r_devbuf, size, …);
```

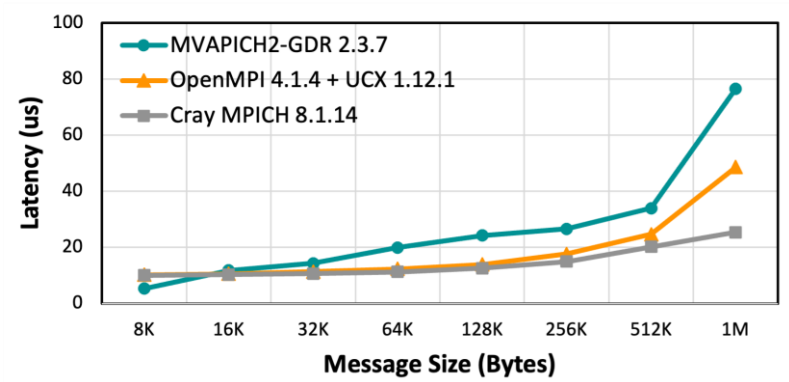
High Performance and High Productivity



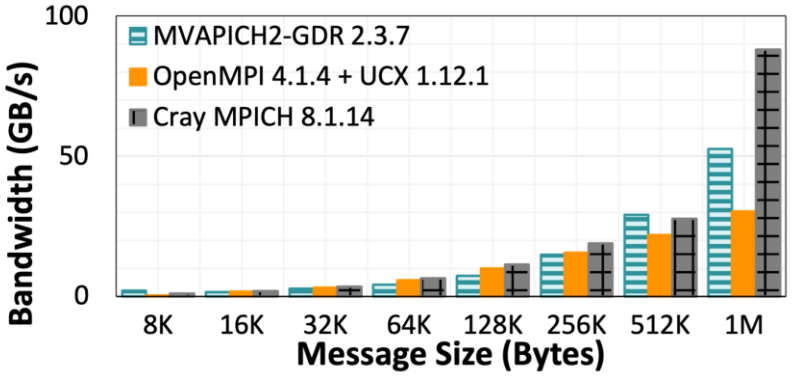
Point-to-Point Performance - Intra-Node GPU



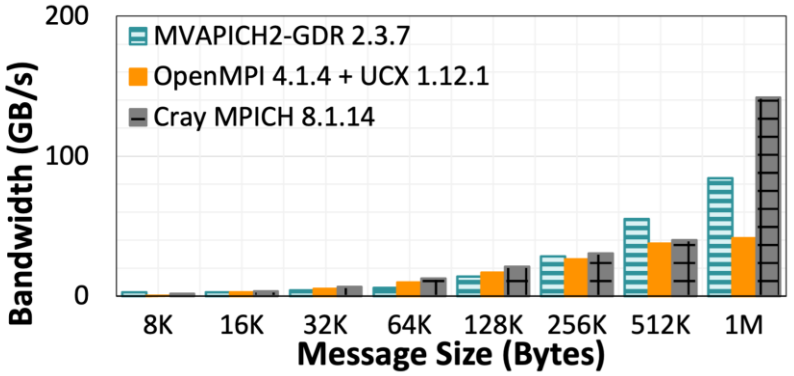
Latency (small messages)



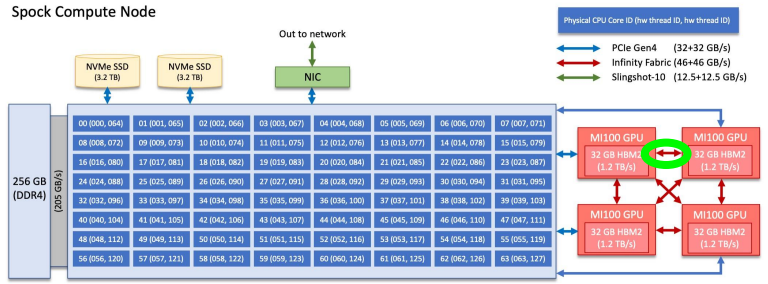
Latency (large messages)



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**

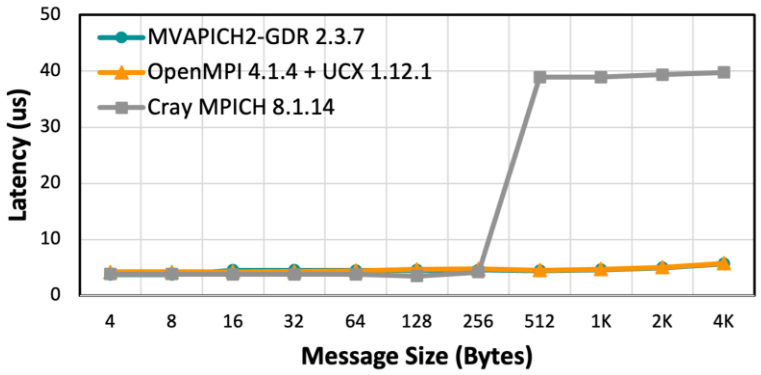
Latency at 4 Bytes:

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

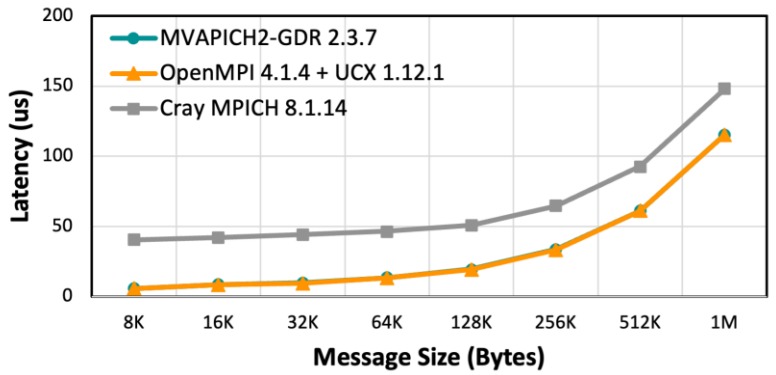
MI100 GPUs on Spock System

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.

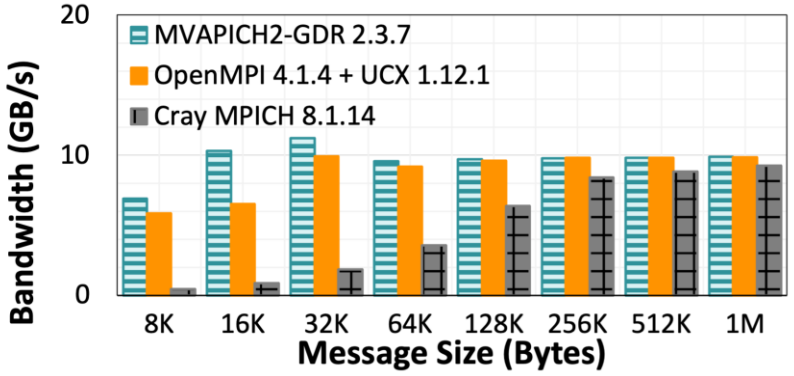
Point-to-Point Performance - Inter-Node GPU



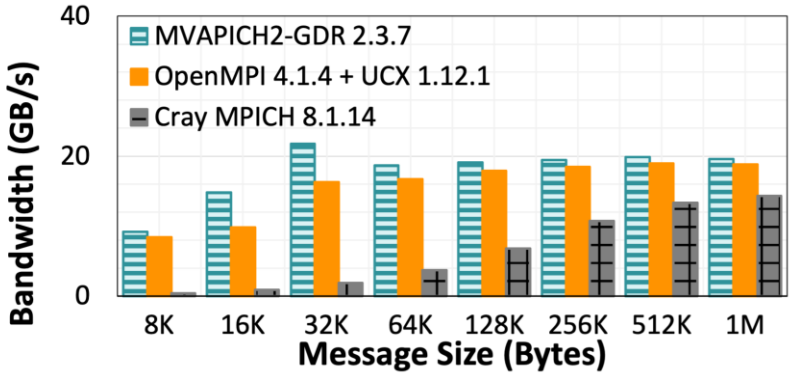
Latency (small messages)



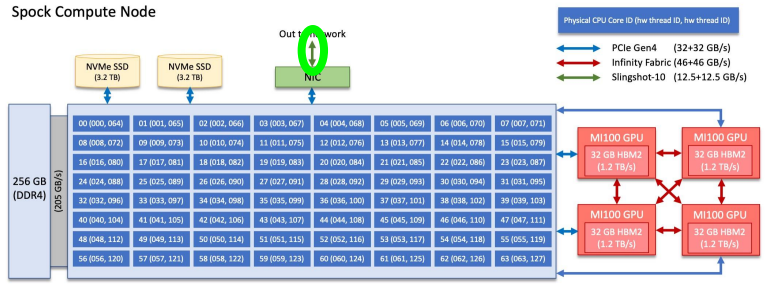
Latency (large messages)



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:

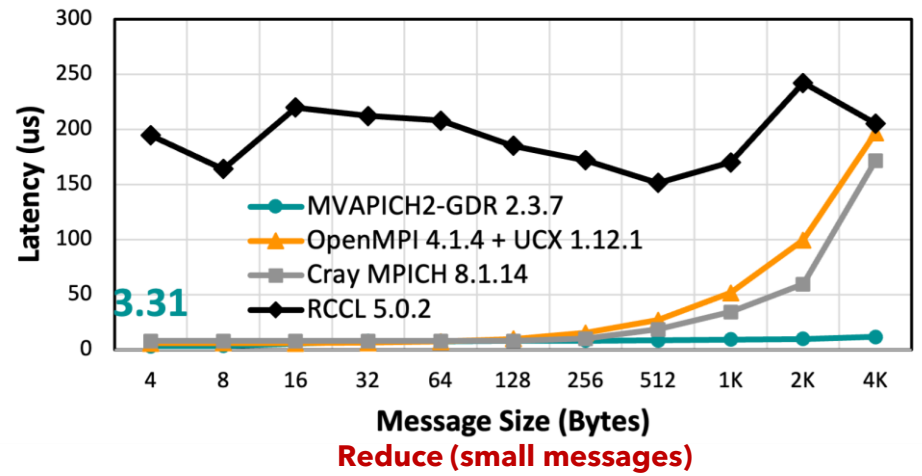
- MVAPICH2-GDR **3.73 us**
- OpenMPI+UCX **4.23 us**
- CrayMPICH **3.8 us**

MI100 GPUs on Spock System

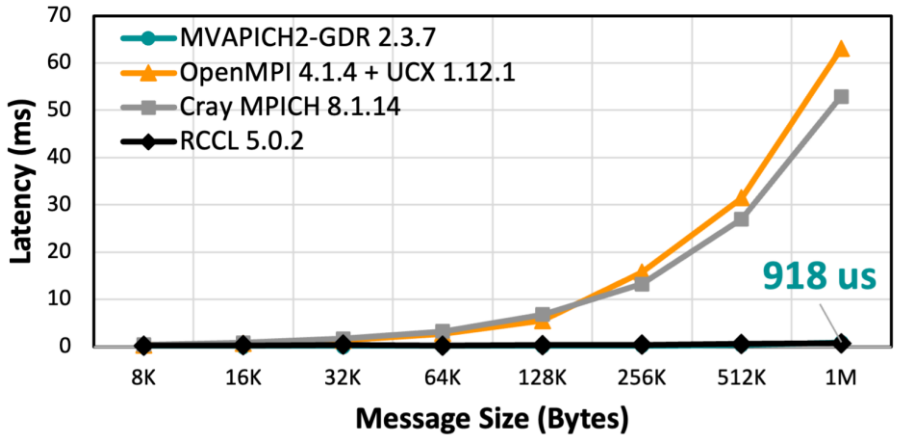
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Collectives Performance - GPU

REDUCE

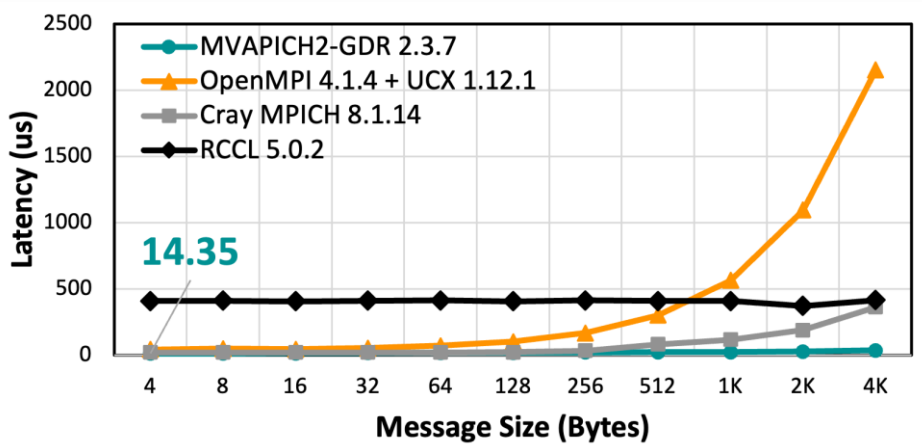


Reduce (small messages)

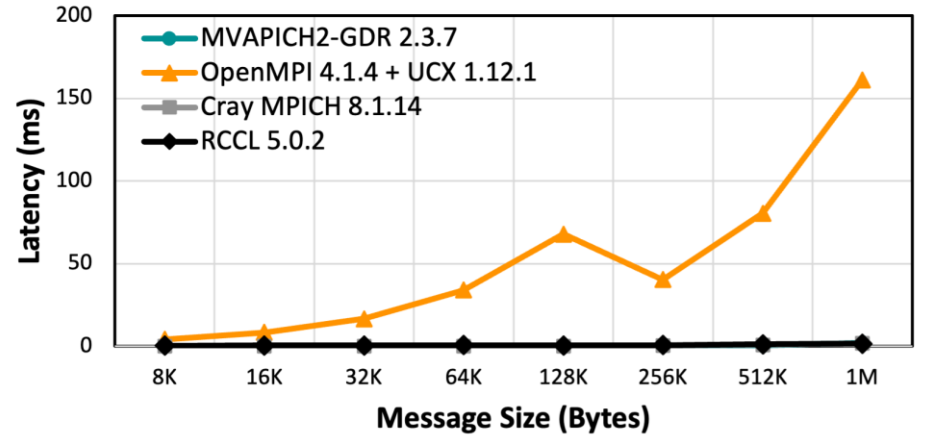


Reduce (large messages)

ALLREDUCE



Allreduce (small messages)



Allreduce (large messages)

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

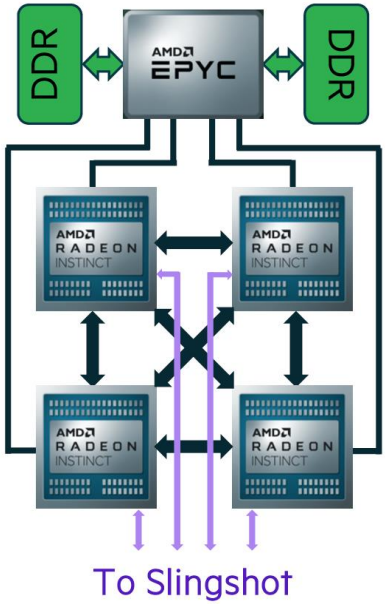
Reference: High Performance MPI over the Slingshot Interconnect: Early Experiences K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda Practice and Experience in Advanced Research Computing, Jul 2022.

Performance Evaluation Slingshot-11

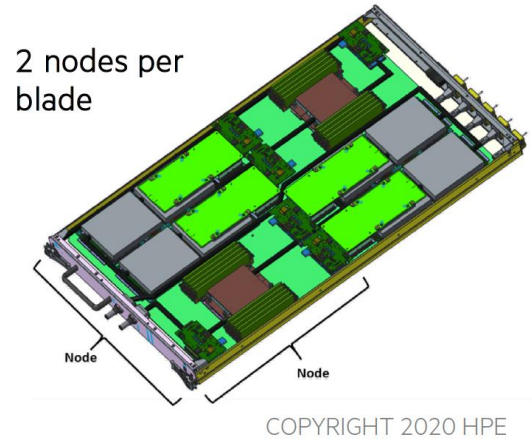
CPU

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.



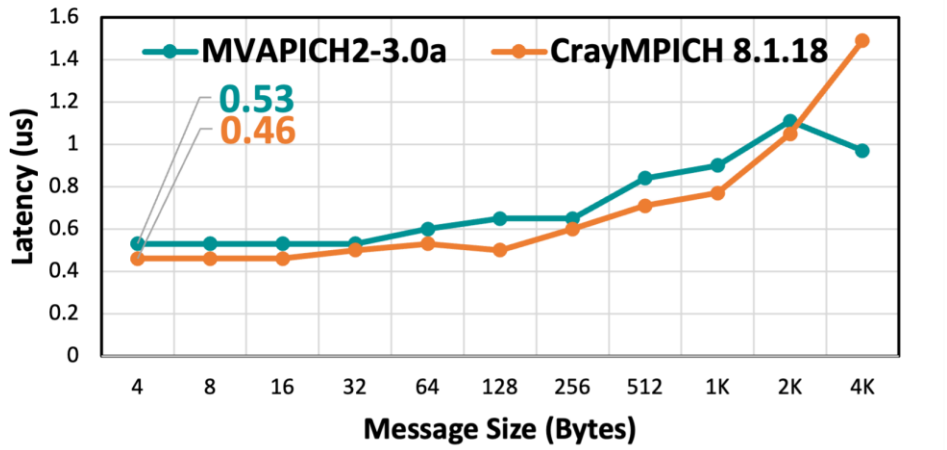
AMD GPU (ORNL)

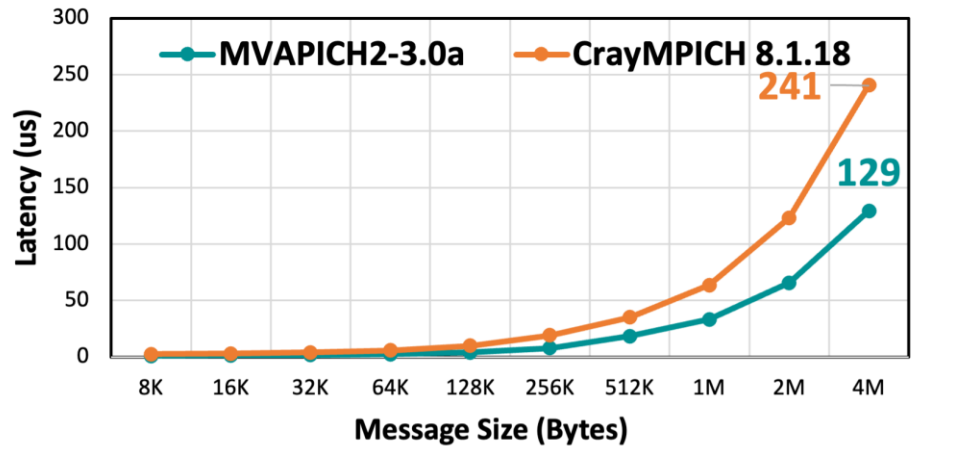


Reference: https://www.olcf.ornl.gov/frontier/

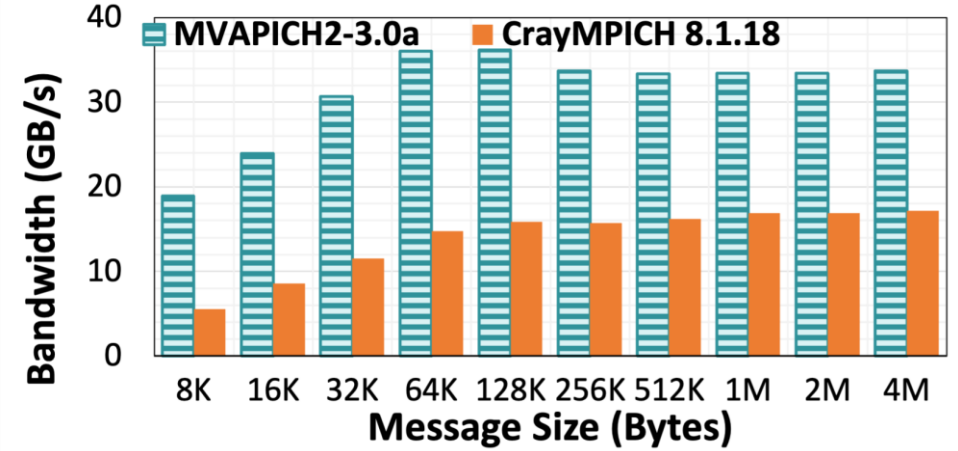
Point-to-Point Performance - Intra-Node CPU

LATENCY

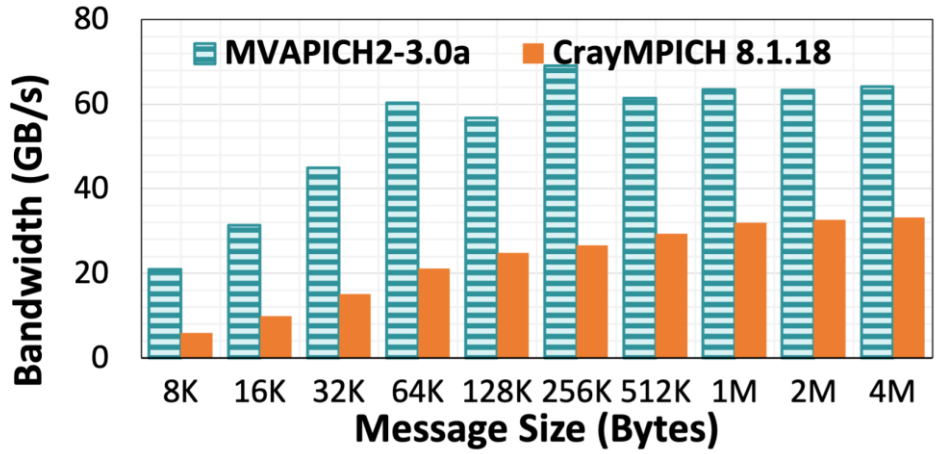




BANDWIDTH



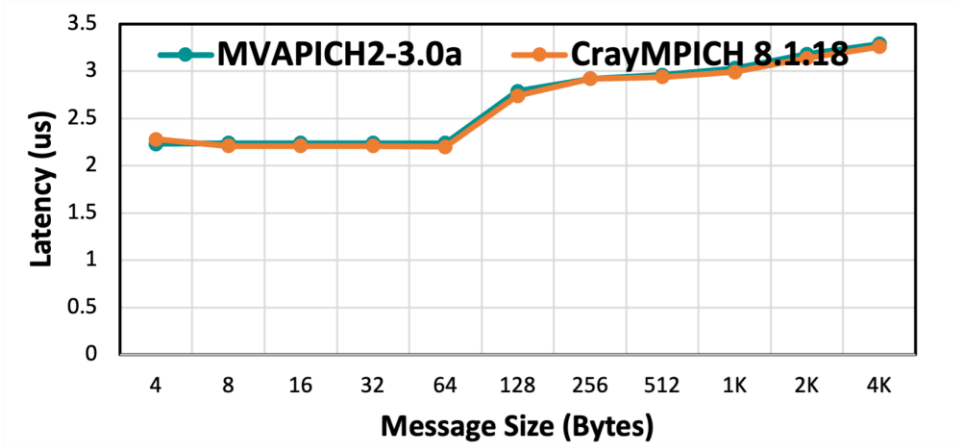
BI-BANDWIDTH

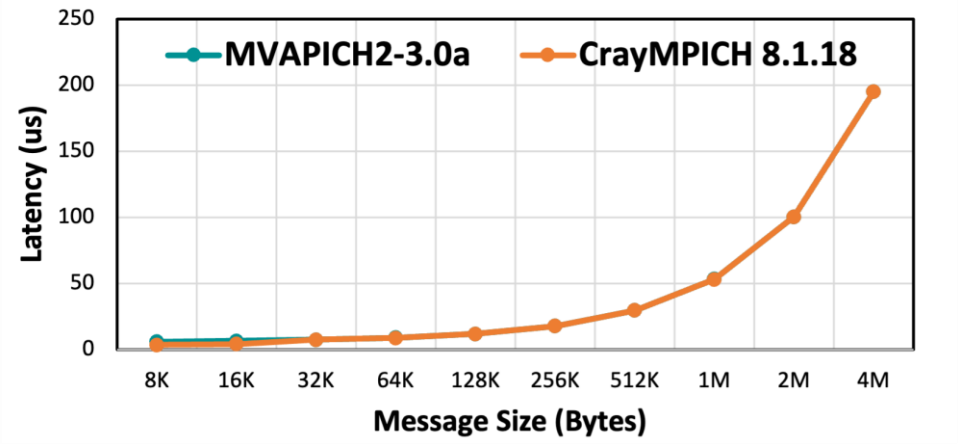


System with Slingshot-11 Networking

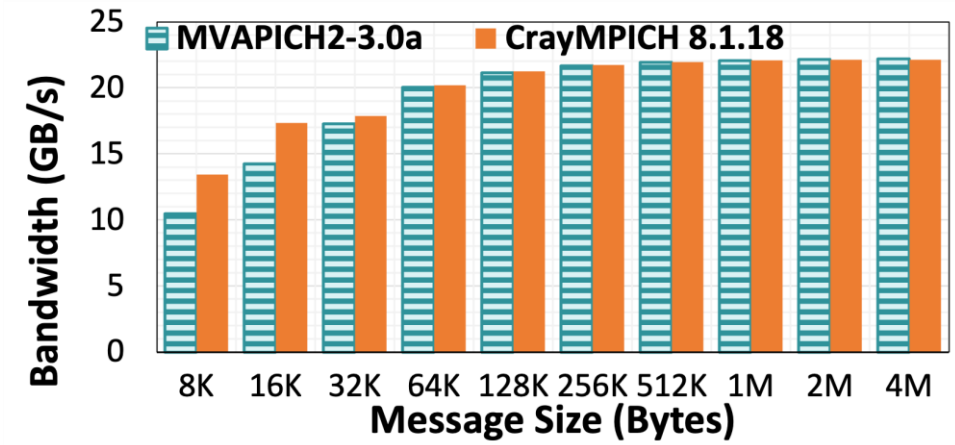
Point-to-Point Performance - Inter-Node CPU

LATENCY

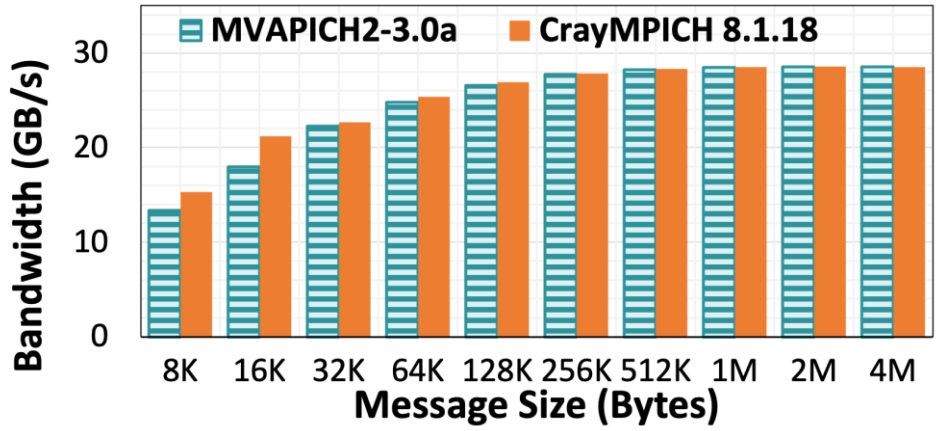




BANDWIDTH



BI-BANDWIDTH



System with Slingshot-11 Networking

Performance Evaluation Slingshot-11

GPU

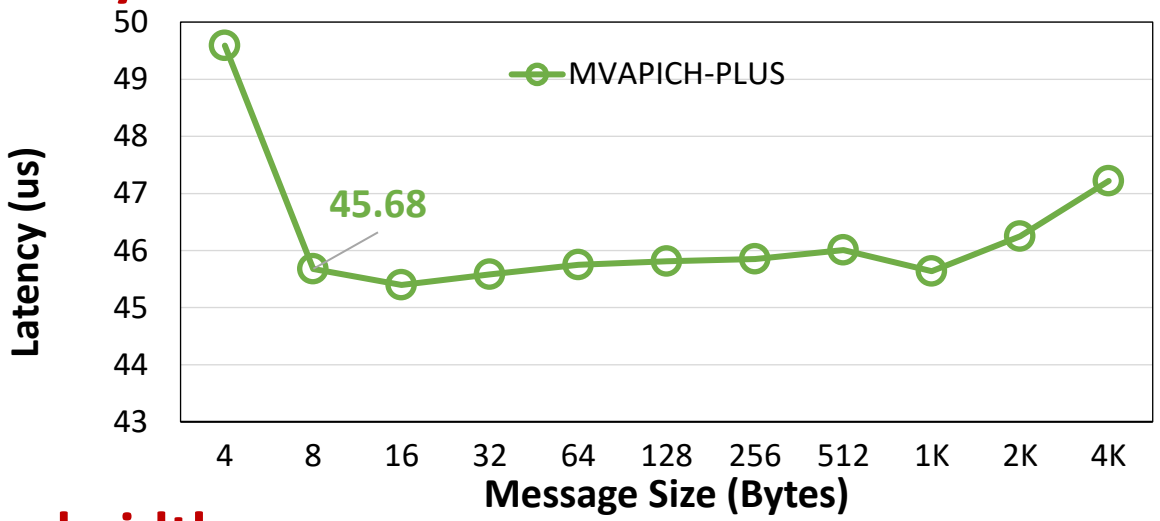
Tioga Compute Node

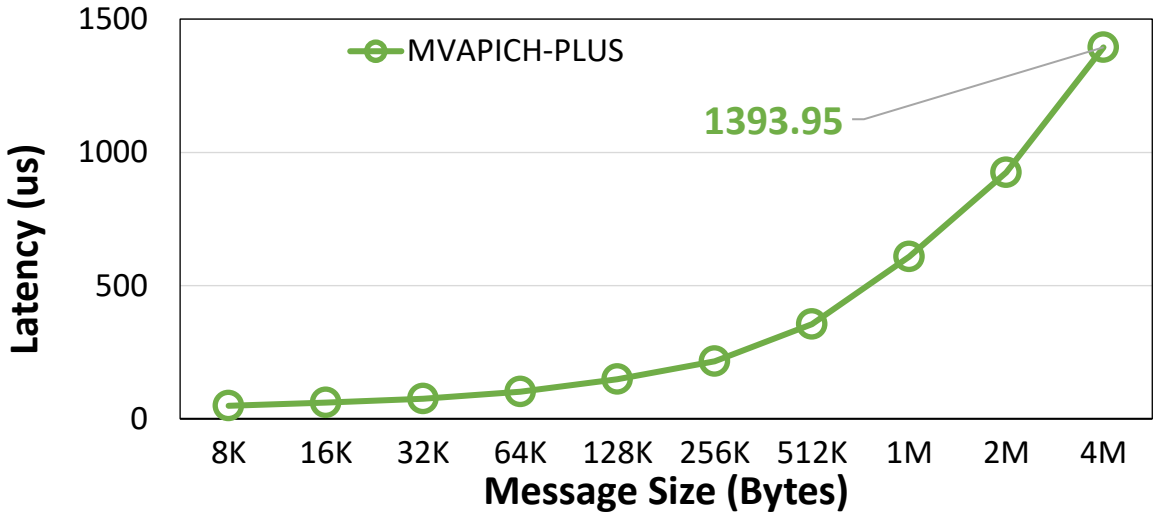
- CPU Architecture: AMD Trento 64 Cores/Nodes
- GPU Architecture: AMD MI-250X
- **CPU-GPU Interconnect:** AMD Infinity Fabric
- **System Interconnect:** HPE Slingshot-11

Reference: https://hpc.llnl.gov/hardware/compute-platforms/tioga

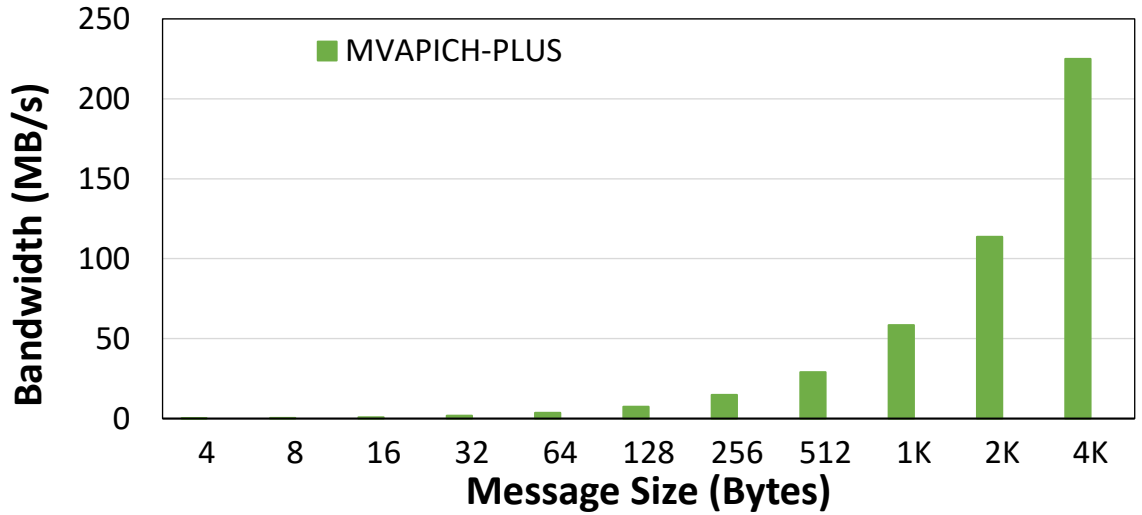
Point-to-Point Inter-Node Performance

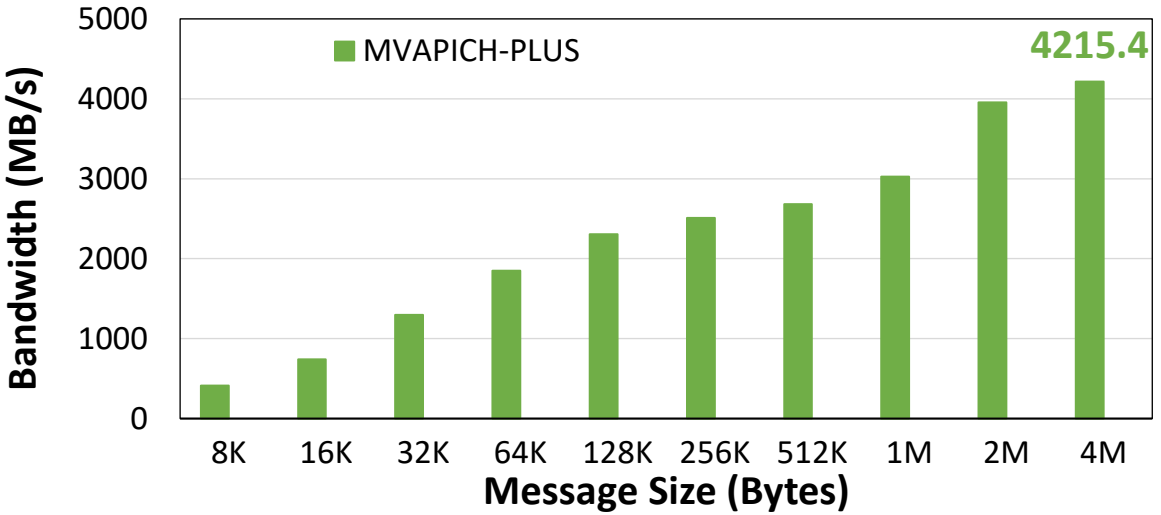
Latency:





Bandwidth:

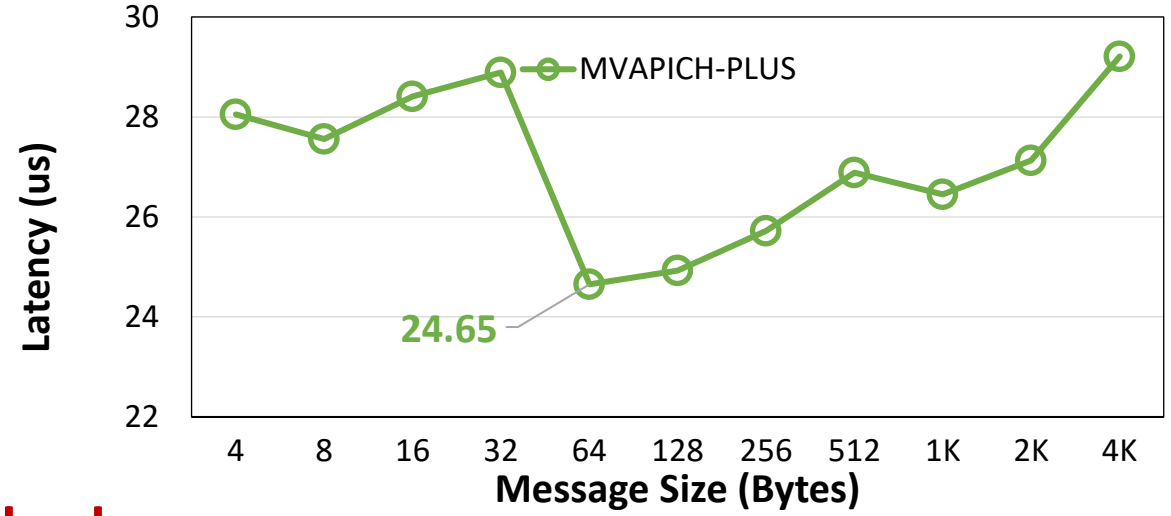


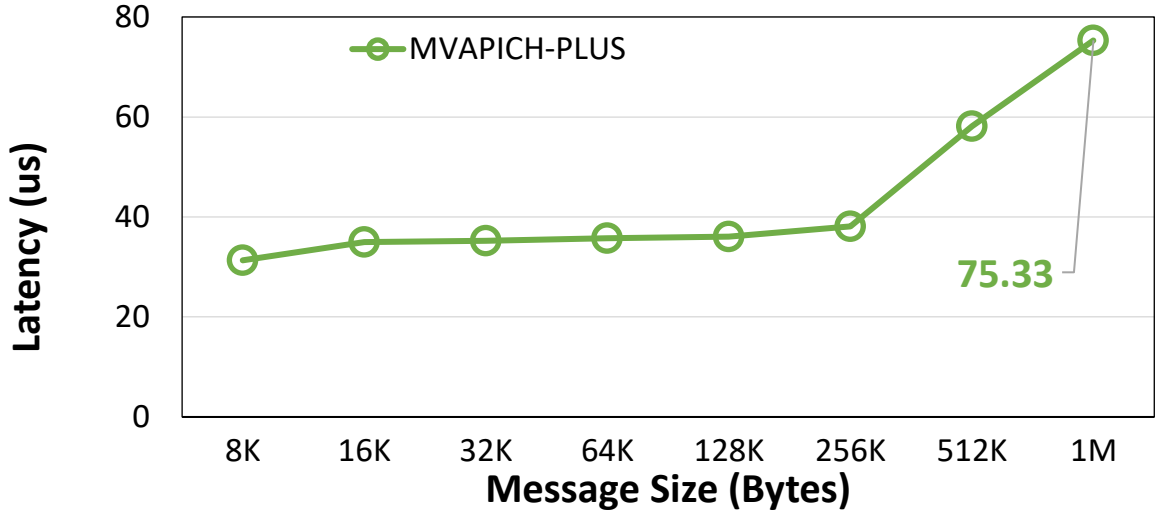


Tioga - ROCm-5.3.0 (MI250-X GPUs)

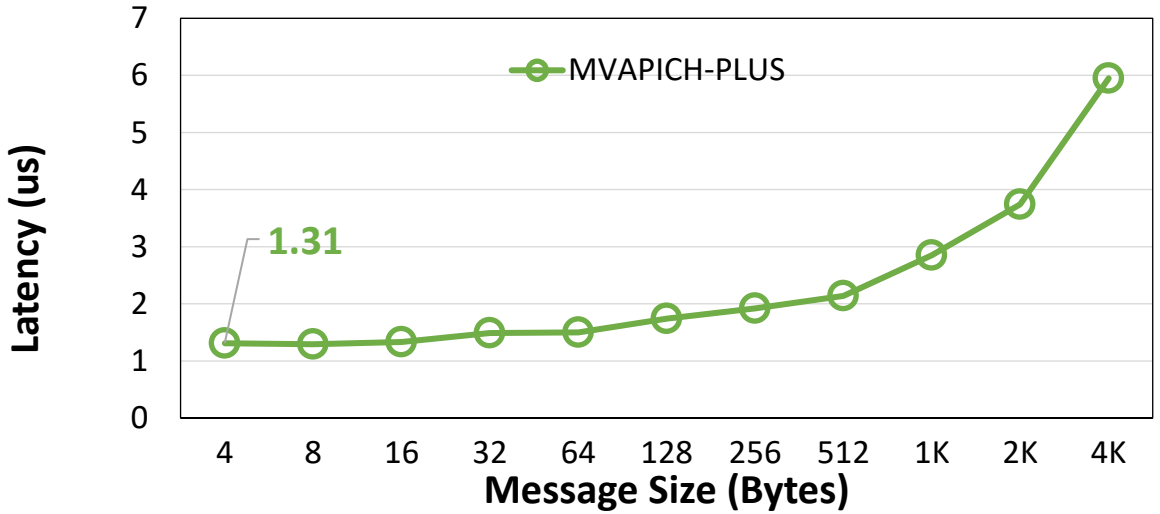
Collective Performance

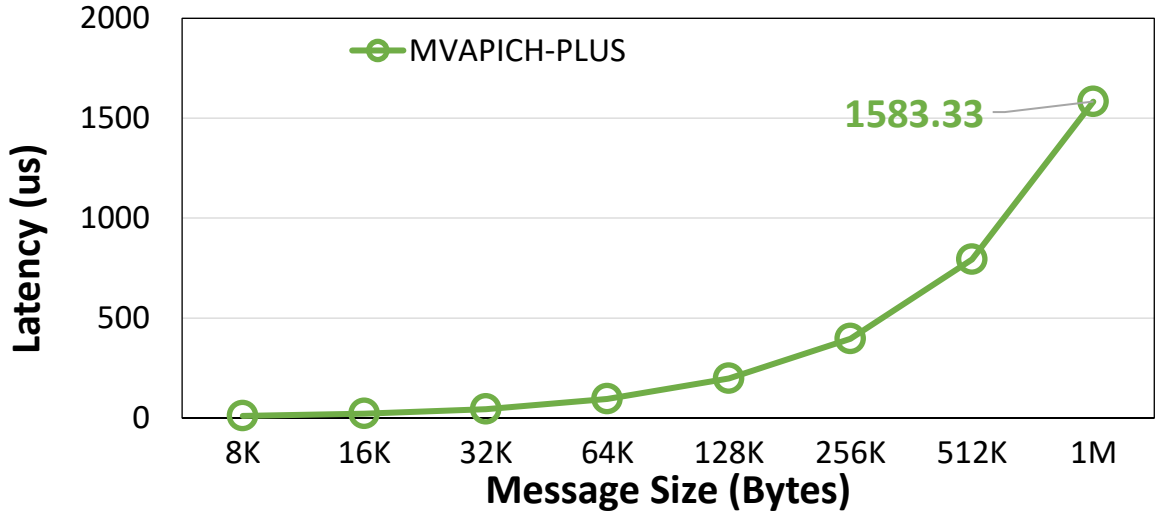
Reduce:





Allreduce:





Tioga - ROCm-5.3.0 (MI250-X GPUs) – 8 GPUs

References

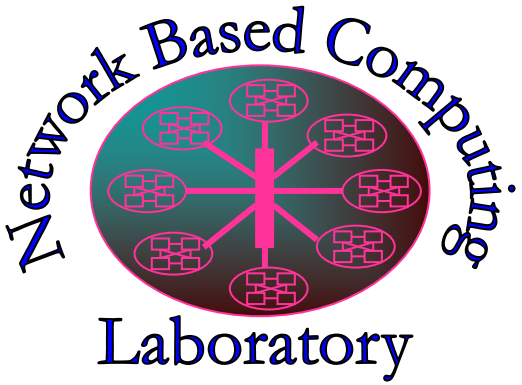
Performance Evaluation of MPI on the Slingshot Interconnect. K. Khorassani, H. Subramoni, D. Panda. OpenFabrics Alliance 2022 Workshop (OFA 2022)., April 2022.

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. [Best Student Paper Award]

High Performance MPI Over Slingshot. K. Khorassani. MVAPICH User Group Conference (MUG 2022), August 2022.

High Performance MPI over the Slingshot Interconnect. K. Khorassani, C. Chen, B. Ramesh, A. Shafi, H. Subramoni, D. Panda. *Accepted at Journal of Computer Science and Technology Special Issue, October 2022.*

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/

Several Talks at OSU Booth #4035

Mvapich.cse.ohio-state.edu