MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

Performance of ROCm-aware MVAPICH2-GDR on LLNL Corona Cluster with AMD GPUs

MUG 2021

Kawthar Shafie Khorassani

shafiekhorassani.1@osu.edu

Network-based Computing Laboratory

Department of Computer Science and Engineering

The Ohio State University

The Ohio State University

Introduction

GPU-aware MPI libraries have been the driving force behind scaling of scientific and Deep Learning applications on GPU-enabled systems

- GPU-accelerator based computing dominated by NVIDIA GPUs & CUDA software stack

Adoption of AMD GPUs in large-scale HPC deployments (i.e. Frontier and El Capitan)

- Radeon Open Compute (ROCm) platform for AMD GPUs
- Lack of support for High-performance communication stacks with AMD
- Need for an optimized ROCm-aware MPI to exploit the capabilities of AMD GPUs

K. Khorassani, J. Hashmi, C. Chu, C. Chen, H. Subramoni, D. Panda Designing a ROCm-aware MPI Library for AMD GPUs: Early Experiences - ISC HIGH PERFORMANCE 2021, Jun 2021.

Background

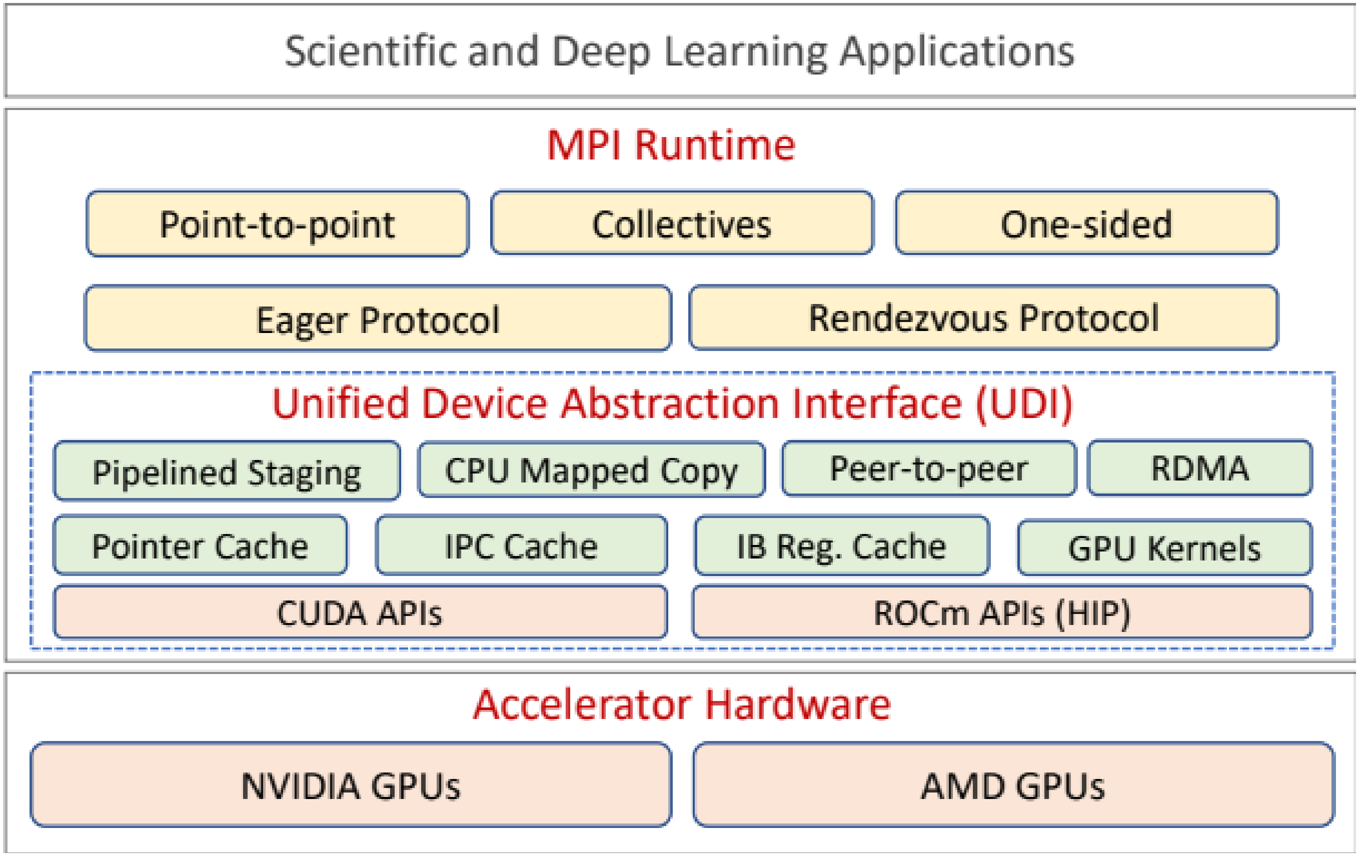
Radeon Open Compute (ROCm)

- AMD developed ROCm to achieve efficient computation and communication performance for applications running on AMD GPUs.
- ROCm platform is an open-source software for AMD GPUs
  - https://github.com/RadeonOpenCompute/ROCm

ROCm-aware MPI –

- Integrate the ROCm runtime into GPU-aware MPI Libraries (i.e. MVAPICH2-GDR) to utilize over AMD GPUs

ROCm-aware MVAPICH2-GDR: Unified Device Abstraction Interface



Scientific and Deep Learning Applications

MPI Runtime

Point-to-point

Collectives

One-sided

Eager Protocol

Rendezvous Protocol

Unified Device Abstraction Interface (UDI)

Pipelined Staging

CPU Mapped Copy

Peer-to-peer

RDMA

Pointer Cache

IPC Cache

IB Reg. Cache

GPU Kernels

CUDA APIs

ROCm APIs (HIP)

Accelerator Hardware

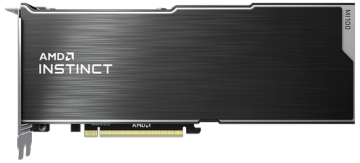
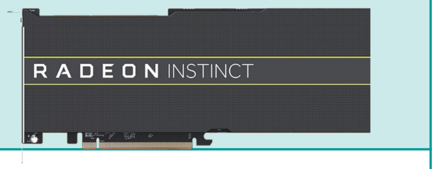
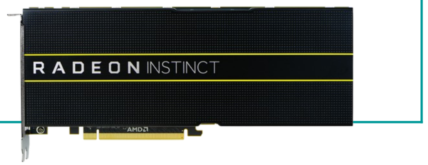
NVIDIA GPUs

AMD GPUs

- **UDI layer in MPI** that abstracts the common operations in a GPU-aware MPI runtime
- Modular design makes it easy to interface with vendor-specific backend implementations (i.e. CUDA or ROCm (HIP) APIs)

K. Khorassani, J. Hashmi, C. Chu, C. Chen, H. Subramoni, D. Panda Designing a ROCm-aware MPI Library for AMD GPUs: Early Experiences - ISC HIGH PERFORMANCE 2021, Jun 2021.

AMD GPUs - MI Series

AMD Instinct™ MI100 Accelerator	- CDNA GPU Architecture - Peak Single-precision (FP32) Performance – **23.1 TFLOPs** - 32 GB HBM2	
AMD Radeon Instinct™ MI50 Accelerator (32GB)	- Vega20 Architecture - Peak Single-precision (FP32) Performance – **13.3 TFLOPs** - 32 GB HBM2	
Radeon Instinct™ MI25 Accelerator	- Vega GPU Architecture - Peak Single-precision (FP32) Performance – **12.29 TFLOPs** - 16 GB HBM2	

Experimental Setup

- Utilized point-to-point and collective benchmarks from the OSU-Microbenchmarks 5.8 suite with ROCm extensions for evaluation on AMD GPUs
  - http://mvapich.cse.ohio-state.edu/benchmarks/
  - `MV2_USE_ROCM=1` and `-d rocm` passed to benchmark
- Corona Cluster at Lawrence Livermore National Laboratory (LLNL)
  - 291 AMD EPYC 7402 series CPU nodes
  - 82 nodes with **4 MI50 AMD GPUs** per node
  - 82 nodes with **4 MI60 AMD GPUs** per node
  - 123 nodes with **8 MI50 AMD GPUs** per node
  - Dual-socket Mellanox IB HDR-100
  - ROCm Version 4.3.0
- ROCm-aware MVAPICH2-GDR v2.3.6
  - http://mvapich.cse.ohio-state.edu/downloads/
- OpenMPI 4.1.1 + UCX 1.11.0
  - https://www.open-mpi.org

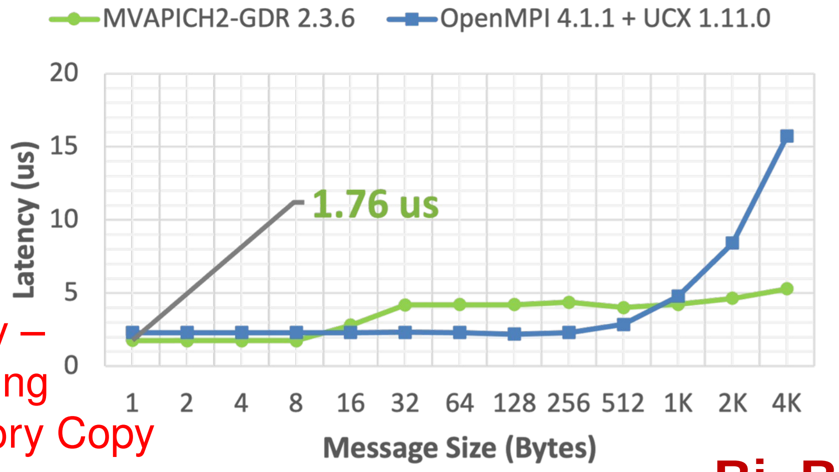
Peak Achievable Performance

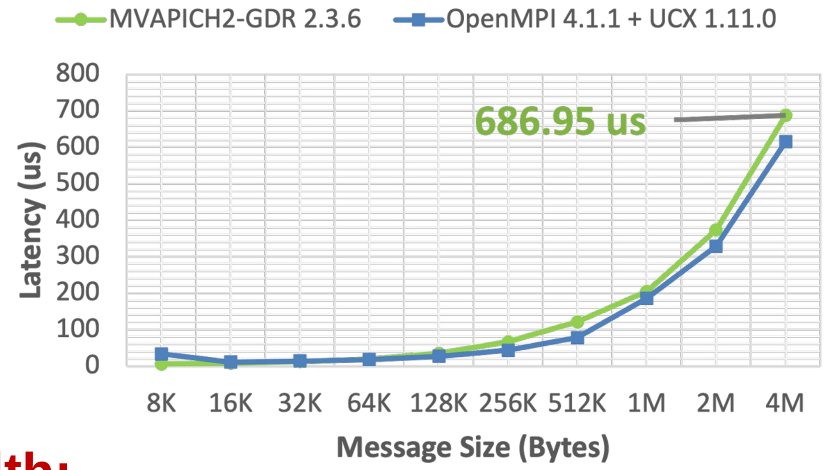
To evaluate the performance on AMD GPUs compared to the peak achievable performance:

- **ROCm Bandwidth Test:** evaluate the performance between two GPUs on a node (displays the peak achievable bandwidth by performing a uni/bi -directional copy involving two devices).
  - https://github.com/RadeonOpenCompute/rocm_bandwidth_test
- **Infiniband Perftest:** utilize *ibreadbw* and *ibreadlat* to measure the peak achievable bandwidth and minimum achievable latency of communicating data across two nodes
  - https://github.com/linux-rdma/perftest

Intra-Node Pt2pt Performance

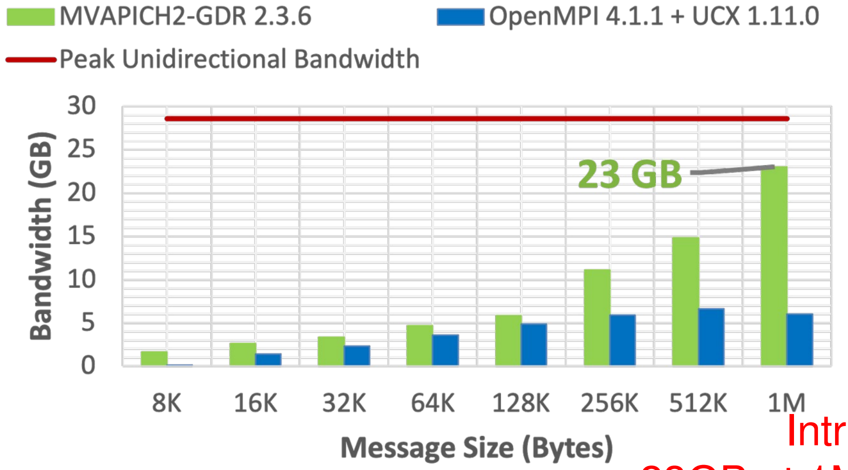
Latency:



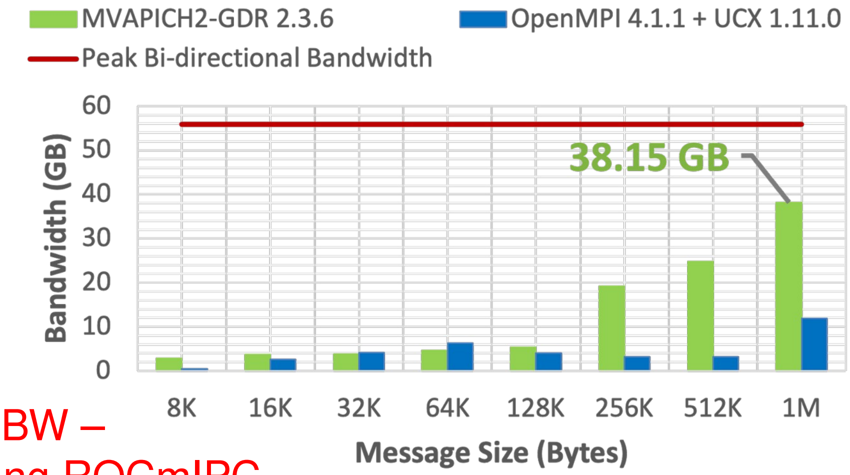


Intra-Node latency –
1.76us at 1B utilizing
CI Bar Mapped Memory Copy

Bandwidth:



Bi- Bandwidth:

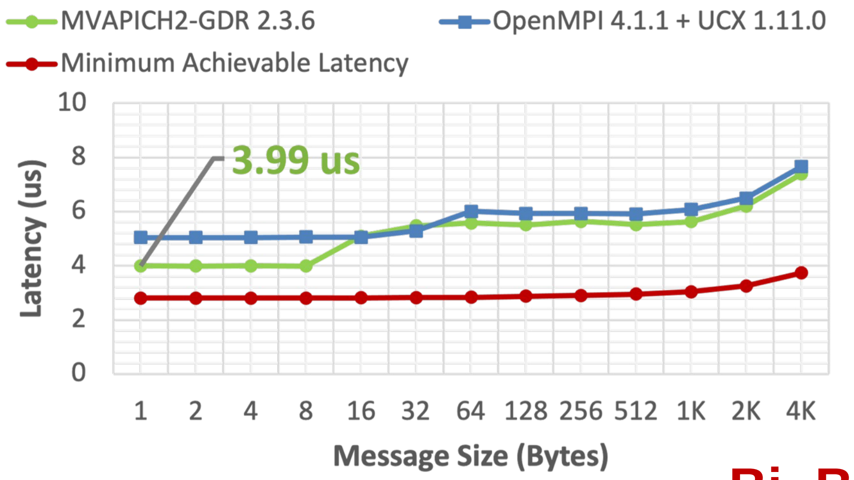


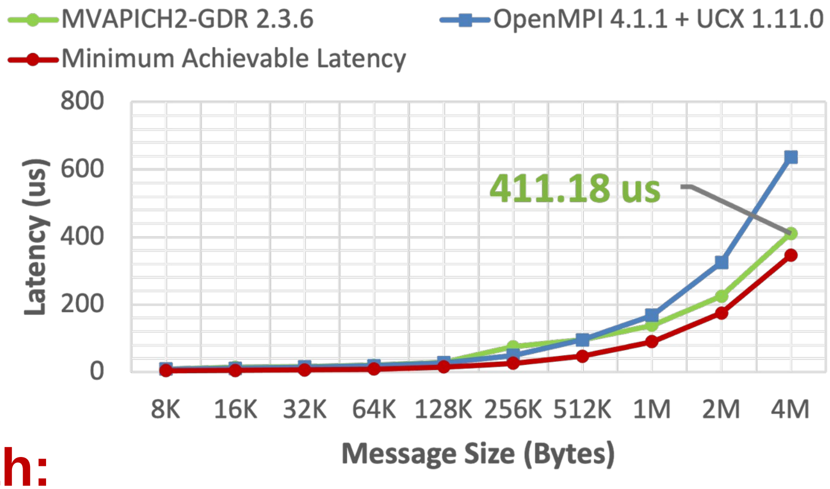
Intra-Node BW –
23GB at 1MB utilizing ROCmIPC

Corona Cluster – (mi50 GPUs) ROCm 4.3.0

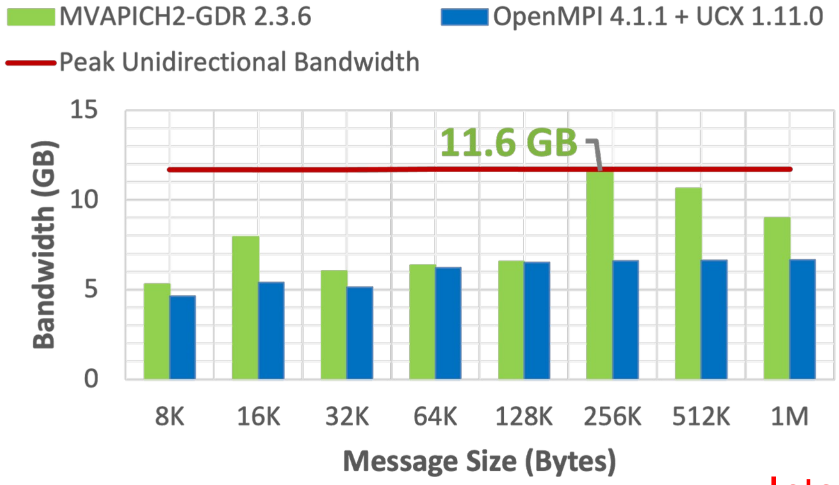
Inter-Node Pt2pt Performance

Latency:

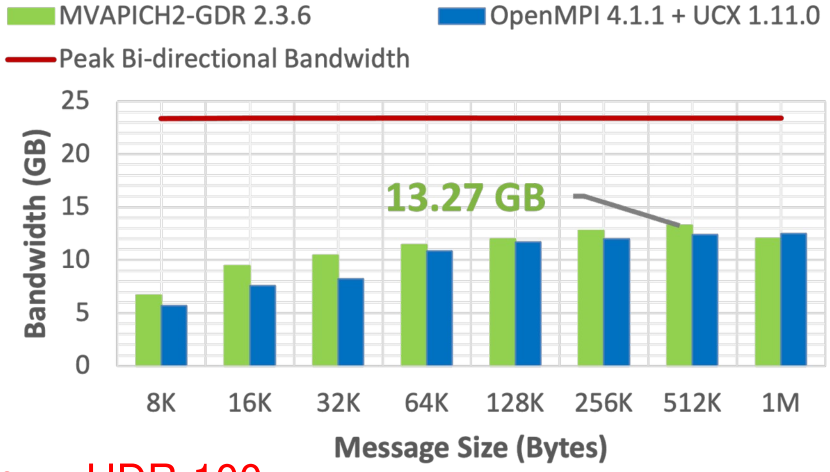




Bandwidth:



Bi- Bandwidth:

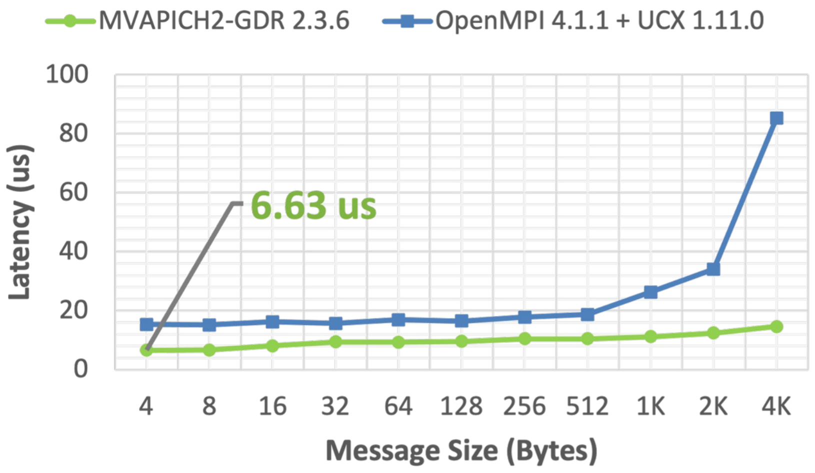


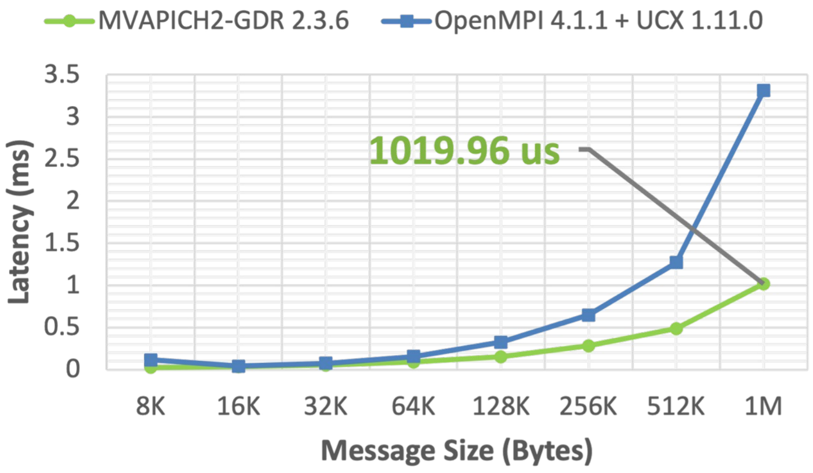
Inter-Node Latency – HDR-100 Interconnect, utilizing ROCmRDMA (PeerDirect)

Corona Cluster – (mi50 GPUs) ROCm 4.3.0

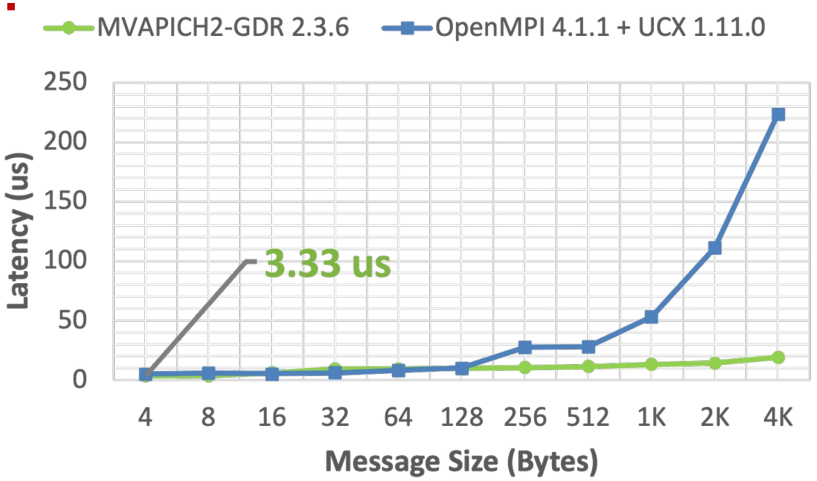
Collectives Performance

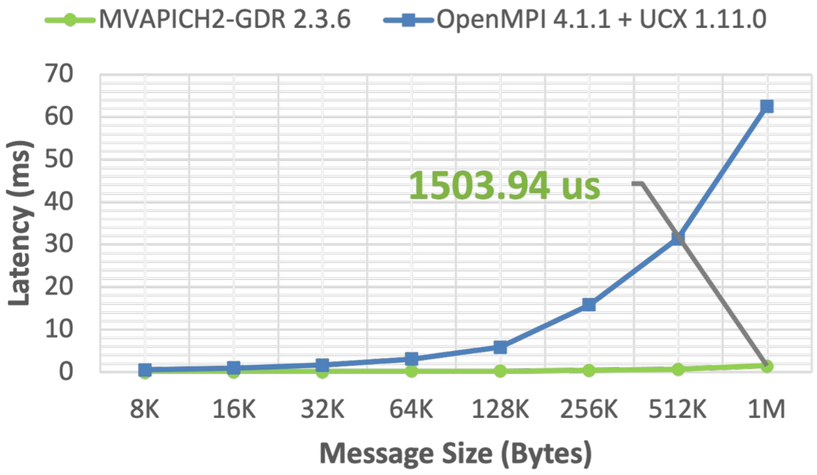
MPI_Bcast:





MPI_Reduce:

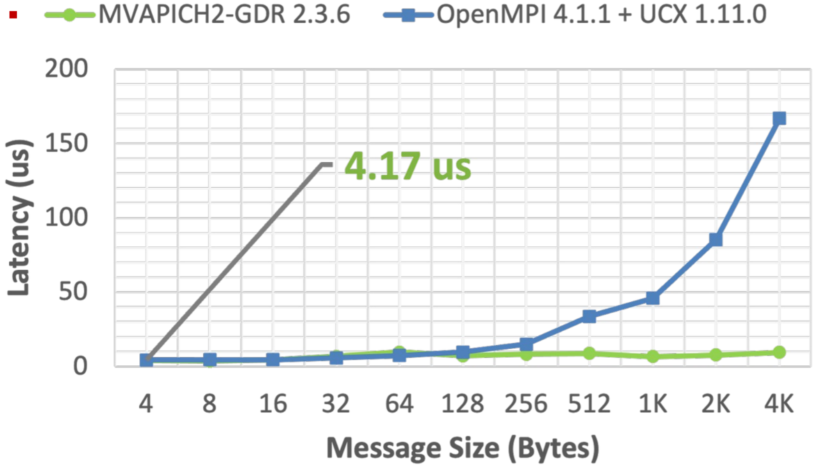


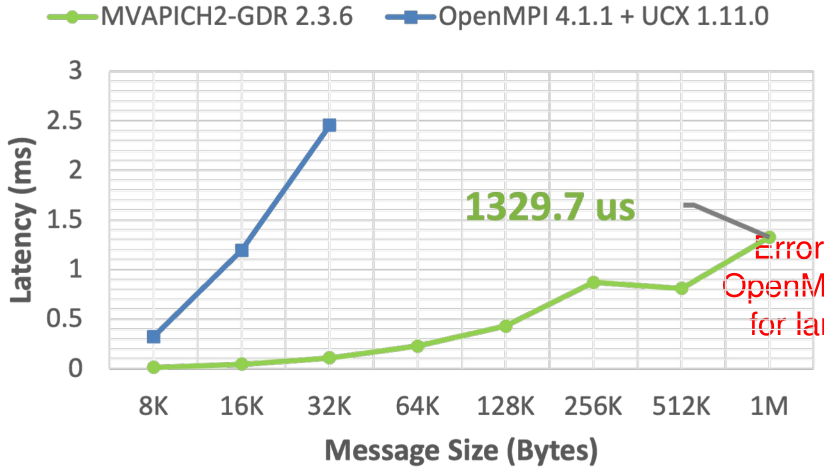


Corona Cluster – (mi50 GPUs) ROCm 4.3.0 **– 8** Nodes 8 PPN **(64 GPUs)**

Collectives Performance

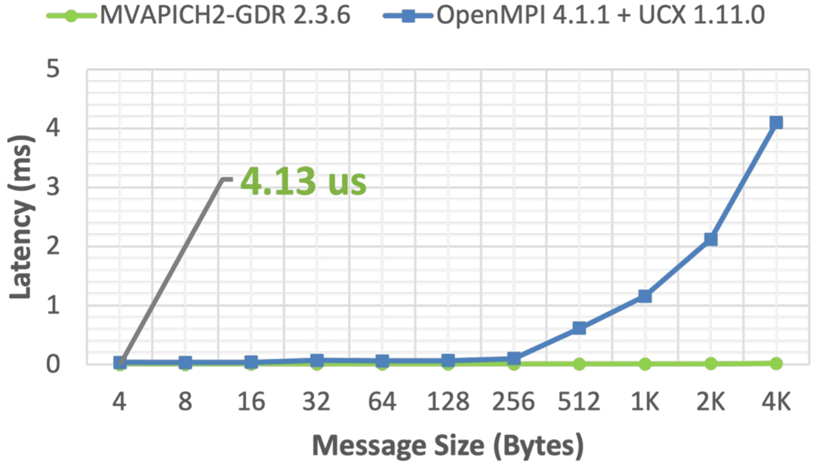
MPI_Gather:

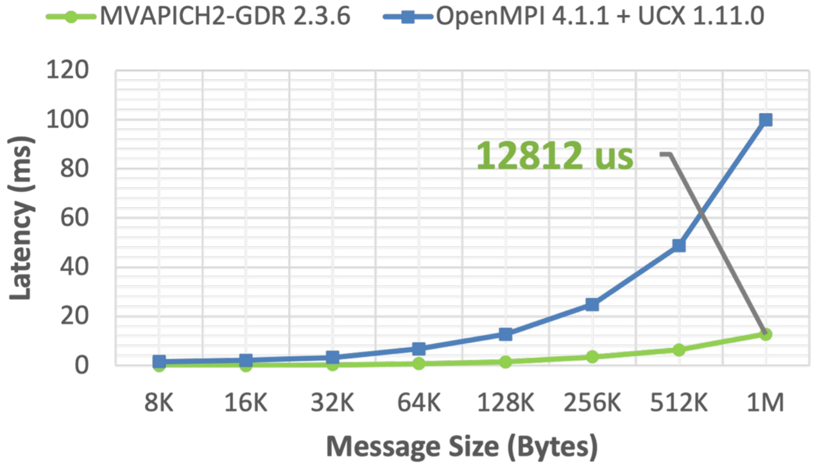




Error with Gather using OpenMPI 4.1.1 + UCX 1.11 for large message sizes

MPI_Allgather:

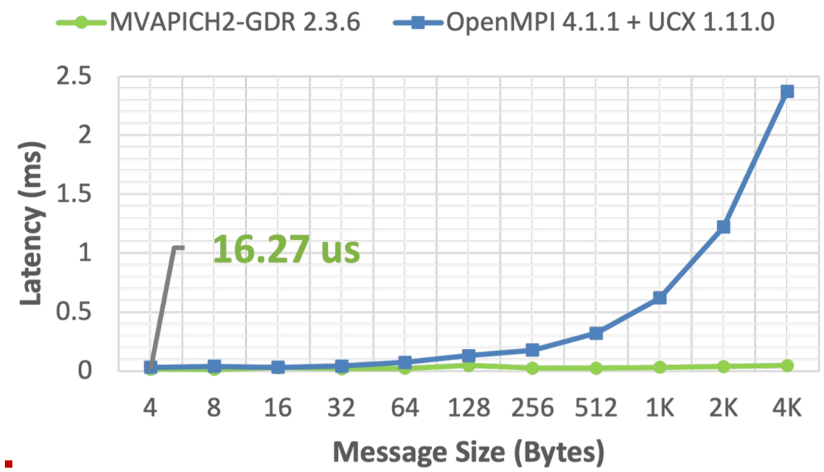


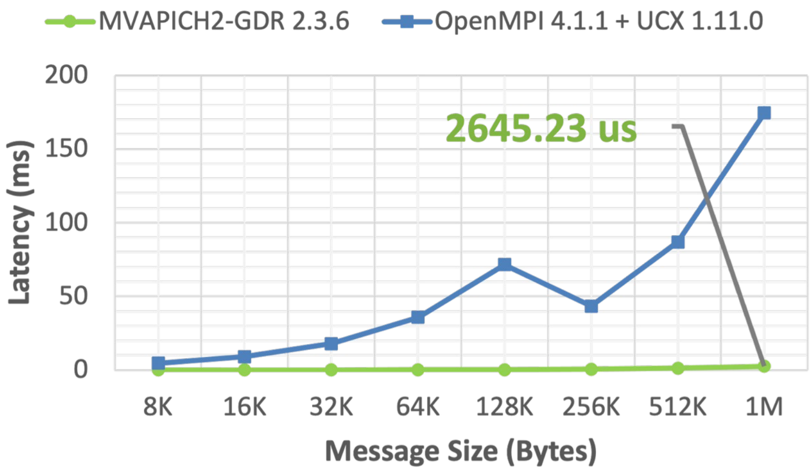


Corona Cluster – (mi50 GPUs) ROCm 4.3.0 **– 8** Nodes 8 PPN **(64 GPUs)**

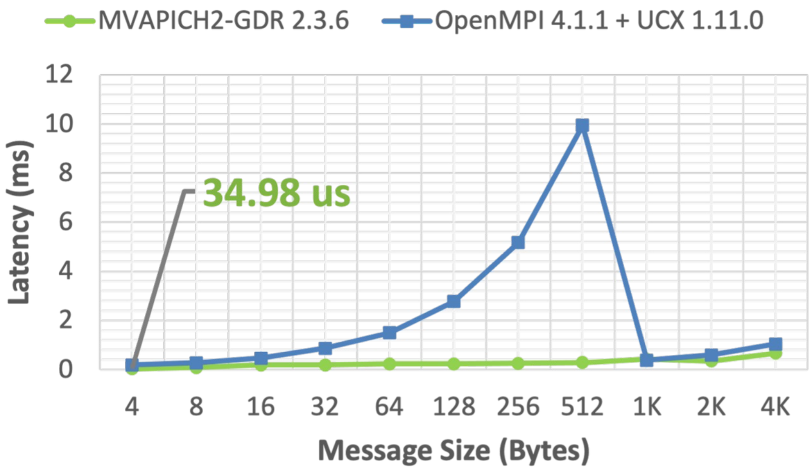
Collectives Performance

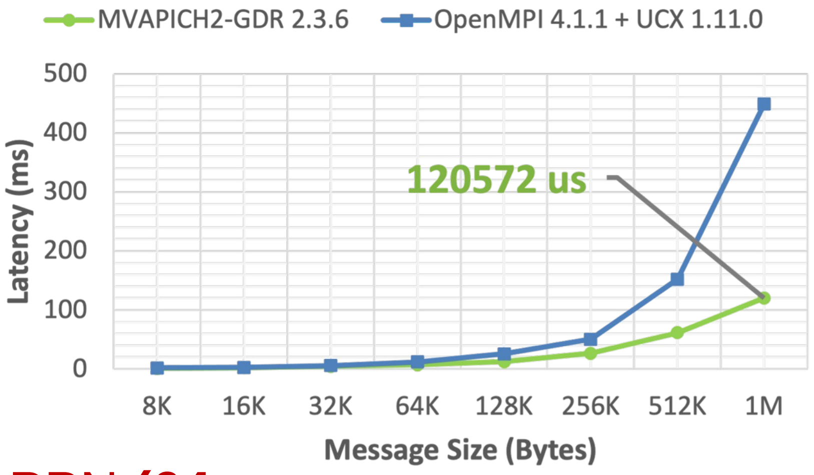
MPI_Allreduce:





MPI_Alltoall:





Corona Cluster – (mi50 GPUs) ROCm 4.3.0 **– 8** Nodes 8 PPN **(64 GPUs)**

Conclusion

Next-generation HPC systems such as Frontier and El Capitan adopting AMD GPUs

- important to ensure that scientific applications and the communication middleware such as MPI are supported and optimized for these systems through a ROCm-aware MPI runtime (i.e. MVAPICH2-GDR).
- Utilize Features provided by ROCm driver / Runtime (i.e. ROCmIPC, ROCmRDMA, Large Bar Feature, etc.) in MPI run-time

ROCm-aware MVAPICH2-GDR is available through releases MVAPICH2-GDR 2.3.5+ and optimizations expected in future releases.

THANK YOU!

Network Based Computing Laboratory

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/