MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

HiBD

High-Performance Big Data

HiDL

High-Performance Deep Learning

Performance Evaluation and Optimization of MVAPICH-Plus on SDSC Cosmos: Early Experience

- **Goutham Kalikrishna Reddy Kuncham**, Siyuan Zhang

08/18/2025

Network-based Computing Laboratory

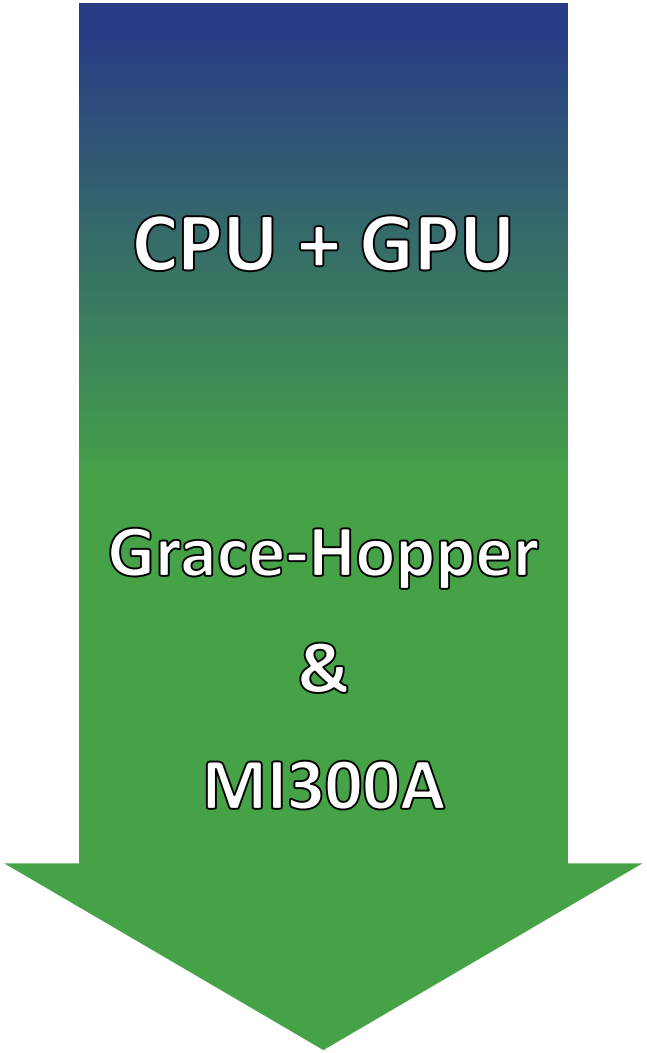
Department of Computer Science and Engineering

The Ohio State University

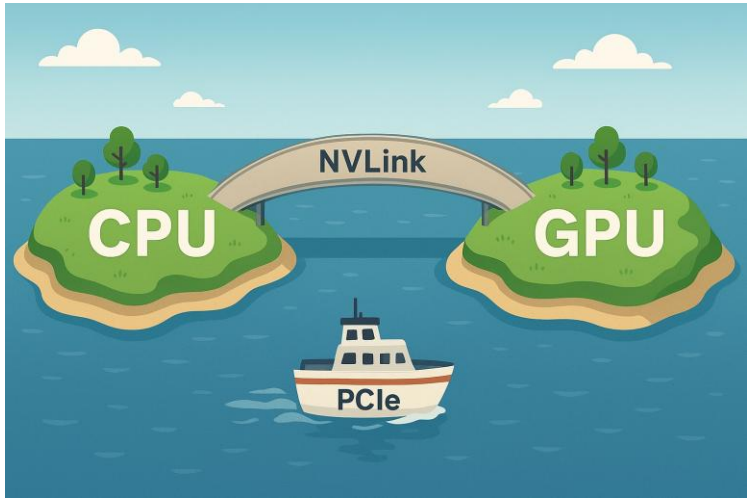
Table of Contents

HPC's Road of Heterogeneous Processor

- **TSUBAME 1.2** (2008) – 1st GPGPU-powered Cluster
  - AMD Opteron Barcelona + NVIDIA Tesla T10
- **Titan** (2012) – #1 in TOP 500 GPGPU-powered Cluster
  - AMD Opteron 6274 + NVIDIA Tesla K20X
- **Alps** (2024) – 1st Grace-Hopper-powered Cluster
  - Grace 72 ARMv9 + Hopper H100 GPU (GH200)
- **El Capitan** (2025) - #1 in TOP 500
  - AMD MI300A
- **Green 500**
  - 4/10 Top 10 are using GH200 (#1, #2, #4, #8 @2025 June)
  - 2/10 Top 10 are using MI300A (#3, #9 @ 2025 June)



Challenges: The "Two-Island" Problem



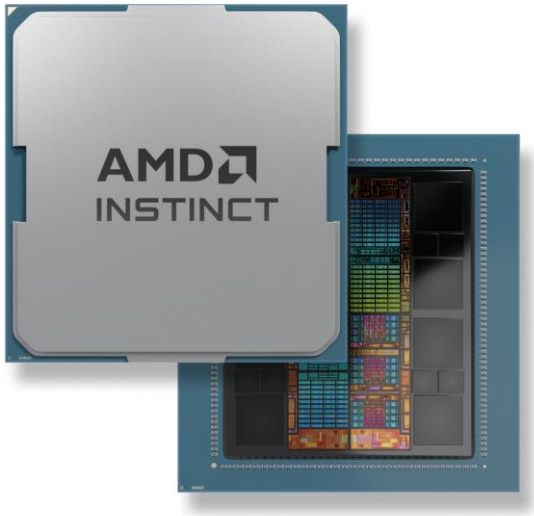
- **Separate Memory Pools**
  - CPU and GPU separate physical memory (DRAM ↔ HBM/GDDR).
- **The "Data Ferry": Expensive & Explicit Copies**
  - Data must be manually copied (`memcpy`) between host and device to be used. **[1]**
  - This "ferry" trip consumes significant **time (latency)** and **energy**, reducing overall efficiency.
- **The "Bigger Bridge" Fallacy**
  - Faster interconnects (NVLink, Infinity Fabric) create a wider bridge, but don't solve the core issue.
  - The two memory islands still exist; copies and synchronization are still required.
- **The Burden on Software**
  - Developers and middleware must explicitly manage data location and traffic.
  - This results in higher code complexity, longer development time, difficult performance trade-offs.

[1] Courtesy: Image generated by ChatGPT

Table of Contents

AMD MI300A APU (Accelerated Processing Unit)

- First data-center APU MI300A integrating CPU and GPU cores within a single package
- MI300A combines x86 CPU cores, CDNA3 GPU compute units
- Shared coherent pool of high-bandwidth HBM3 memory
- Eliminates explicit host-device data transfers
- Allows CPU and GPU to access the same physical memory seamlessly
- Changes communication costs and middleware strategies

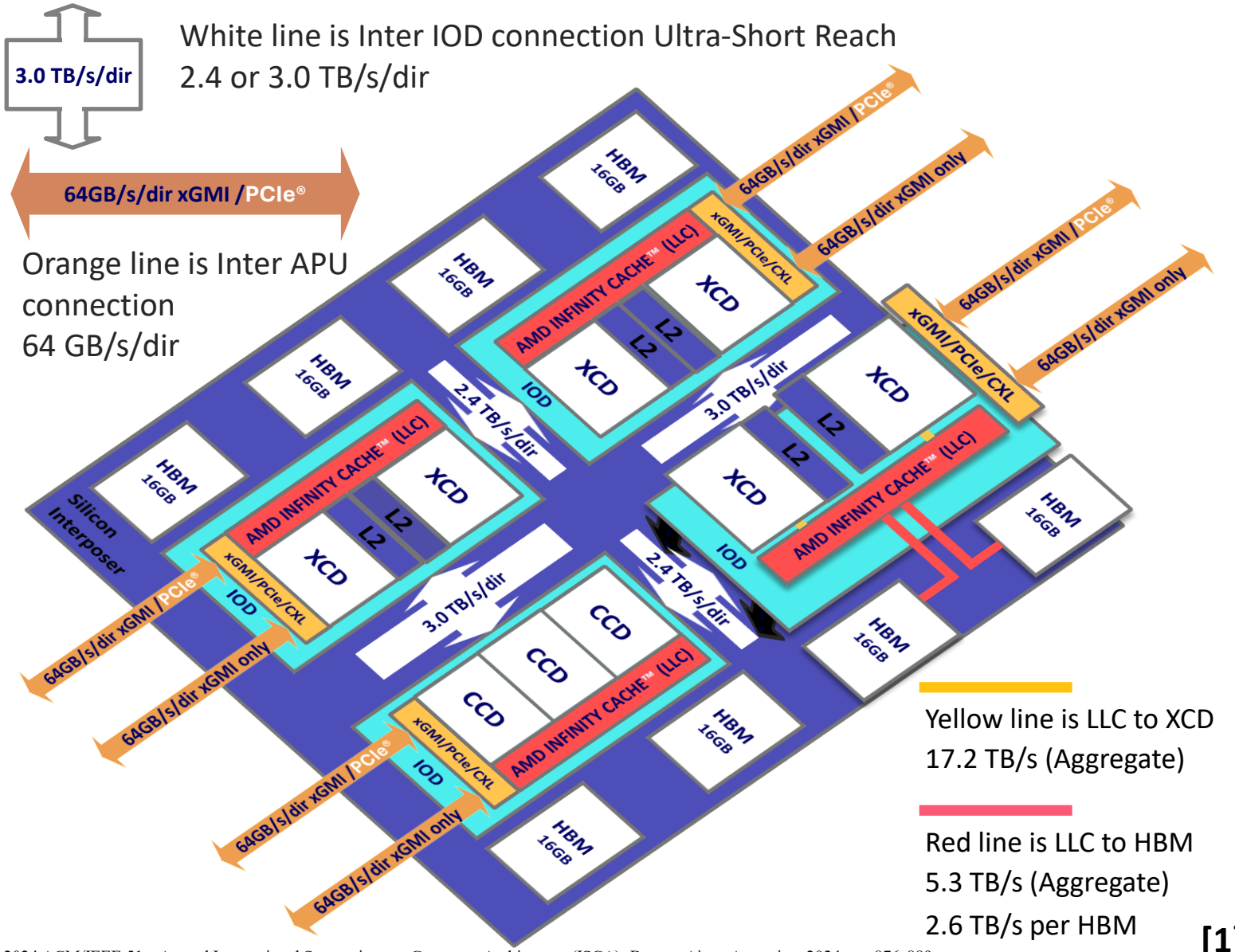


[1]

[1] Courtesy: AMD Instinct™ MI300A Accelerators from AMD website

MI300A Chiplet-Based Architecture Components

- **3** Core Complex Dies (CCDs), each with 8 high-performance Zen 4 CPU cores (24 total)
- **6** Accelerator Complex Dies (XCDs) delivering massive GPU compute (228 CDNA3 compute units)
- **4** I/O Dies (IODs) functioning as cached active interposers

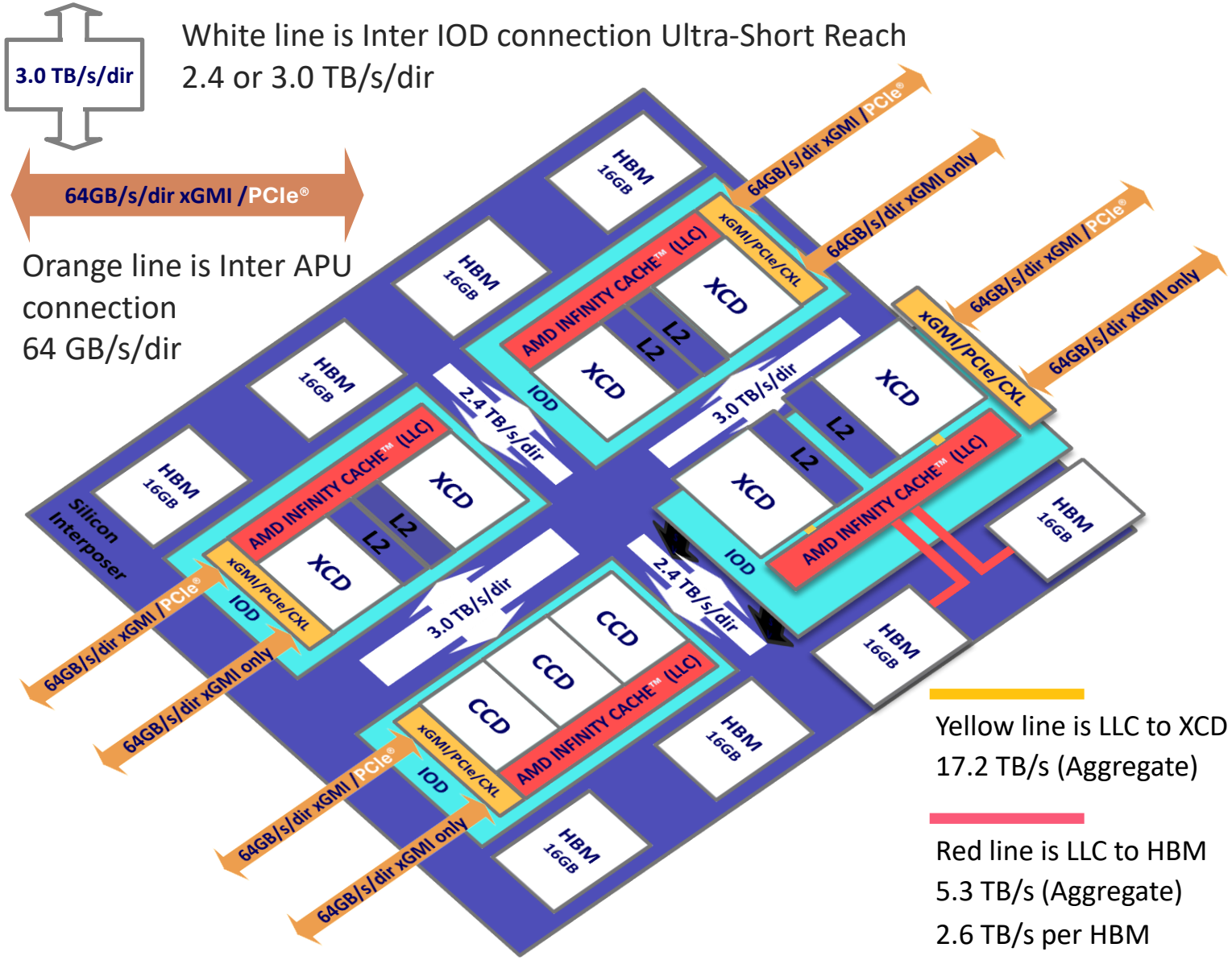


[1] Courtesy: A. Smith et al., "Realizing the AMD Exascale Heterogeneous Processor Vision : Industry Product," 2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA), Buenos Aires, Argentina, 2024, pp. 876-889,

[1]

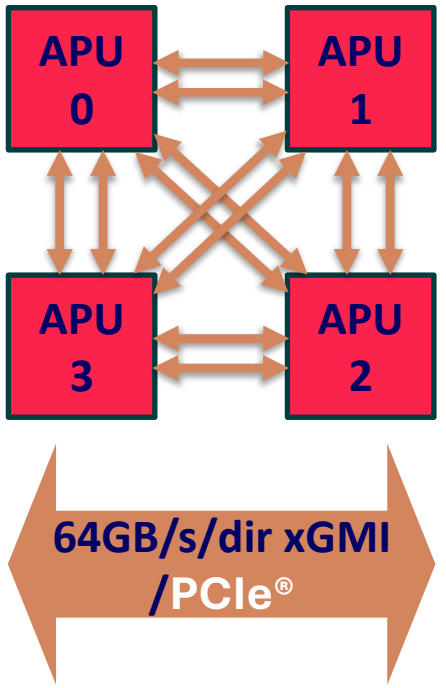
MI300A Chiplet-Based Architecture Components

- 256 MB Infinity Cache
- Peak theoretical bandwidth from the Infinity Cache is 17.2 TB/s
- 8 x 16GB HBM3 memory across 4 IODs (Total of 128GB)
- 5.3 TB/s peak theoretical HBM bandwidth



SDSC COSMOS

- 42 nodes, each with 4 APUs total of 168 APUs
- Nodes are interconnected by a fully connected network based on AMD's Infinity fabric and xGMI technology
- Interconnect delivers up to 768 GB/s aggregate and 256 GB/s peer-to-peer bi-directional bandwidth between APUs
- Every node is provisioned with four HPE Cray Slingshot-11 interconnects, offering an aggregate bidirectional bandwidth of 200 GB/s (equivalent to 25GB/s per link in each direction).

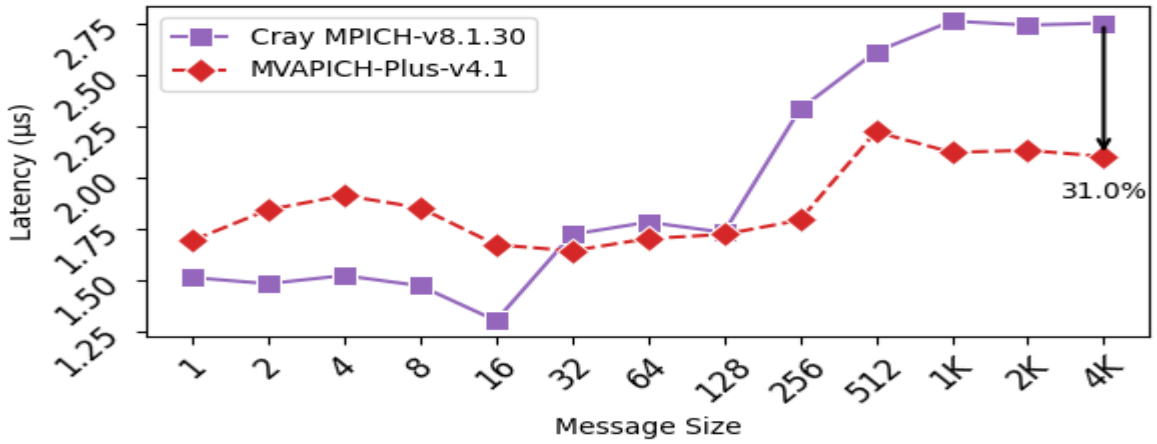


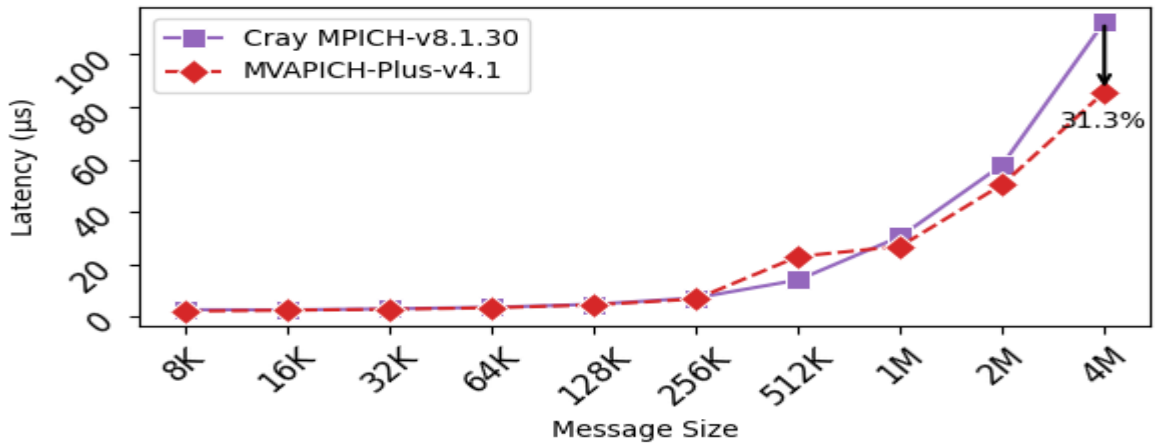
Orange line is Inter APU connection
64 GB/s/dir

Table of Contents

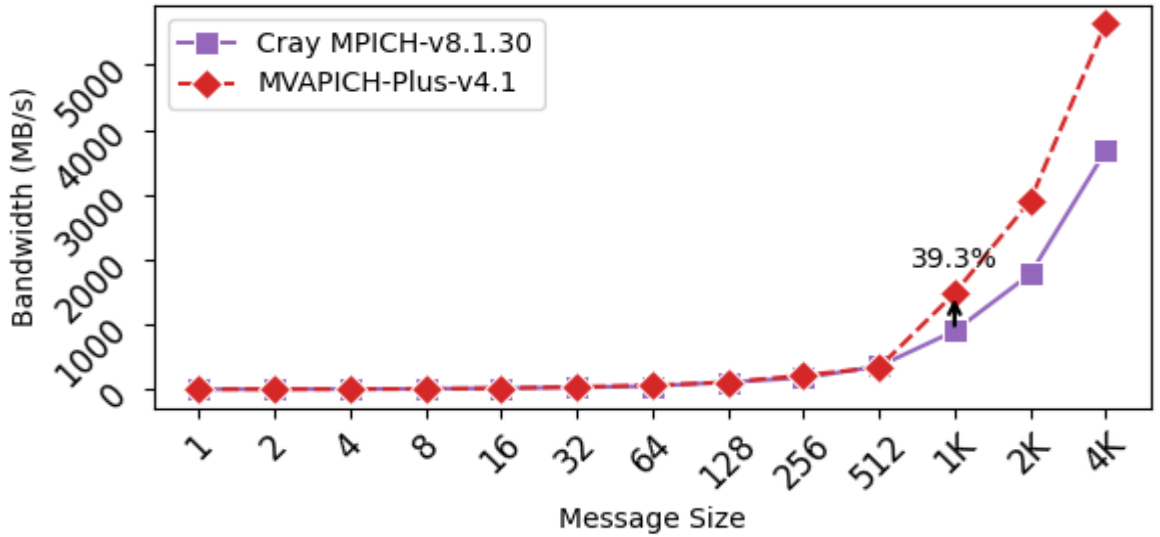
MVAPICH-Plus Performance on MI300A PT2PT Intra Node - CPU

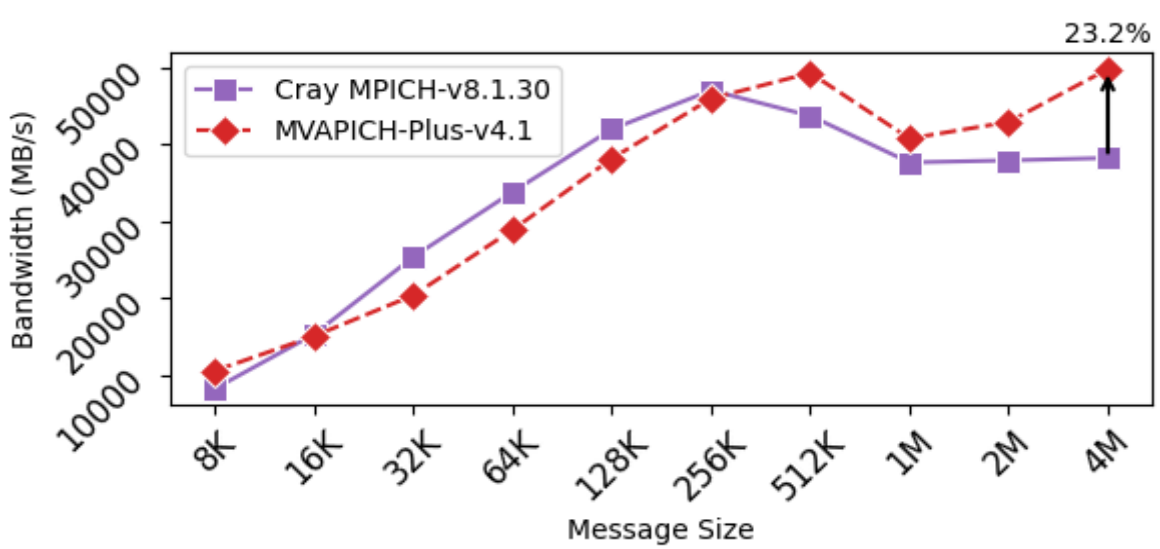
Latency





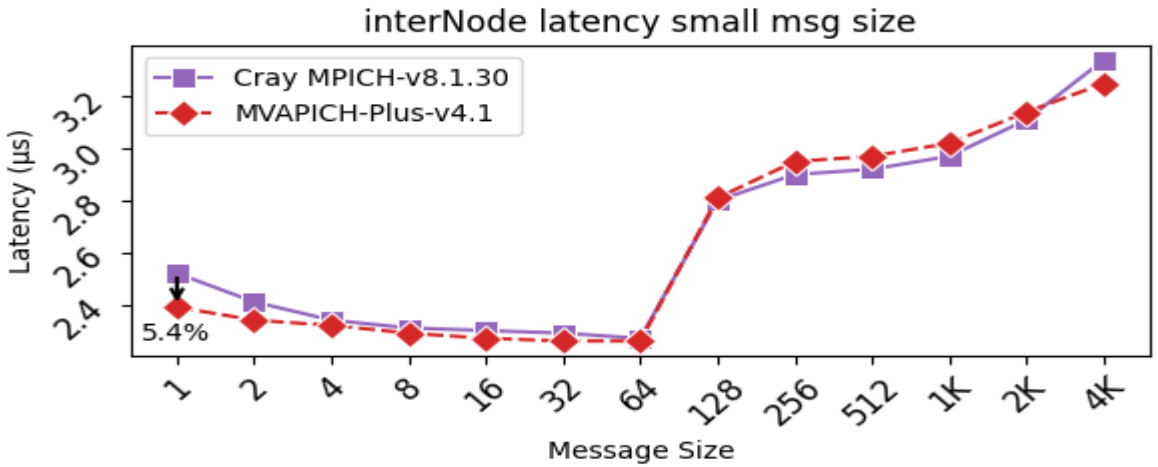
Bandwidth

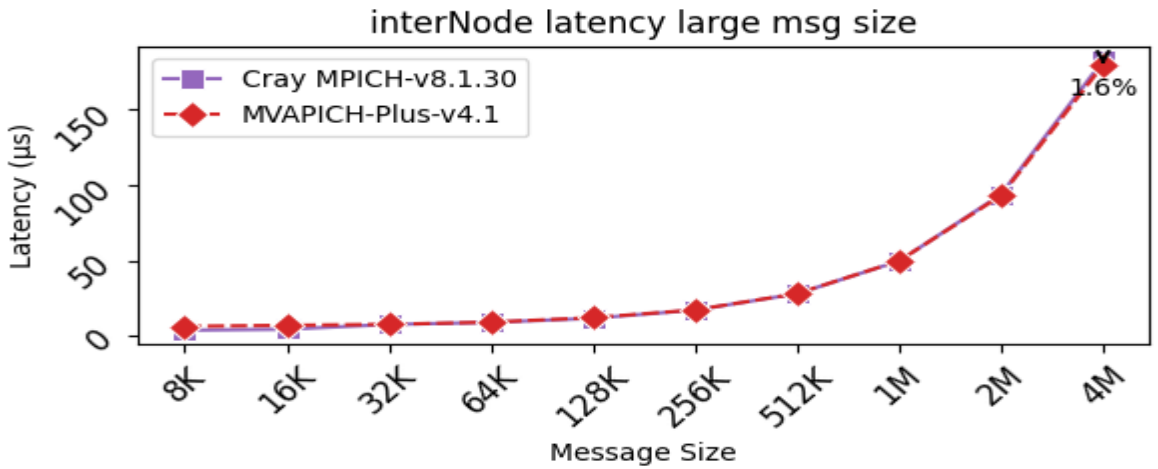




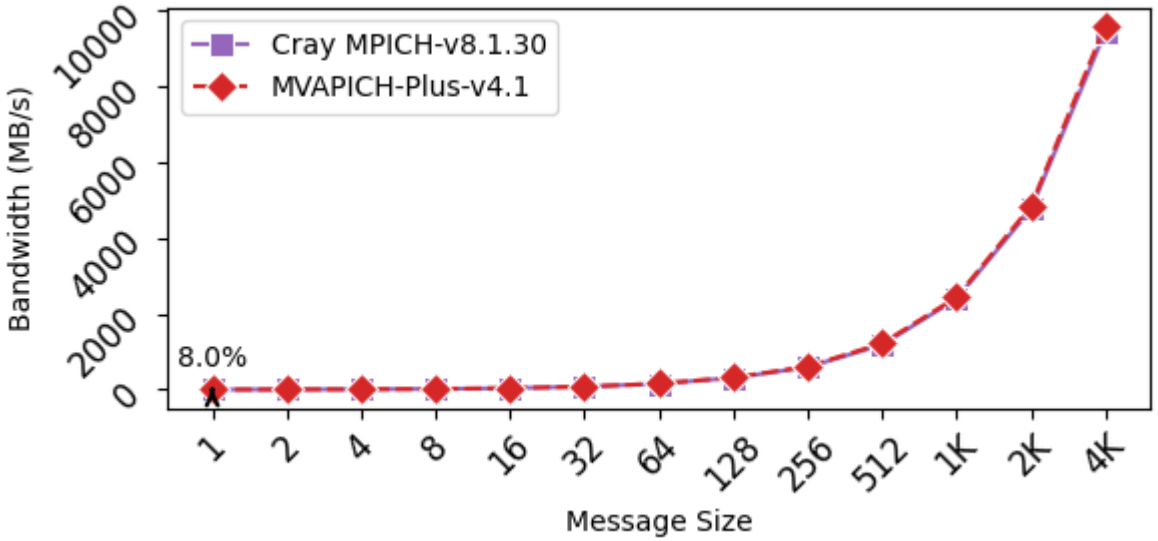
MVAPICH-Plus Performance on MI300A PT2PT Inter Node - CPU

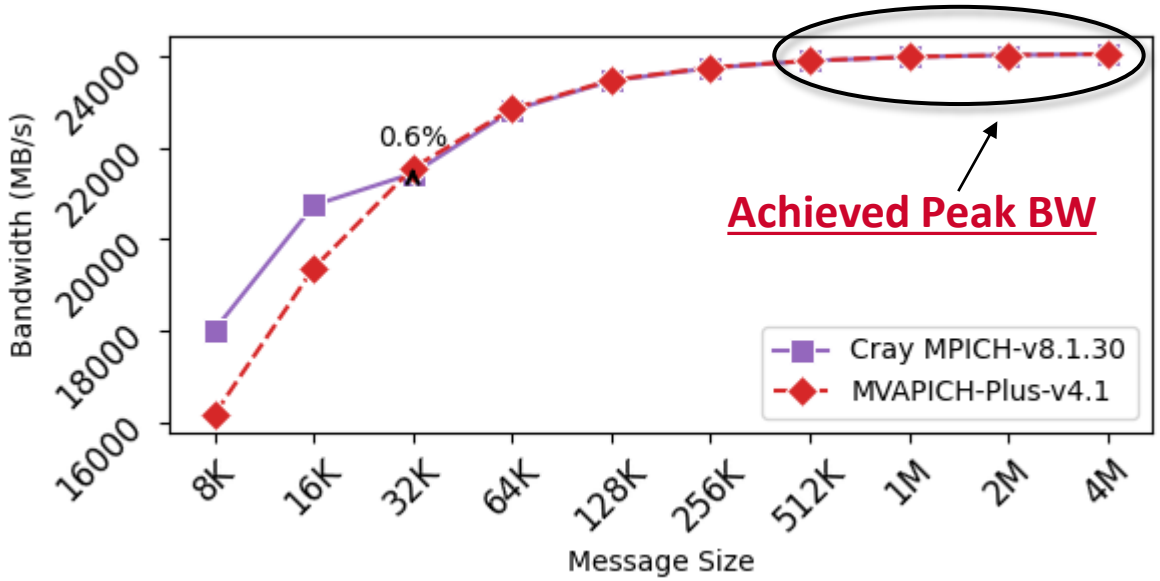
Latency





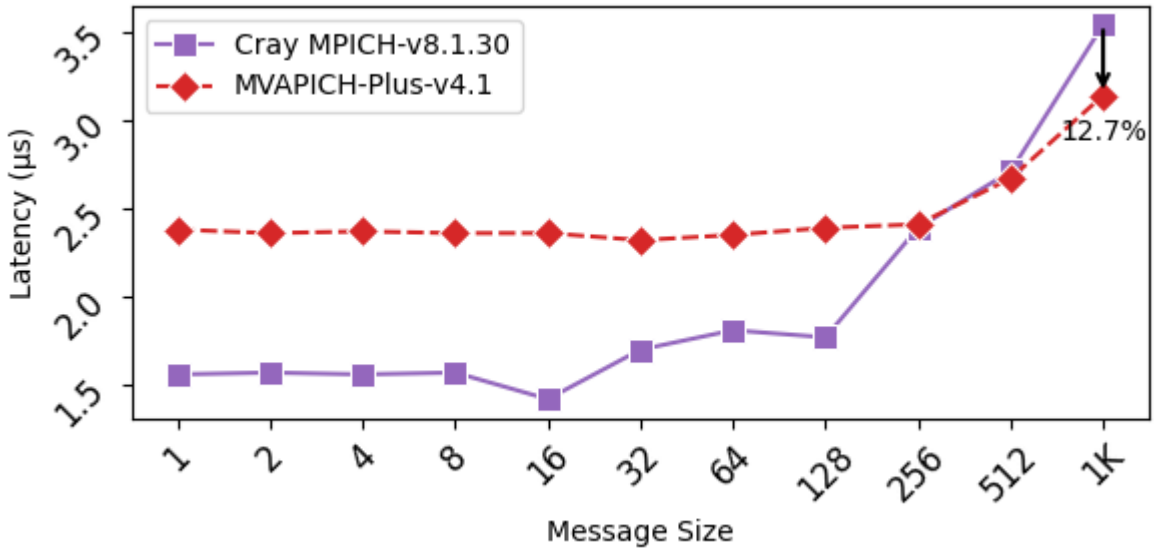
Bandwidth

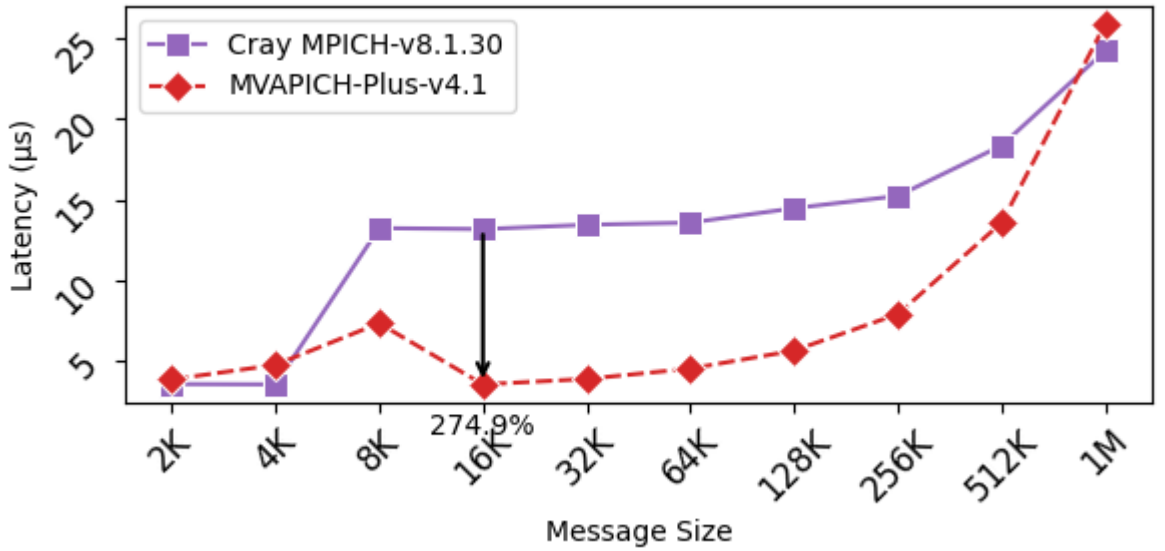




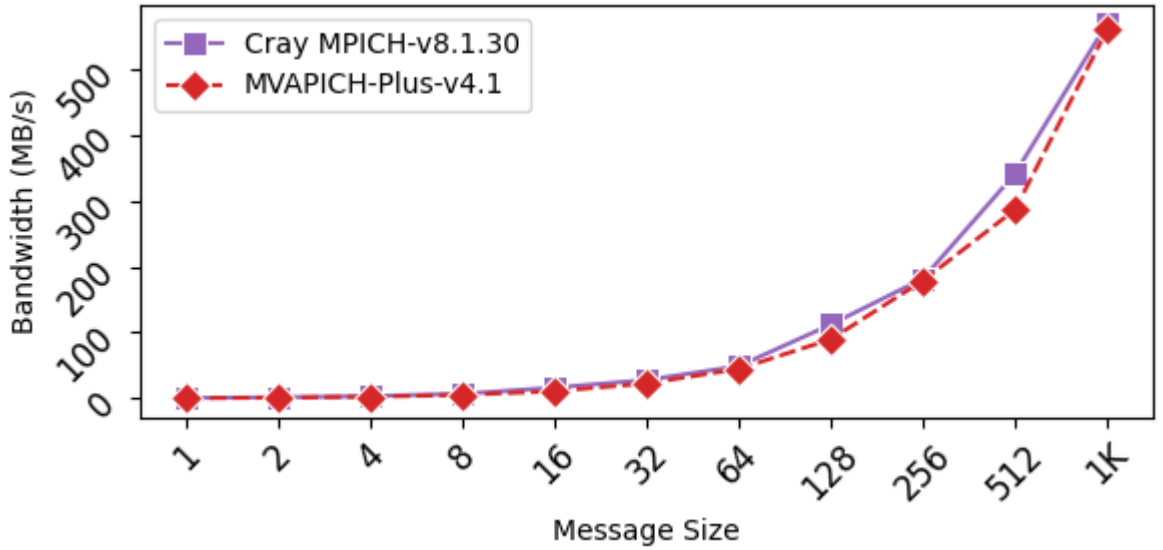
MVAPICH-Plus Performance on MI300A PT2PT Intra Node - GPU

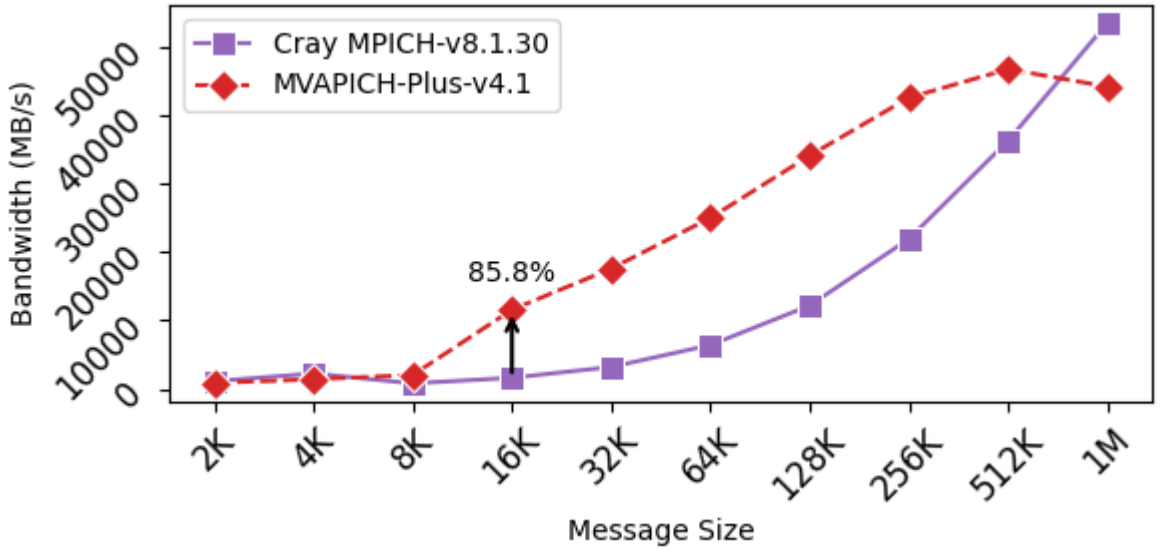
Latency





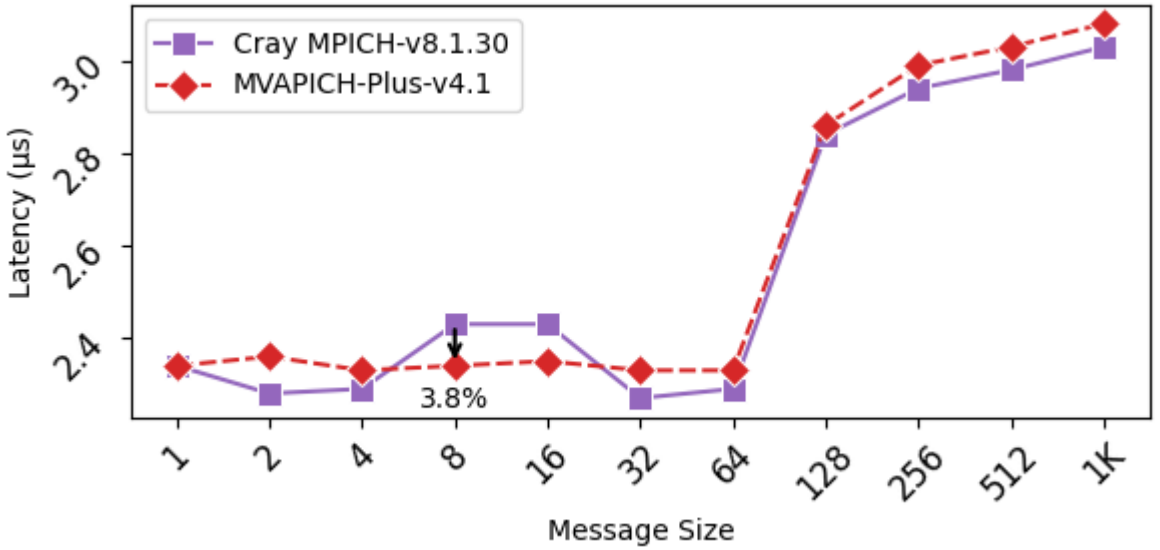
Bandwidth

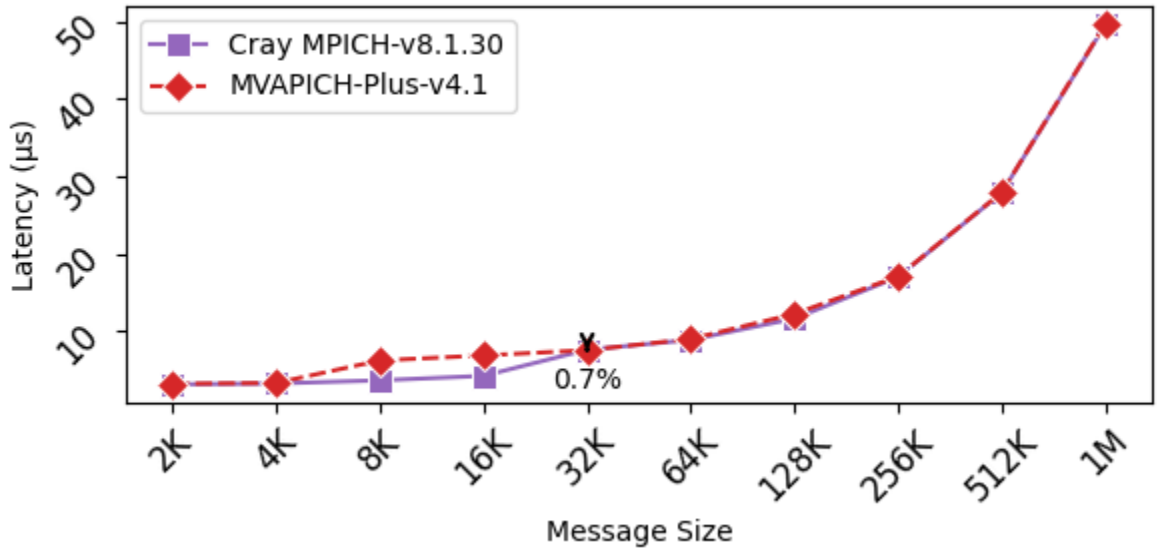




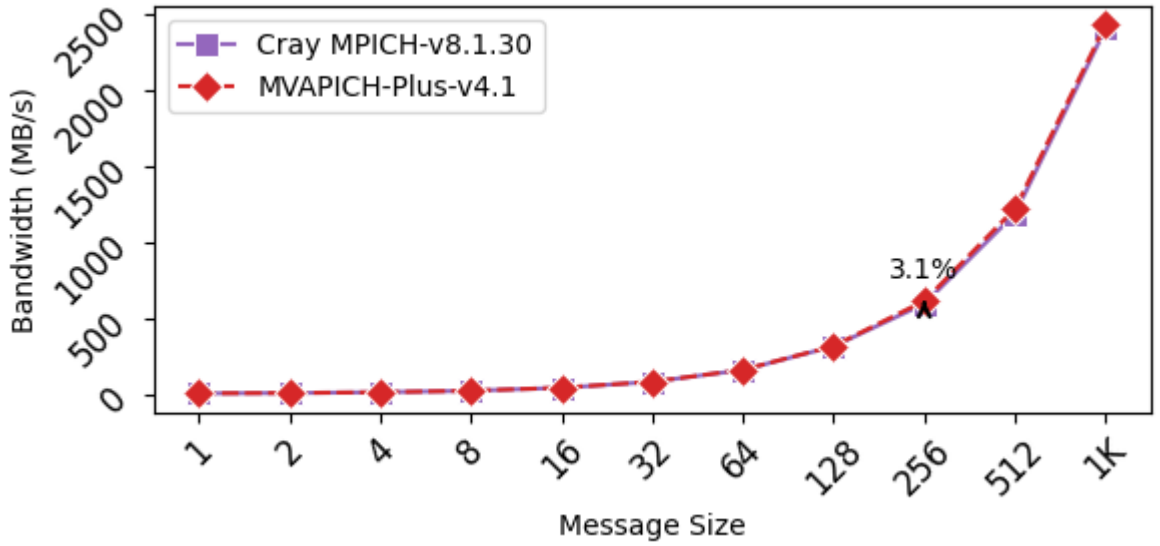
MVAPICH-Plus Performance on MI300A PT2PT Inter Node - GPU

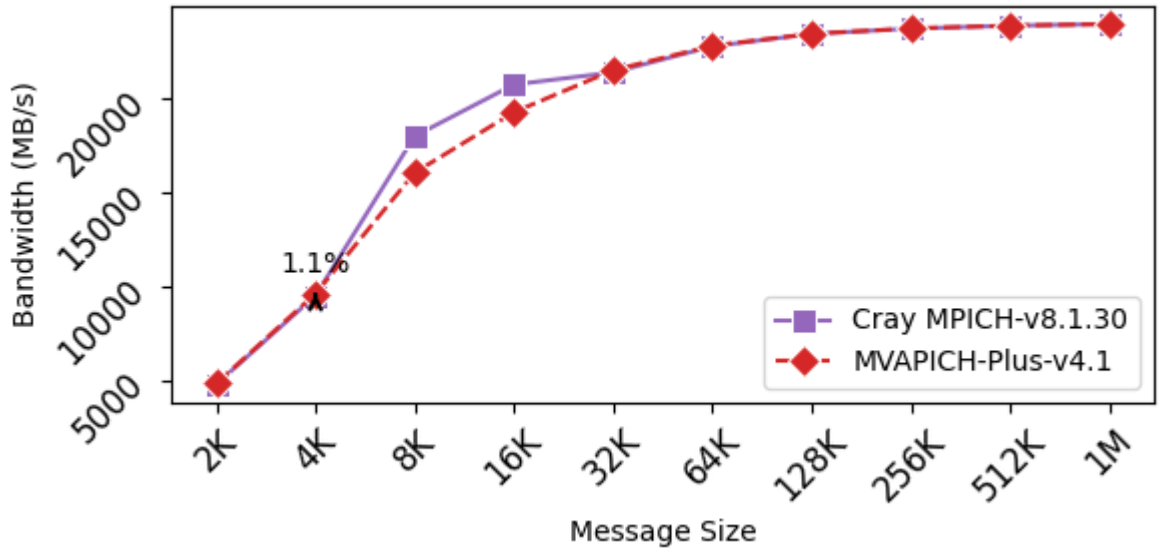
Latencv





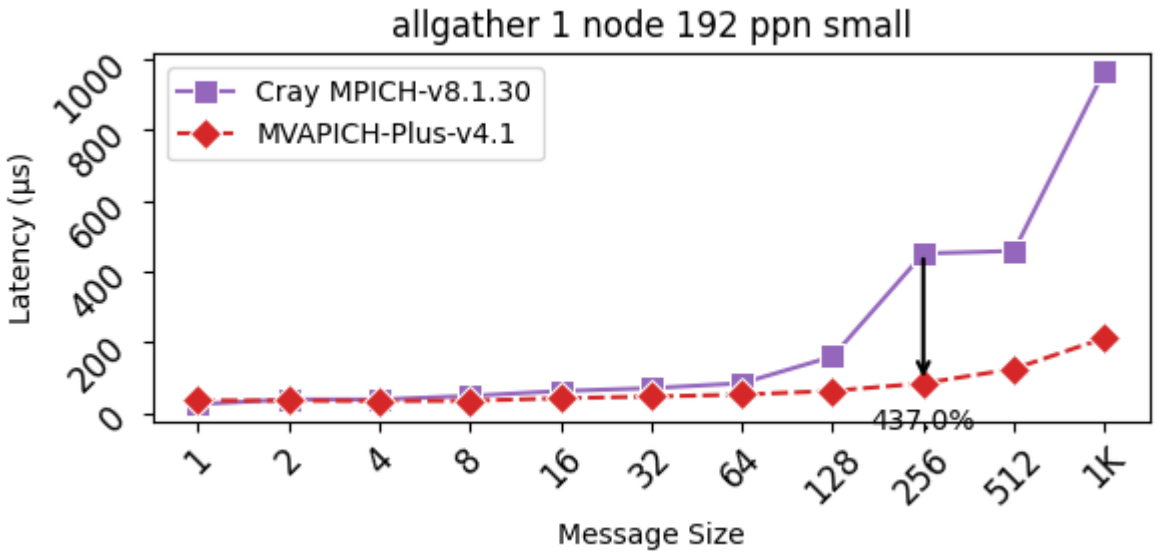
Bandwidth

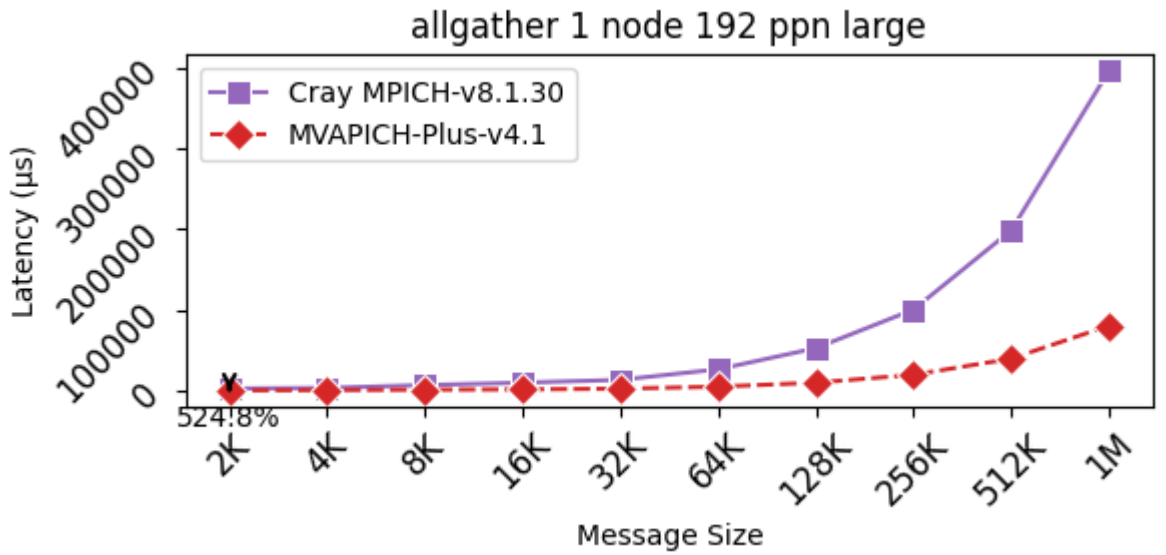




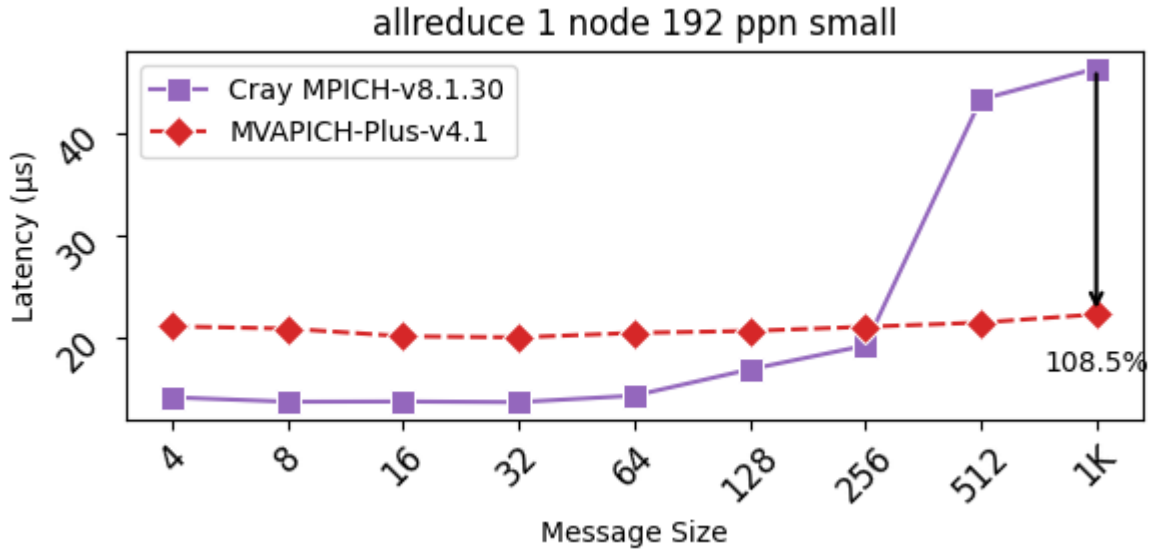
MVAPICH-Plus Performance on MI300A Collective 1 Node 192 PPN – CPU

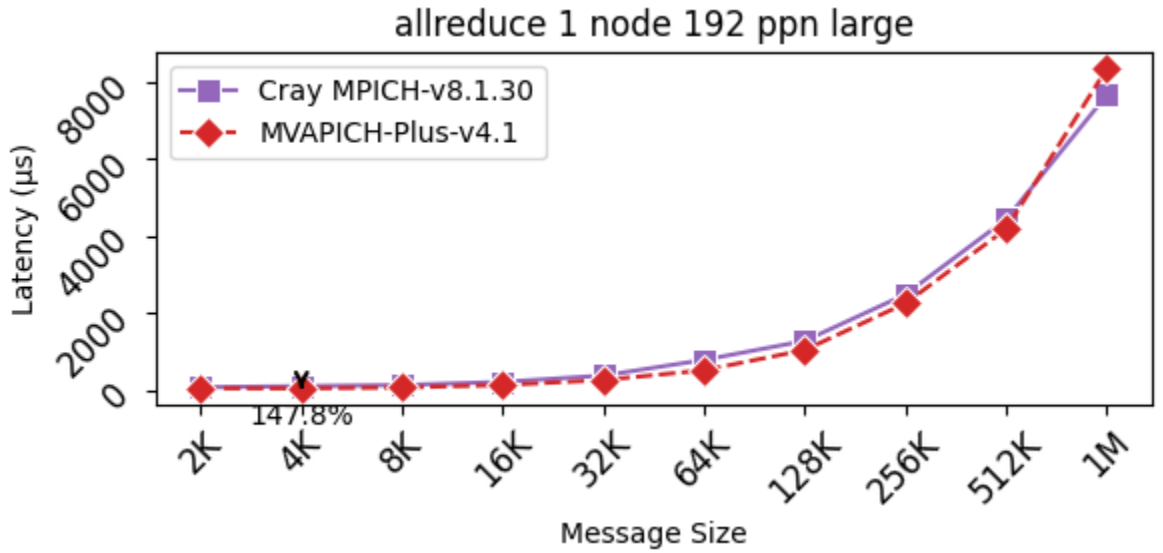
Allgather





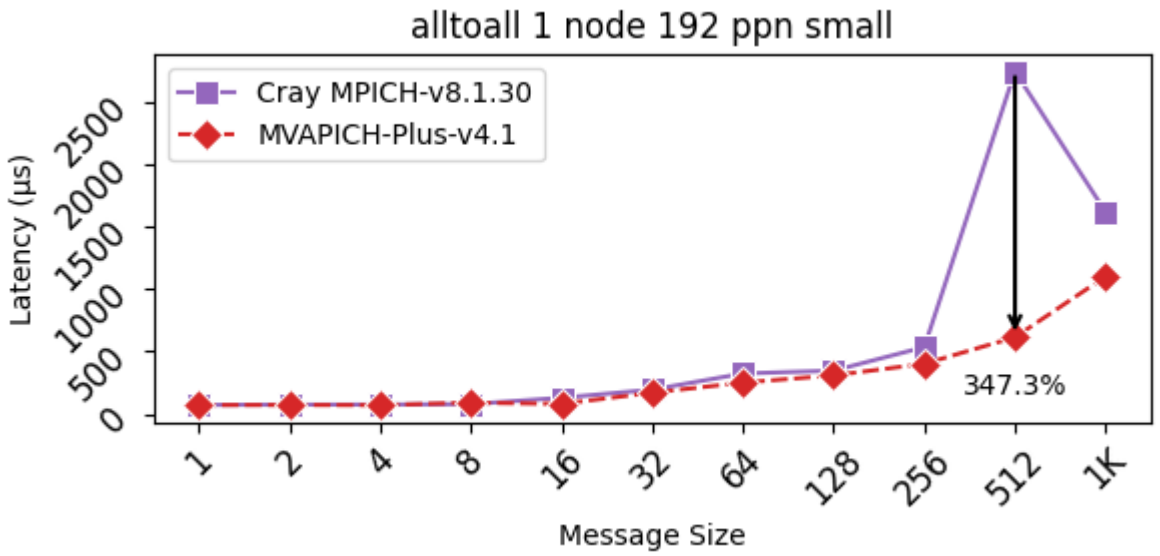
Allreduce

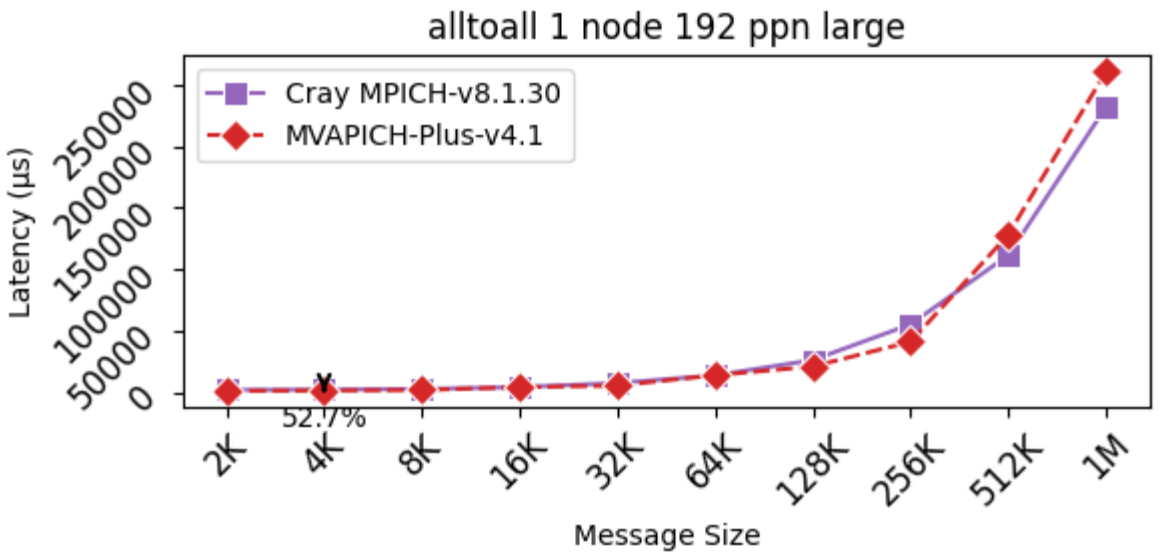




MVAPICH-Plus Performance on MI300A Collective 1 Node 192 PPN – CPU

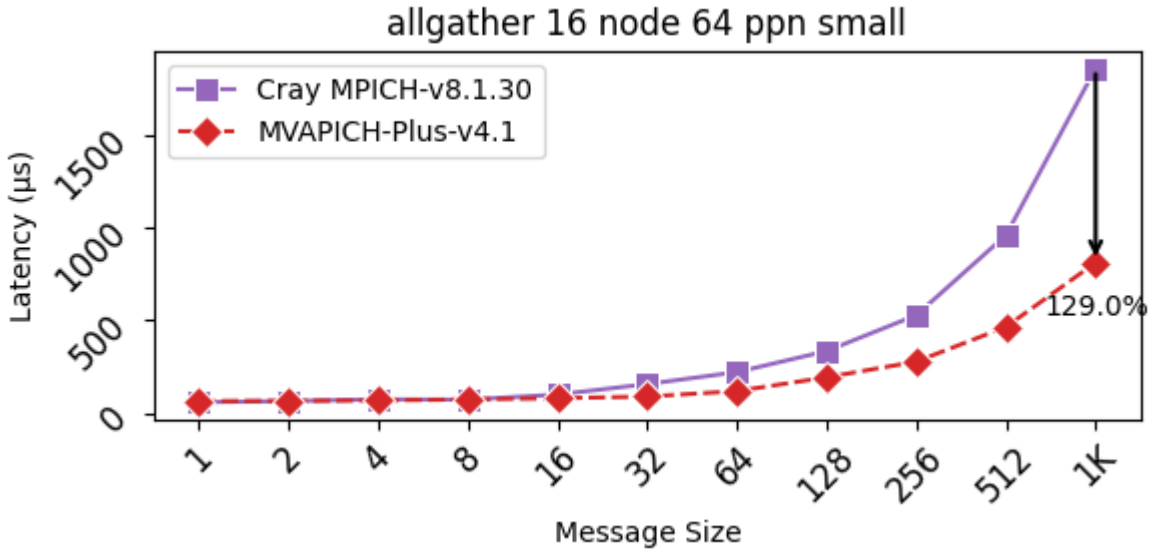
Alltoall

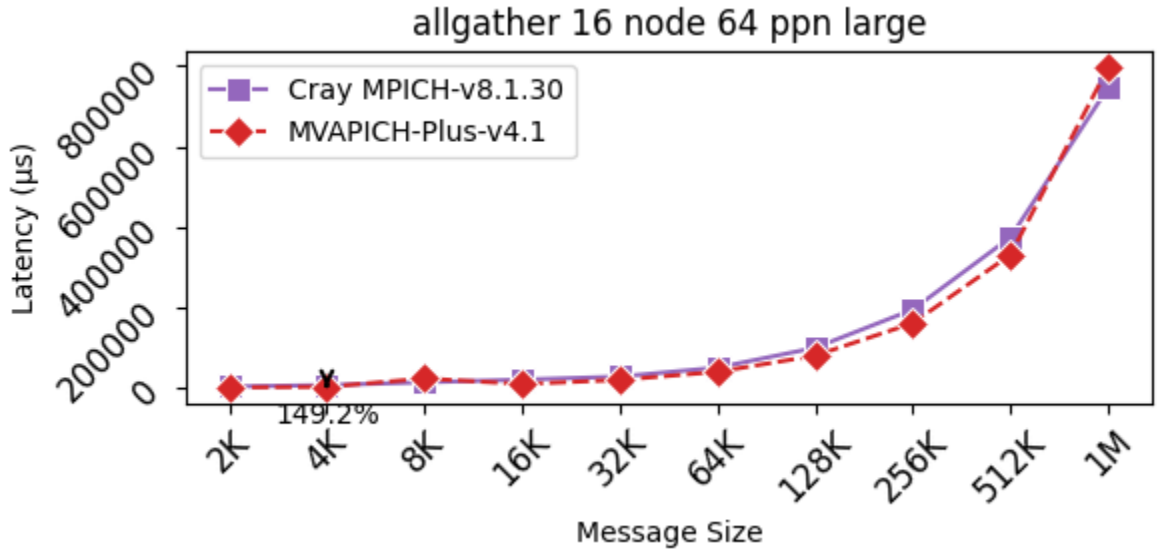




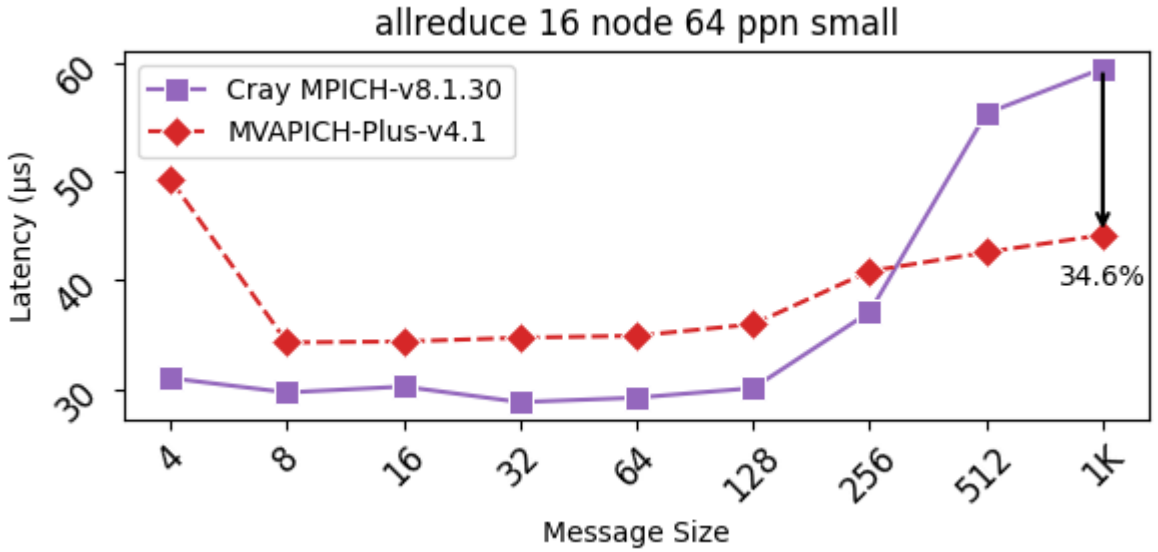
MVAPICH-Plus Performance on MI300A Collective 16 Node 64 PPN – CPU

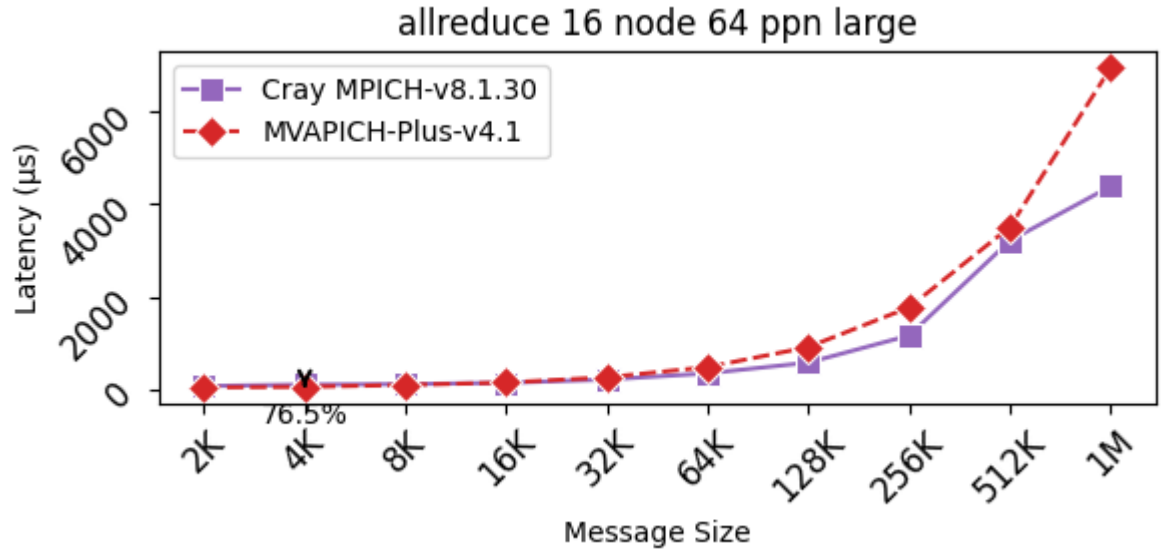
Allgather





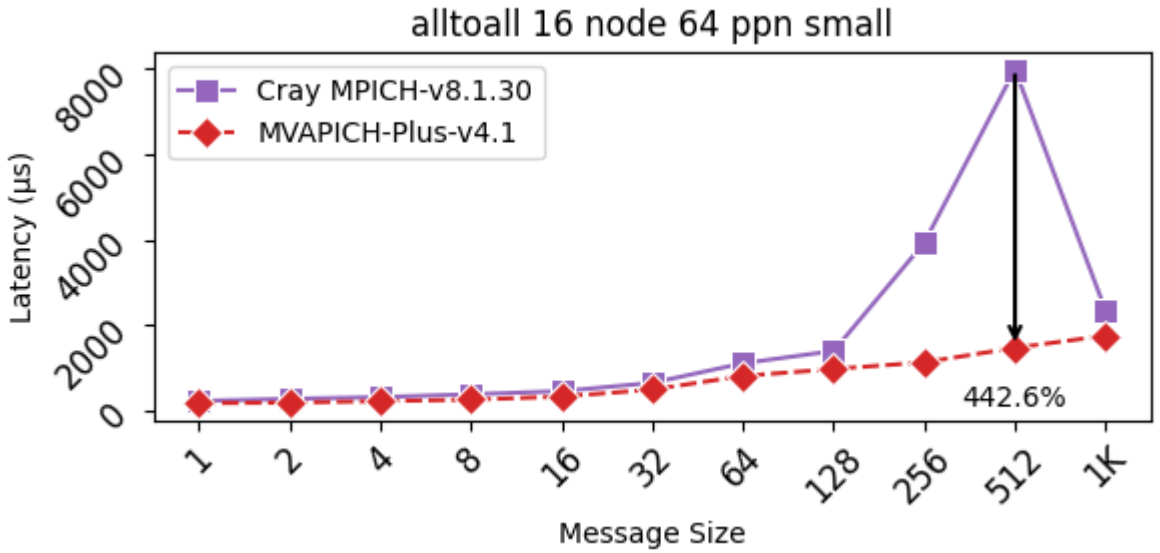
Allreduce

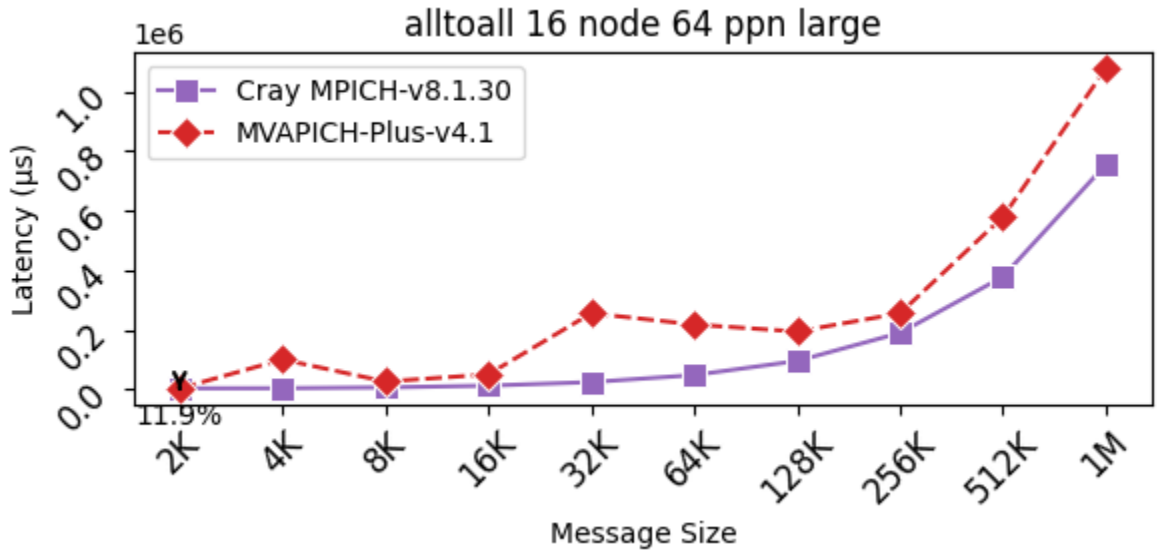




MVAPICH-Plus Performance on MI300A Collective 16 Node 64 PPN – CPU

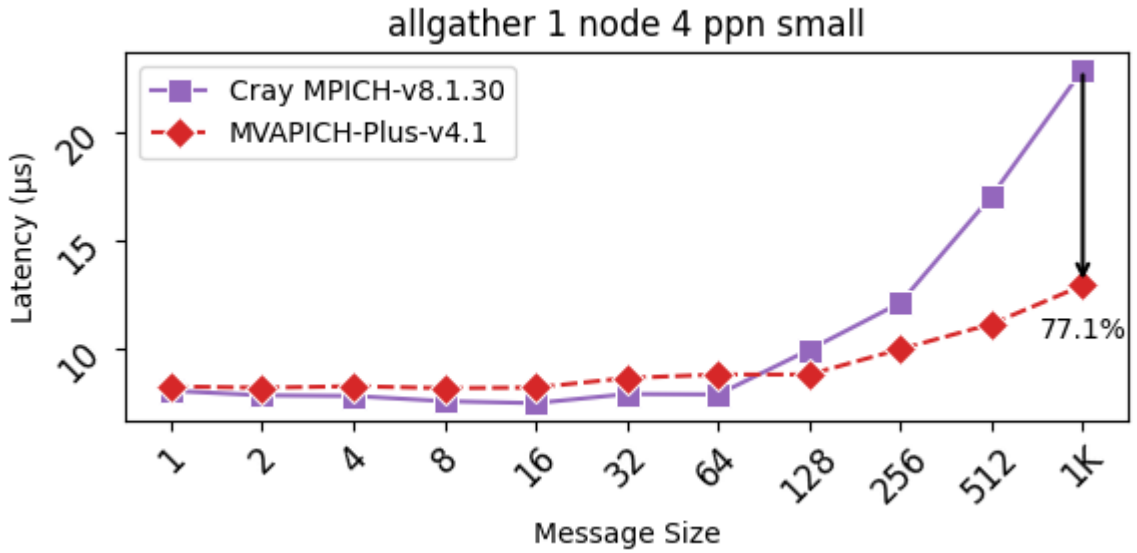
Alltoall

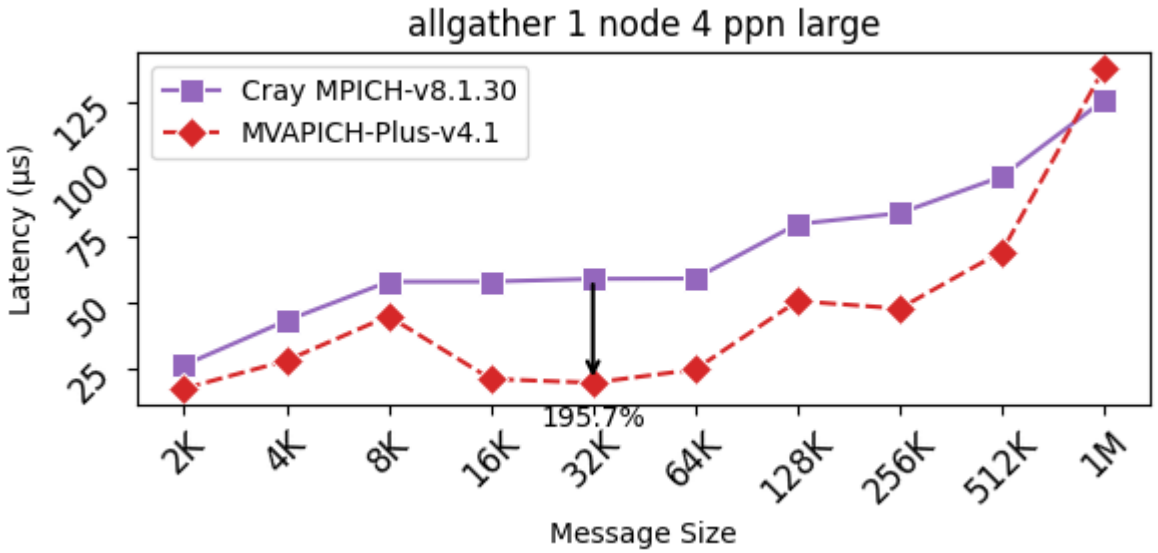




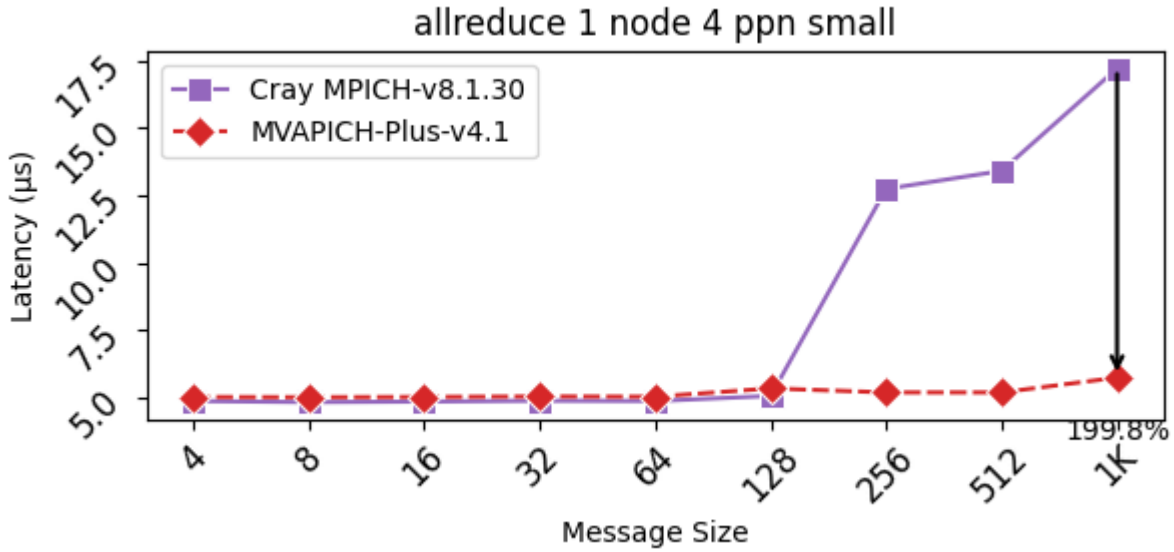
MVAPICH-Plus Performance on MI300A Collective 1 Node 4 PPN – GPU

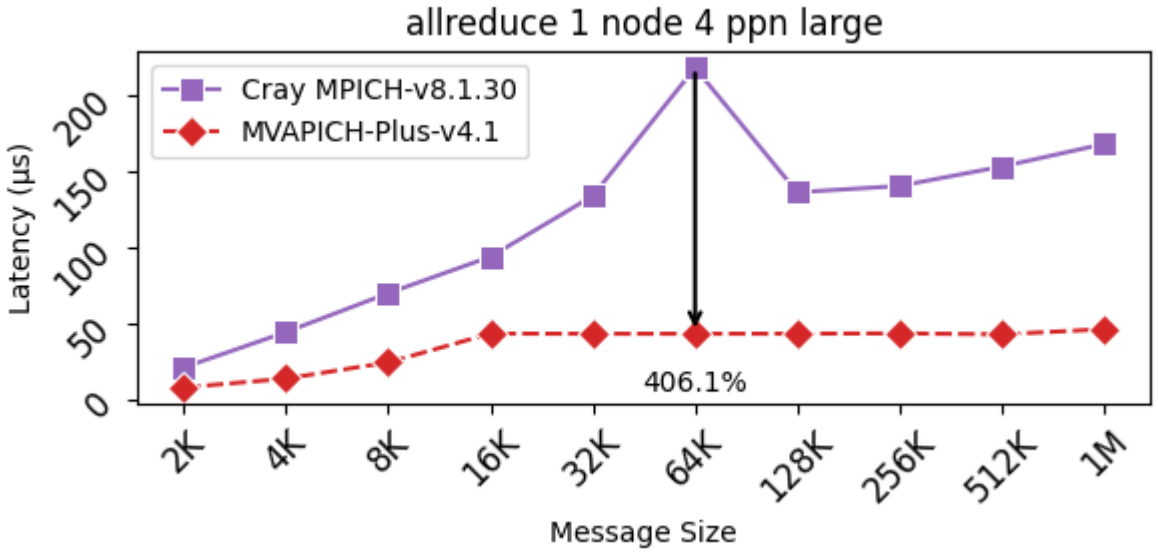
Allgather





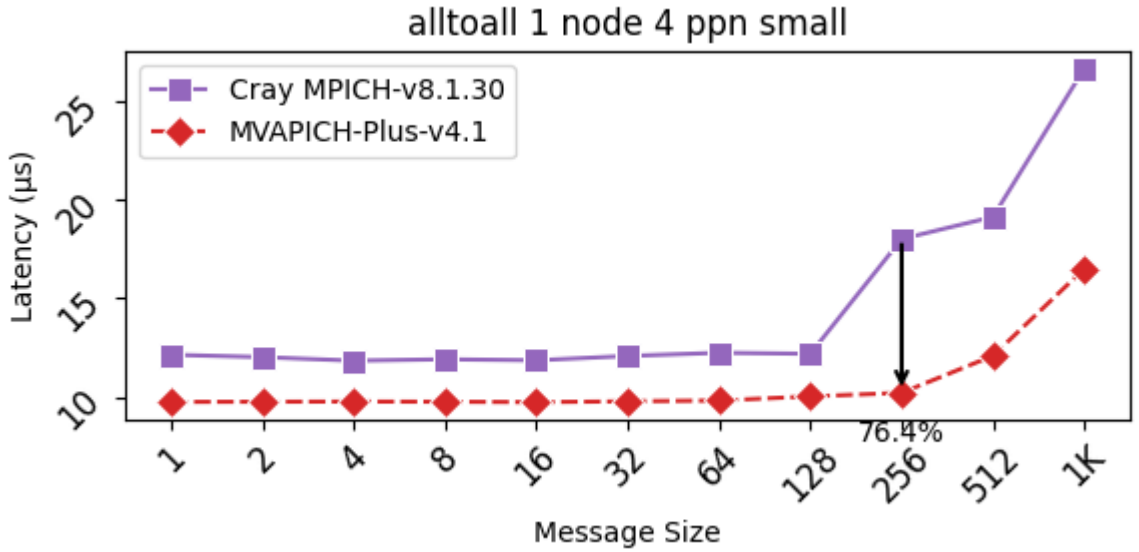
Allreduce

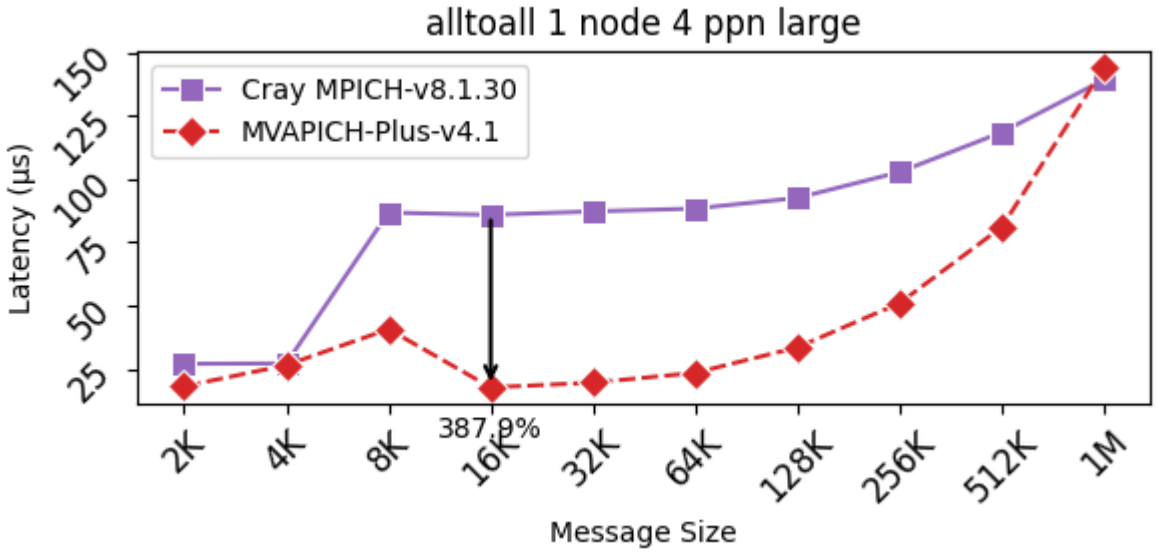




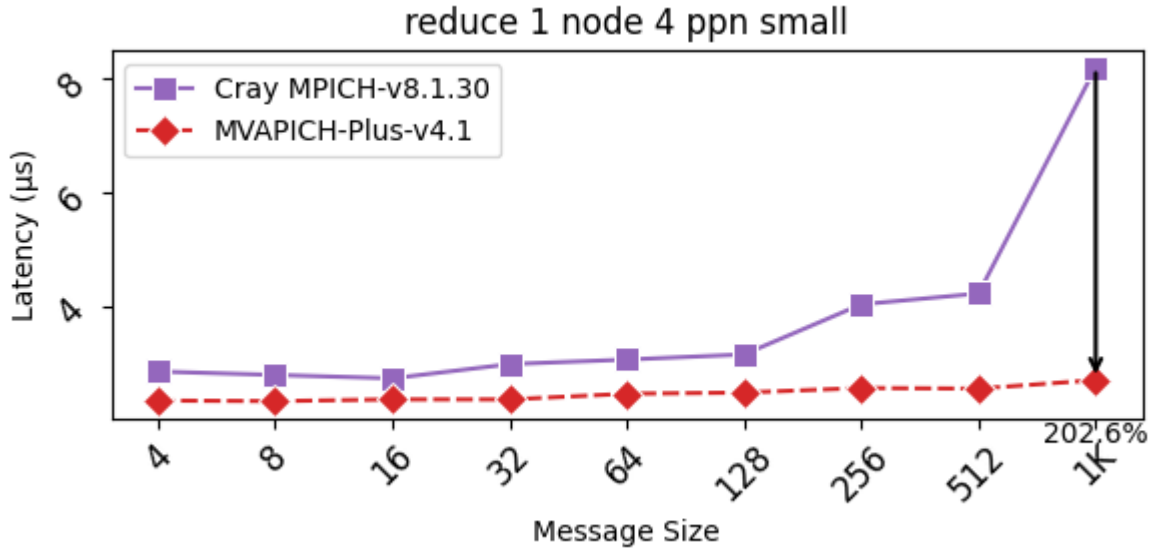
MVAPICH-Plus Performance on MI300A Collective 1 Node 4 PPN – GPU

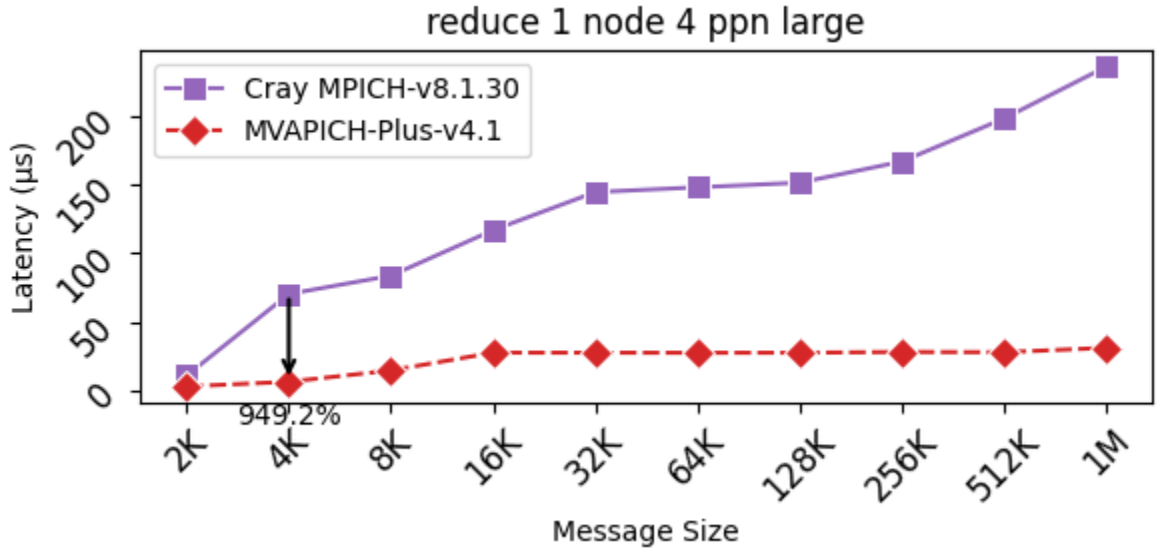
Alltoall





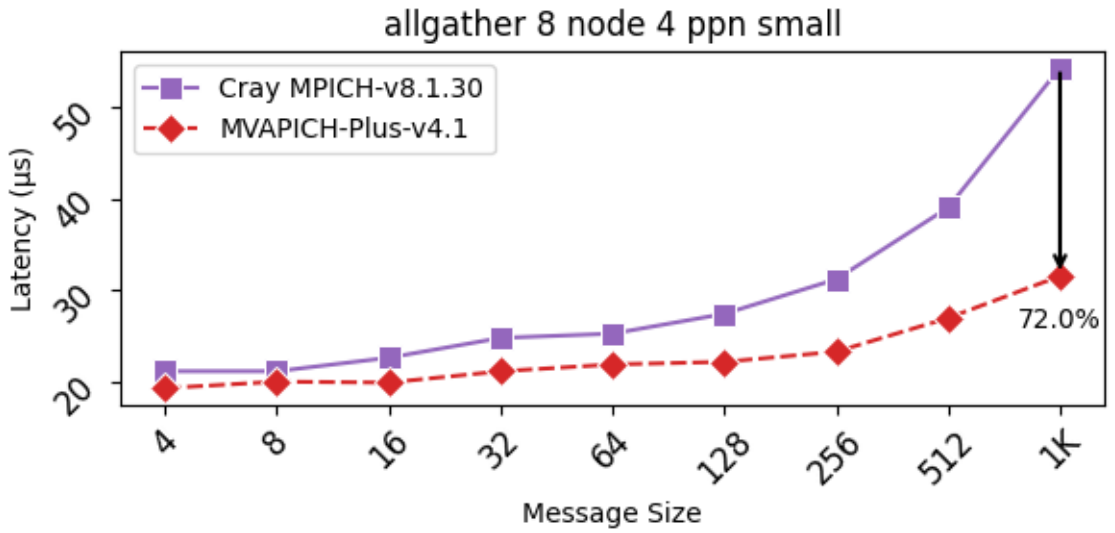
Reduce

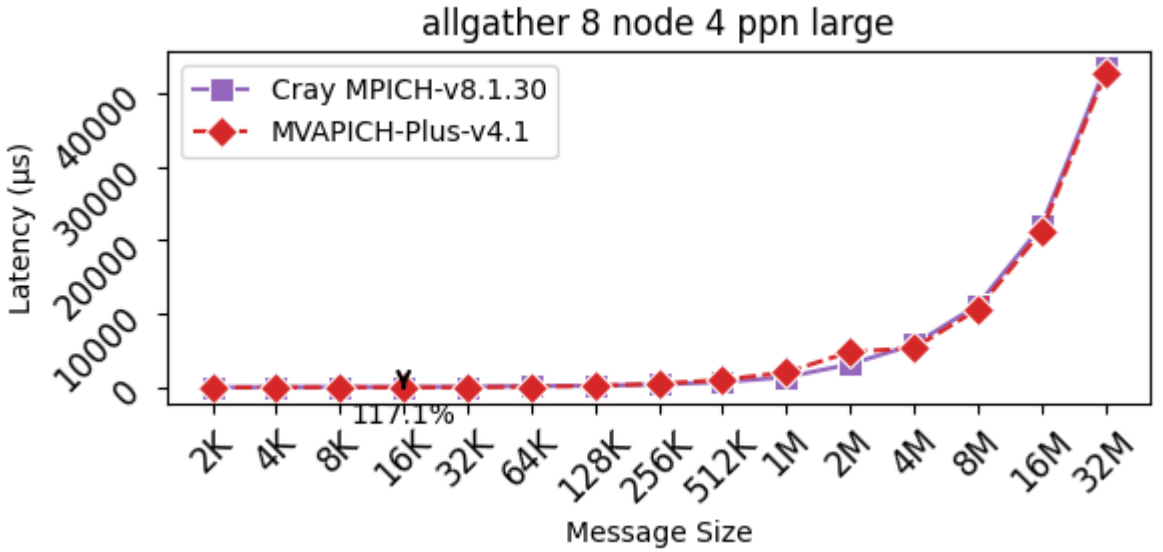




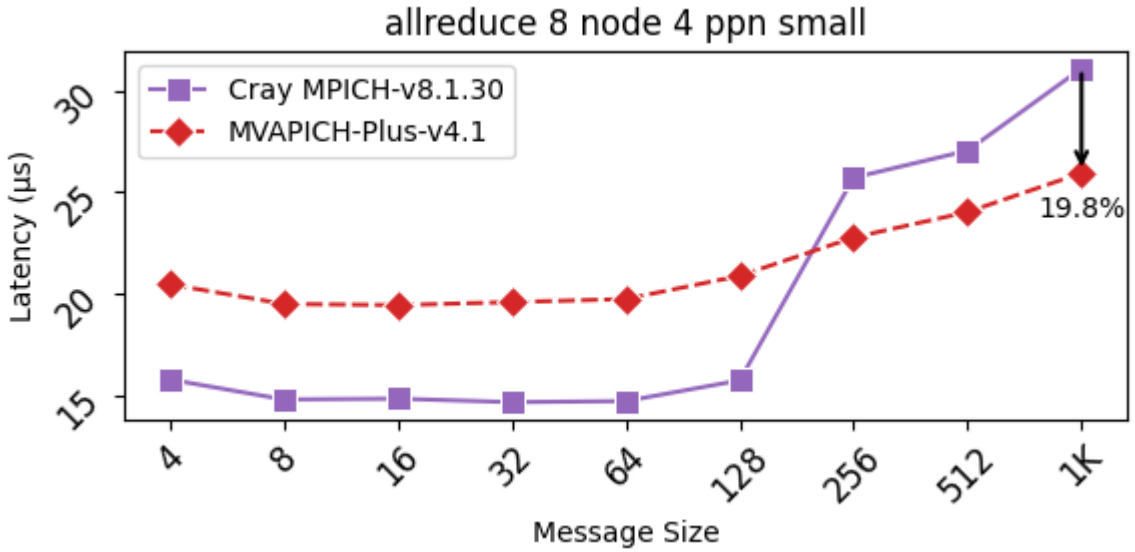
MVAPICH-Plus Performance on MI300A Collective 8 Node 4 PPN – GPU

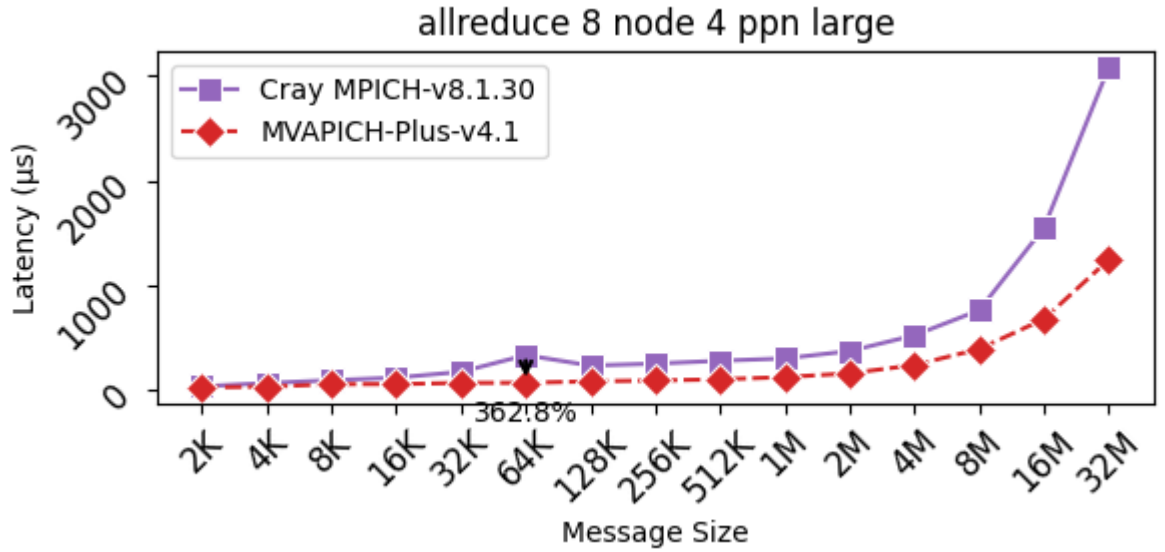
Allgather





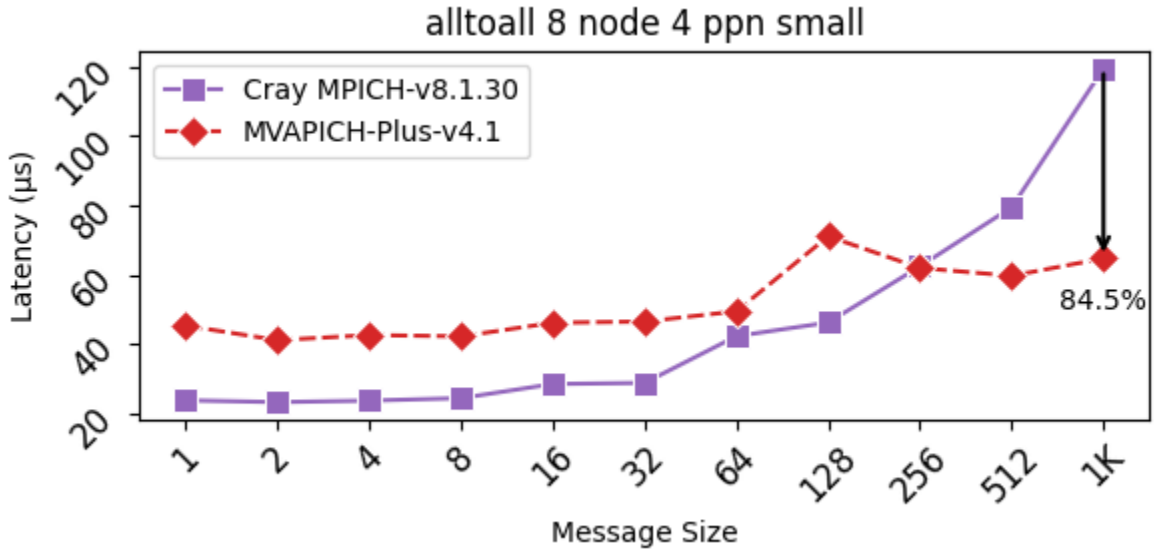
Allreduce

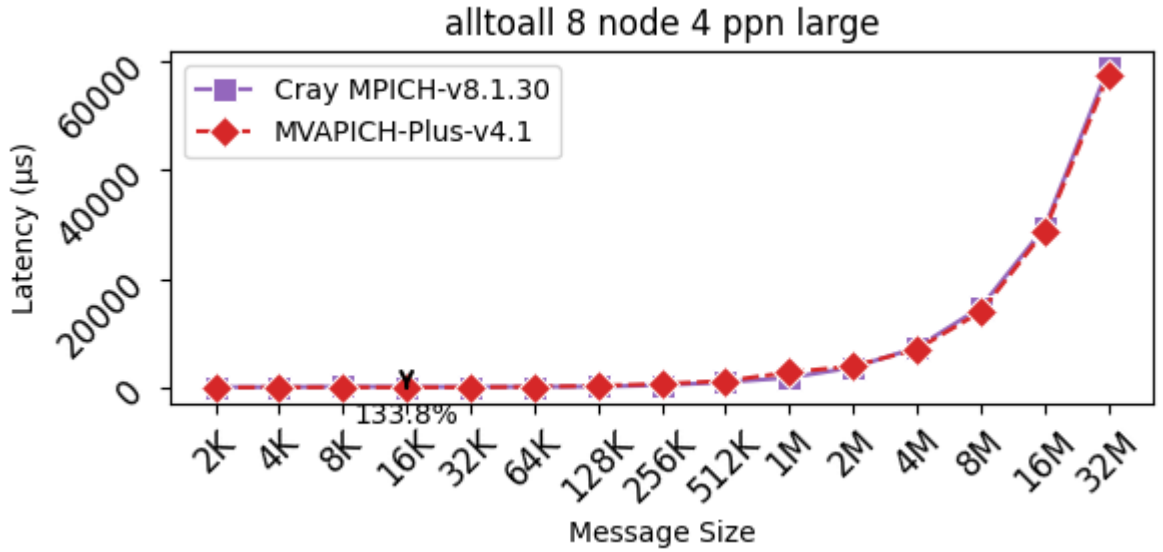




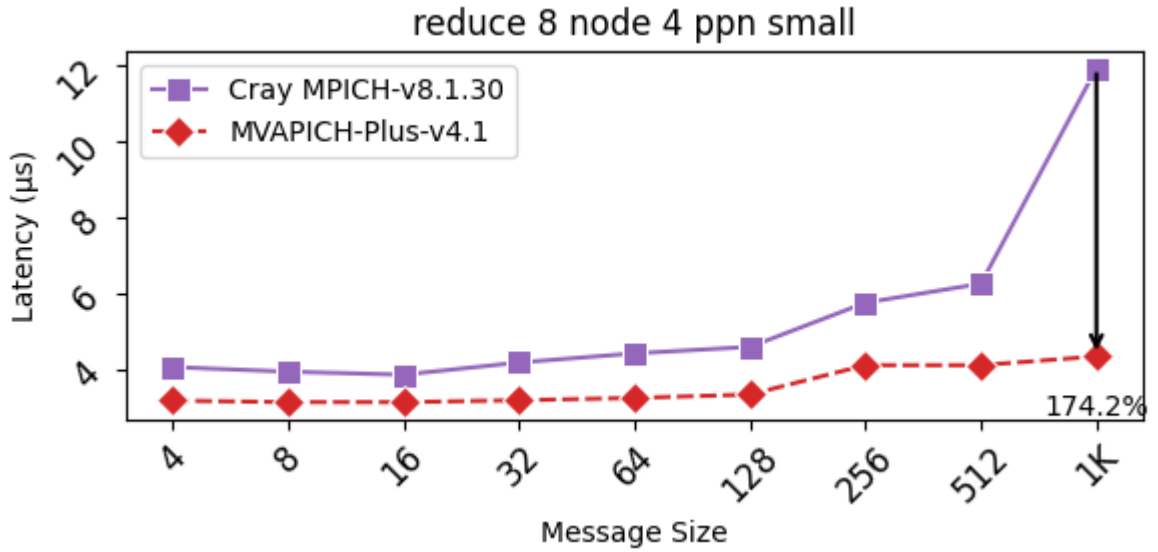
MVAPICH-Plus Performance on MI300A Collective 8 Node 4 PPN – GPU

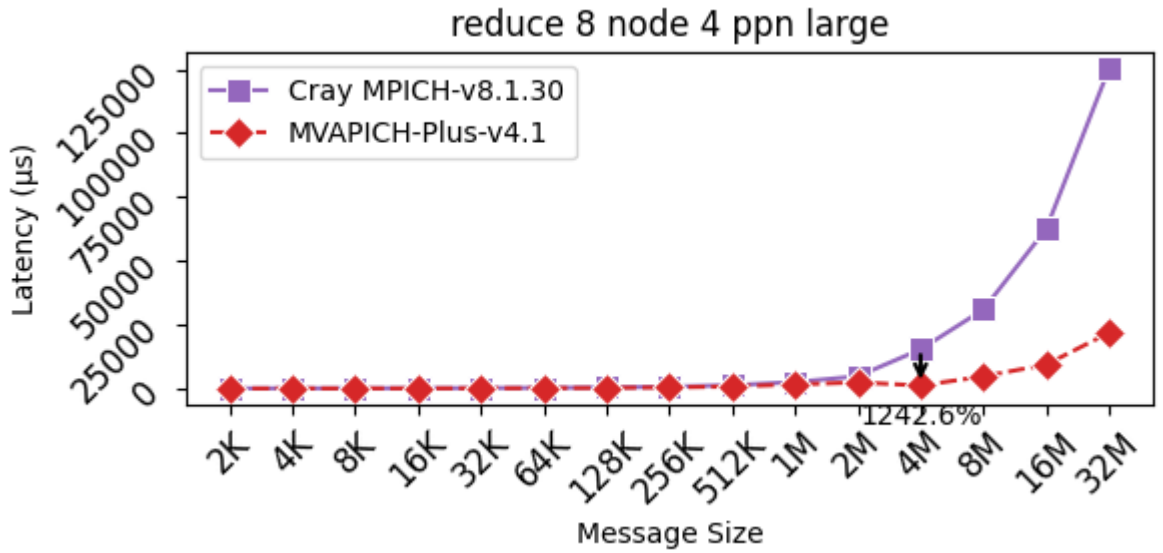
Alltoall





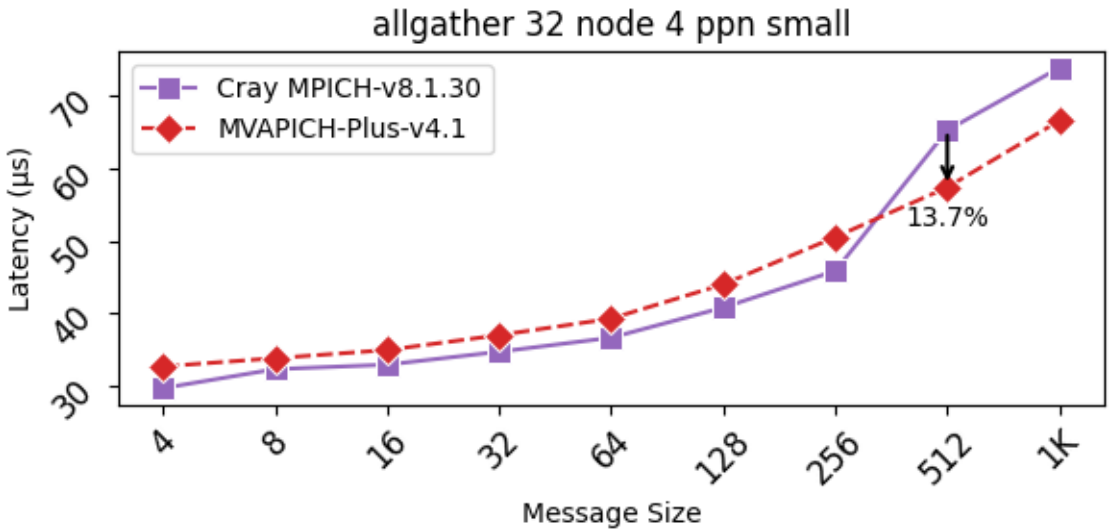
Reduce

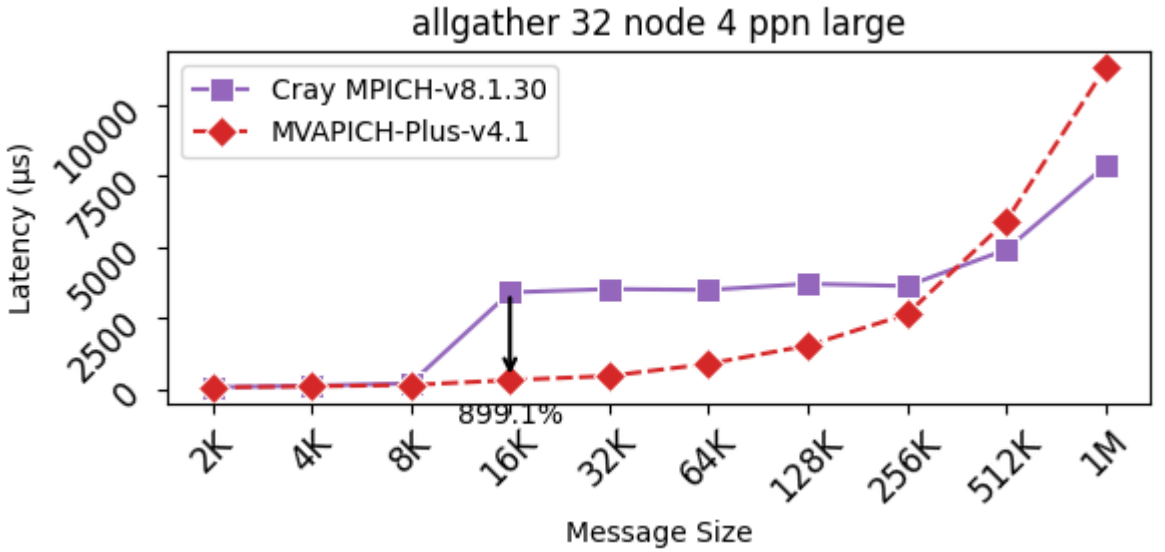




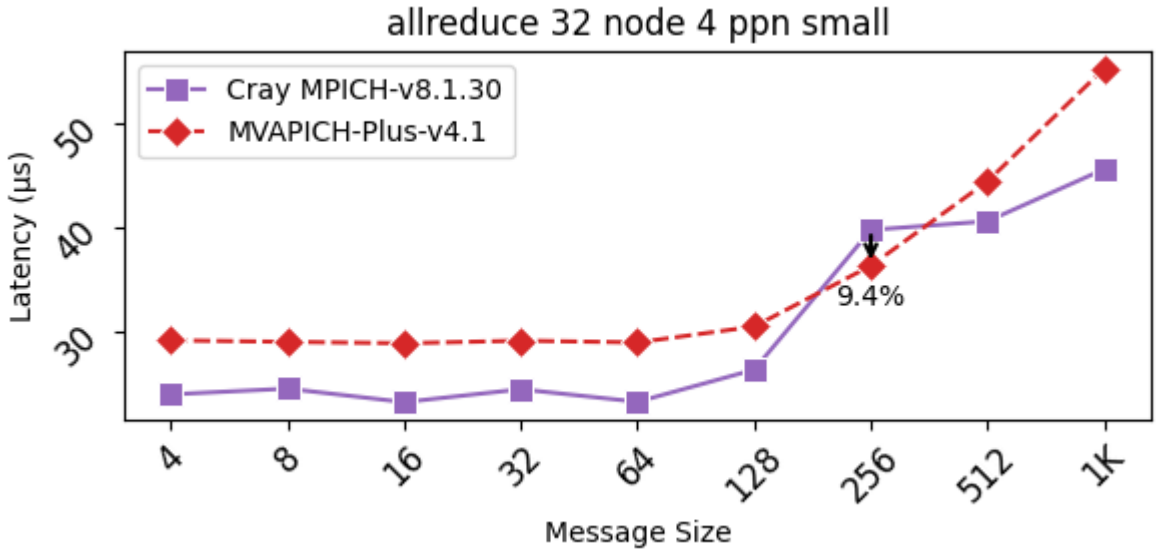
MVAPICH-Plus Performance on MI300A Collective 32 Node 4 PPN – GPU

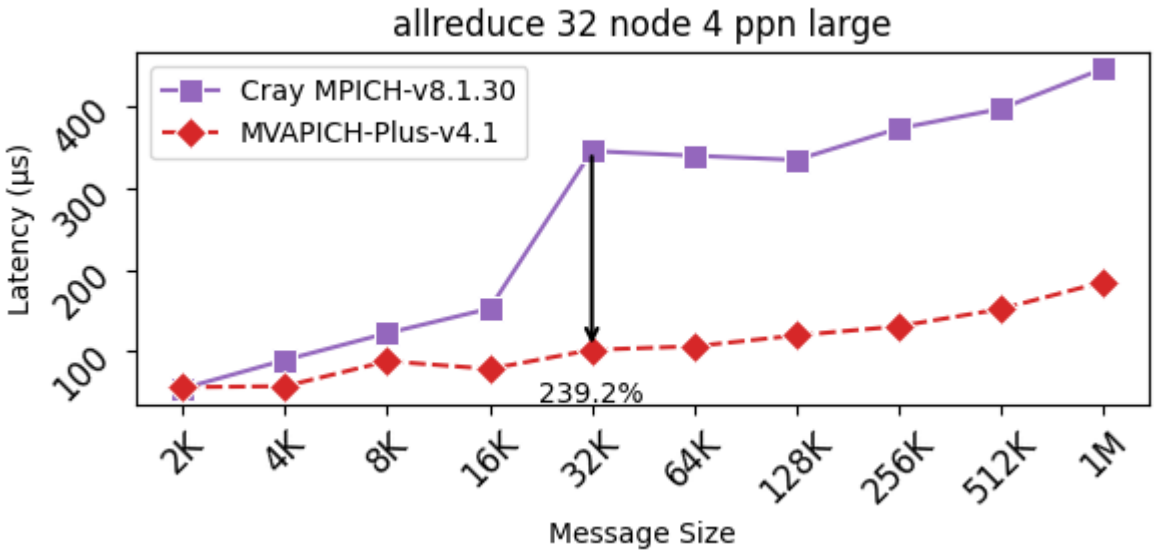
Allgather





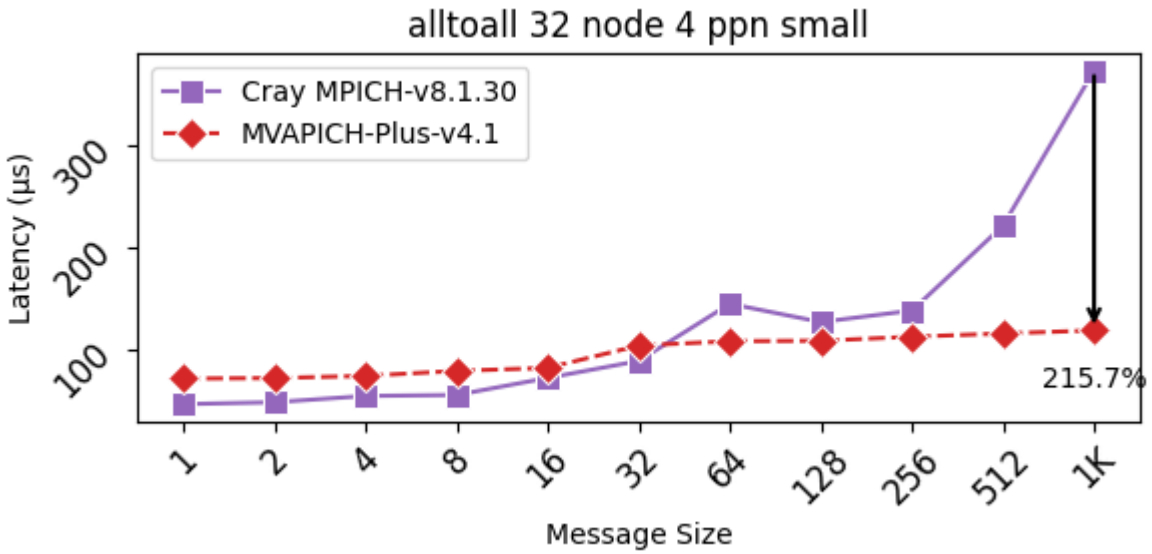
Allreduce

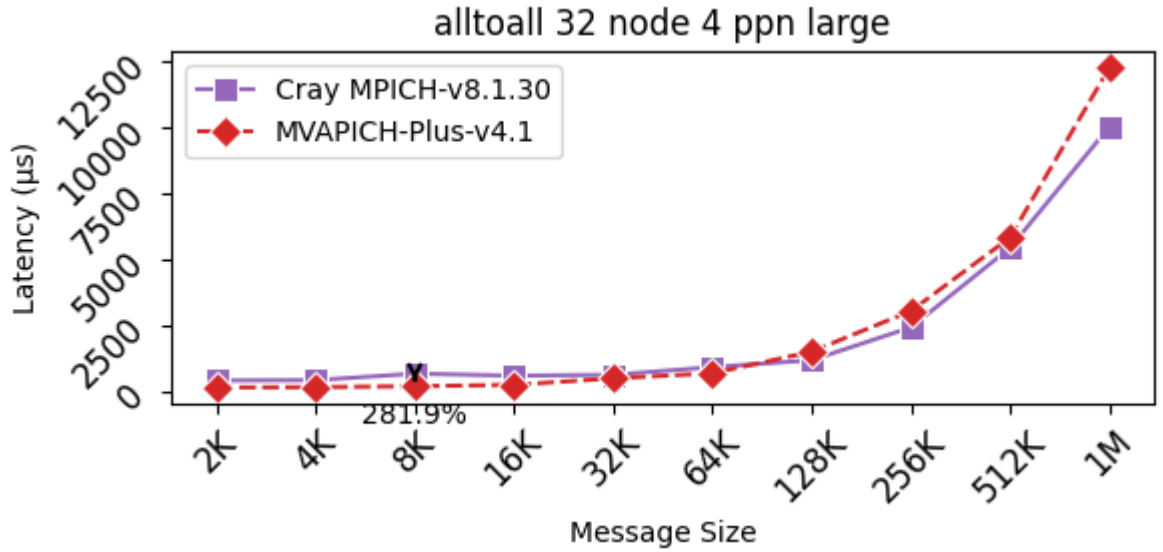




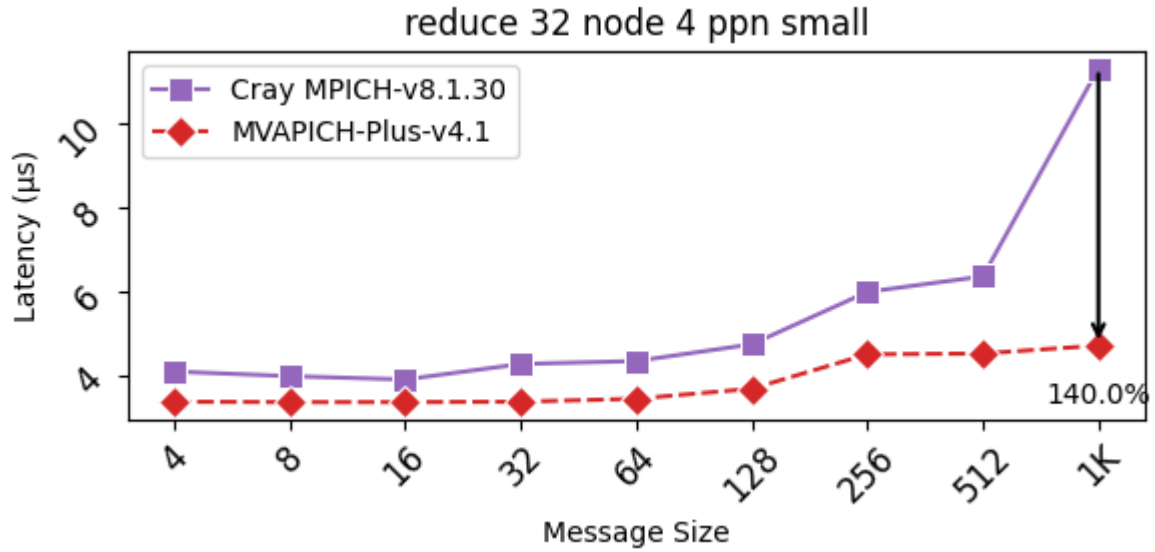
MVAPICH-Plus Performance on MI300A Collective 32 Node 4 PPN – GPU

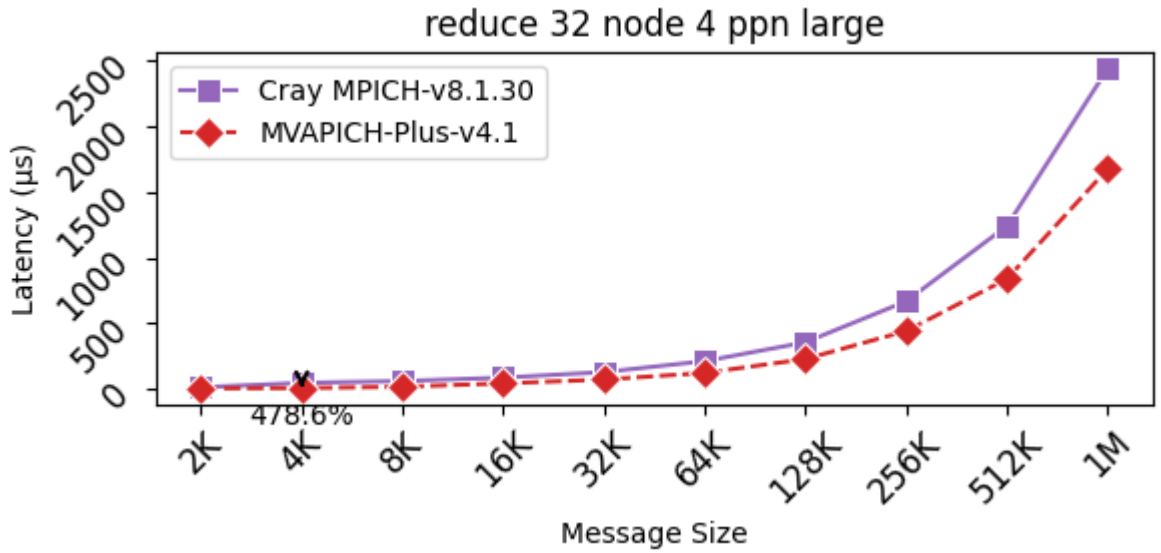
Alltoall





Reduce





Conclusion

- MI300A integrates CPU and GPU chiplets on a single package for true heterogeneous computing.
- Unified coherent HBM3 memory pool eliminates costly data transfers and simplifies programming
- Advanced chiplet design and Infinity Fabric interconnect provide high bandwidth and low latency

THANK YOU!

Questions?

Email : kuncham.2@osu.edu