2021 OFA Virtual Workshop

Designing a ROCm-aware MPI Library for AMD GPUs over High-speed Networks

Kawthar Shafie Khorassani, Jahanzeb Hashmi, Ching-Hsiang Chu, **Hari Subramoni**,
and Dhabaleswar K. (DK) Panda

The Ohio State University
subramon@cse.ohio-state.edu

OUTLINE

- **Introduction**
- GPU-aware Communication Features
- ROCm-aware MPI
- Performance of ROCm-aware MPI
- Conclusion and Future Work

INTRODUCTION

GPU-aware MPI libraries have been the driving force behind scaling applications on GPU-enabled HPC and cloud systems

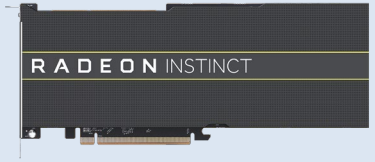
- Dominated by NVIDIA GPUs and CUDA software stack

The lack of support for high-performance communication stacks with AMD limited the adoption of AMD GPUs in large-scale HPC deployments

- Upcoming DOE exascale systems (Frontier and El Capitan) will be driven by AMD GPUs
- Radeon Open Compute (ROCm) platform for AMD GPUs
- Need for a ROCm-aware MPI to exploit the capabilities of AMD GPUs

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

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	

https://www.amd.com/en/graphics/servers-radeon-instinct-mi

OUTLINE

- Introduction
- **GPU-aware Communication Features**
- ROCm-aware MPI
- Performance of ROCm-aware MPI
- Conclusion and Future Work

GPU-AWARE MPI LIBRARY (MVAPICH2-GDR)

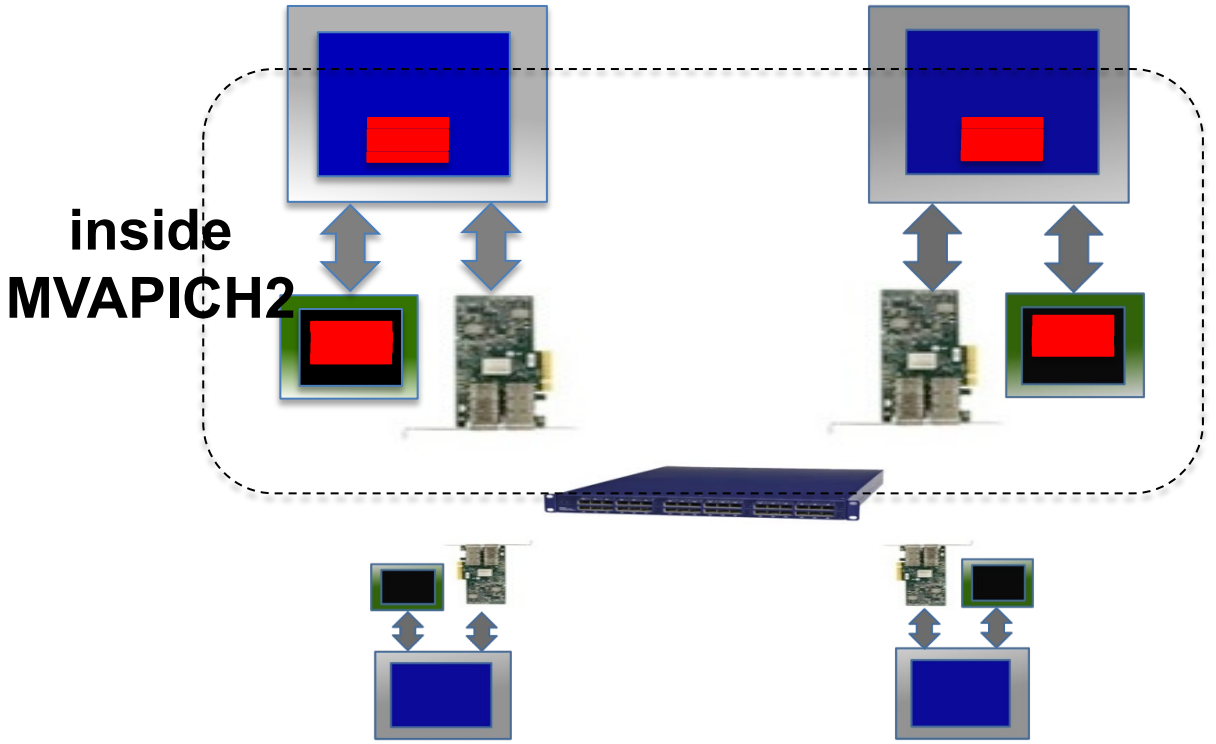
- Standard MPI interface used for GPU-aware data movement
- Overlaps data movement from GPU with RDMA transfers

At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

```
MPI_Recv(r_devbuf, size, …);
```



GPU-AWARE MPI LIBRARIES - COMMUNICATION FEATURES

State-of-the-art GPU-aware MPI libraries have support for NVIDIA GPUs through the CUDA toolkit. They utilize features including:

- NVIDIA GPUDirect RDMA
- CUDA Inter-Process Communication (IPC)
- GDRCopy

AMD GPUs utilizing the ROCm Driver and Run-time:

- ROCm RDMA (PeerDirect)
- ROCm IPC
- Large BAR Feature

OUTLINE

- Introduction
- GPU-aware Communication Features
- **ROCm-aware MPI**
- Performance of ROCm-aware MPI
- Conclusion and Future Work

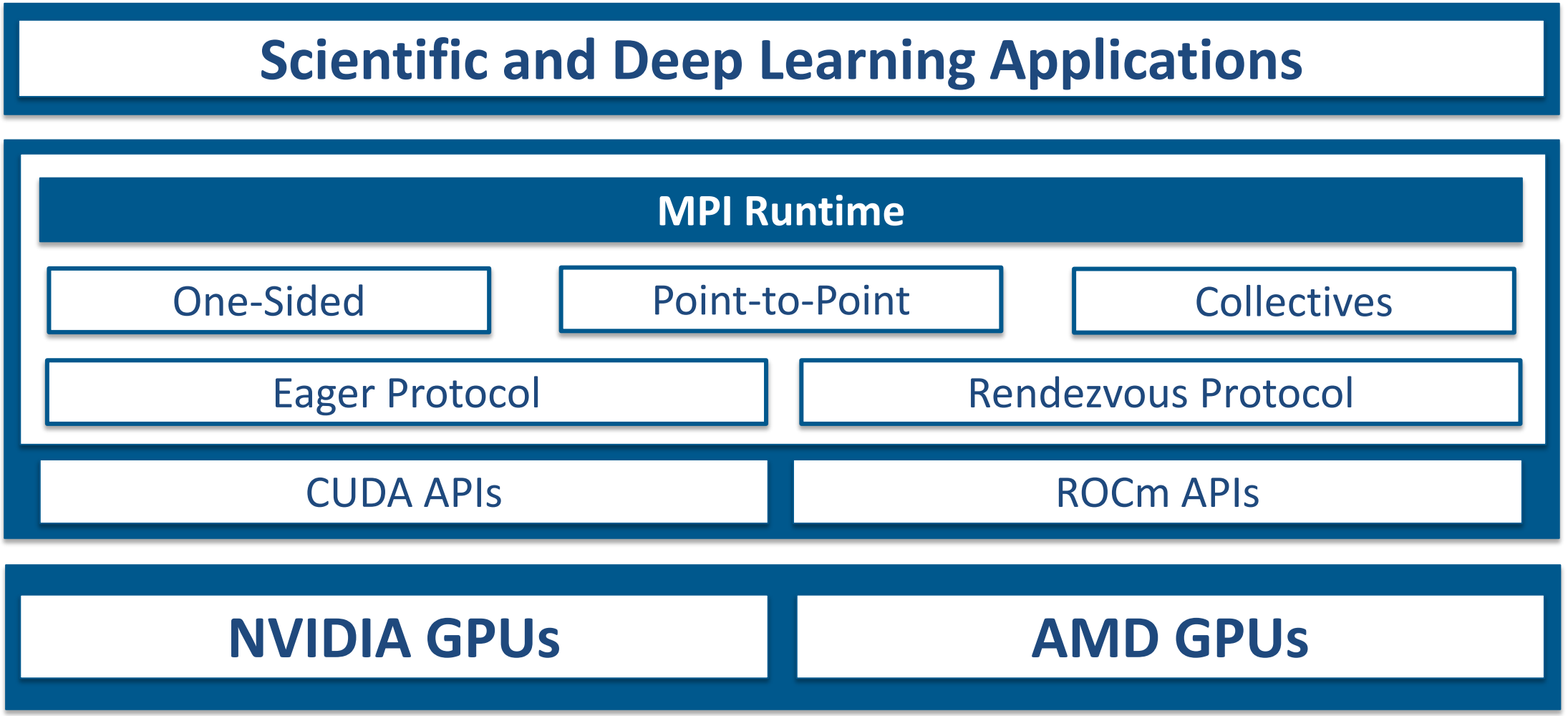
RADEON OPEN COMPUTE (ROCM)

- AMD developed Radeon Open Compute (ROCm) tailored towards achieving 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

- **Design and demonstrate early performance results of a ROCm-aware MPI library (MVAPICH2-GDR) that utilizes high-performance communication features like ROCmRDMA to achieve scalable multi-GPU performance**

INTEGRATING ROCM INTO MPI (MVAPICH2-GDR 2.3.5)



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), and AWS EFA
- **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 GPGPUs, since 2014
  - **MVAPICH2-GDR with support for AMD GPUs since MVAPICH2-GDR-2.3.5**
  - 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
    - **Support for AMD GPUs with ROCm-aware MPI in Release Version 5.7**
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015

- Used by more than 3,150 organizations in 89 countries
- More than 1.26 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov '20 ranking)
  - 4th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 9th, 448, 448 cores (Frontera) at TACC
  - 14th, 391,680 cores (ABCI) in Japan
  - 21st, 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 9th ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

OUTLINE

- Introduction
- GPU-aware Communication Features
- ROCm-aware MPI
- **Performance of ROCm-aware MPI**
- Conclusion and Future Work

MVAPICH2-GDR 2.3.5 WITH ROCM-AWARE MPI SUPPORT

MVAPICH2-GDR Release Version 2.3.5 - AMD GPU ROCm-aware MPI Support

Features:

- Support for AMD GPUs via Radeon Open Compute (ROCm) platform
- Support for ROCm PeerDirect, ROCm IPC, and unified memory based device-to-device communication for AMD GPUs
- GPU-based point-to-point tuning for AMD Mi50 and Mi60 GPUs
- http://mvapich.cse.ohio-state.edu/downloads/

OSU-Microbenchmarks Version 5.7

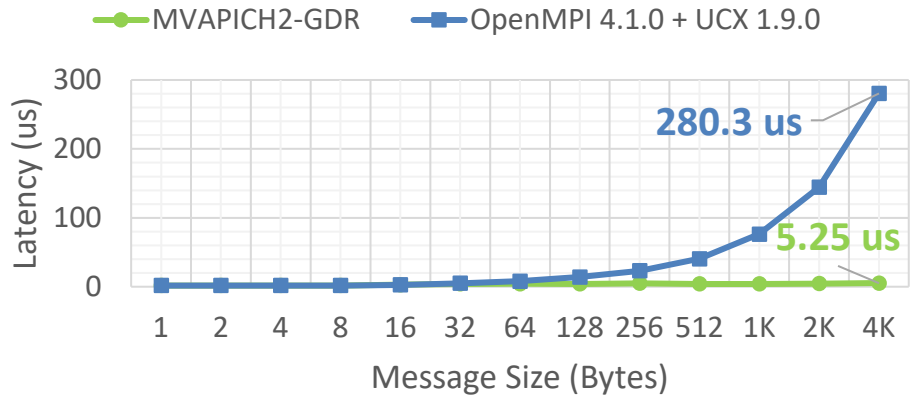
- Add support to OMB to evaluate the performance of various primitives with AMD GPU device and ROCm support
- http://mvapich.cse.ohio-state.edu/benchmarks/

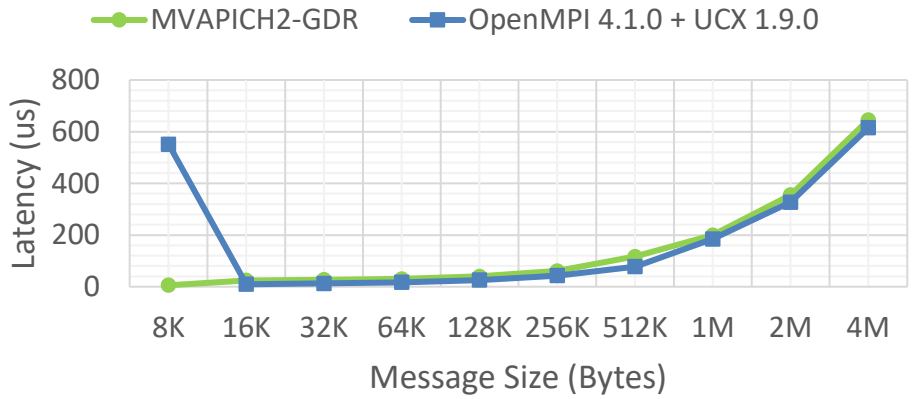
EXPERIMENTAL SETUP

- **Utilized point-to-point and collective benchmarks from the OSU-Microbenchmarks 5.7 suite with ROCm extensions for evaluation on AMD GPUs**
  - http://mvapich.cse.ohio-state.edu/benchmarks/
- **Corona Cluster at Lawrence Livermore National Laboratory (LLNL)**
  - 291 AMD EPYC 7002 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-200
  - Mellanox OFED 5.0
  - ROCm Version 3.10.0
- **MVAPICH2-GDR 2.3.5 with ROCm-aware MPI**
  - http://mvapich.cse.ohio-state.edu/downloads/
- **OpenMPI 4.0.1 + UCX 1.9.0**
  - https://www.open-mpi.org

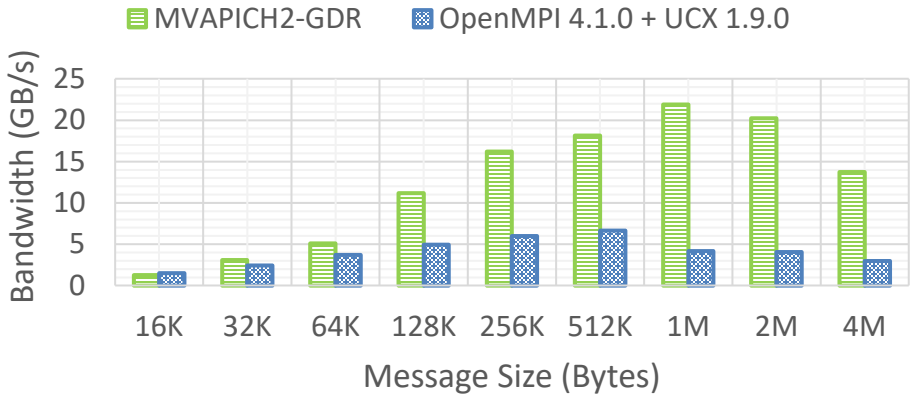
INTRA-NODE – MVAPICH2-GDR & OPENMPI + UCX

Latency: MVAPICH2-GDR intra-node latency at 1 Byte utilizing **PCI Bar Mapped Memory** is 1.78 us

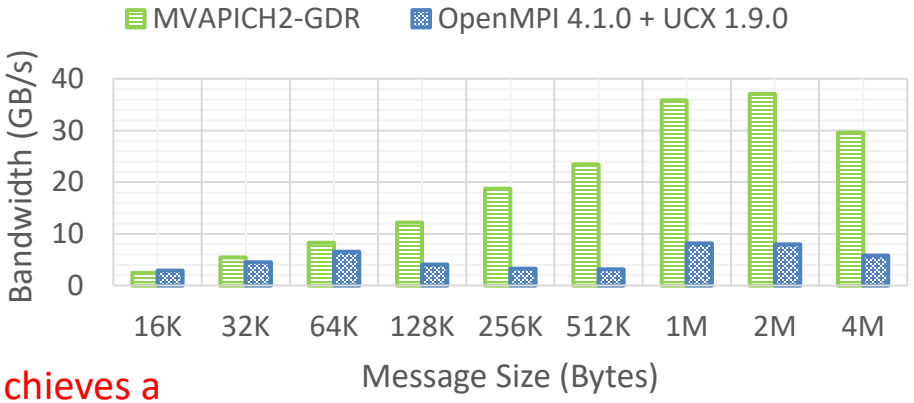




Bandwidth:



Bi-Directional Bandwidth:

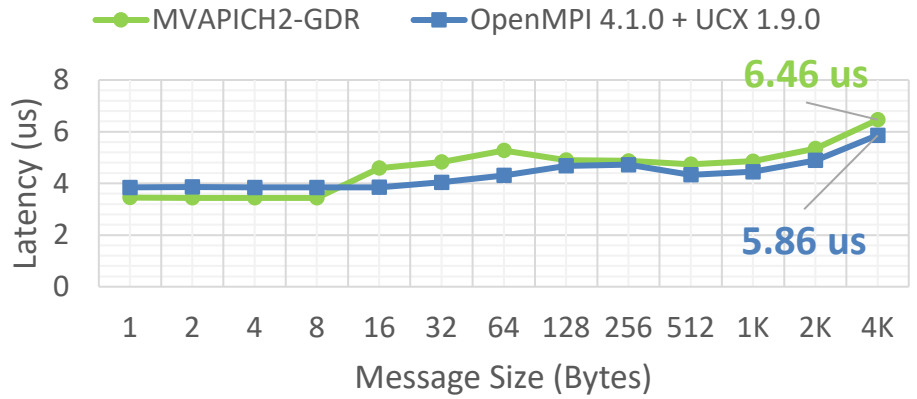


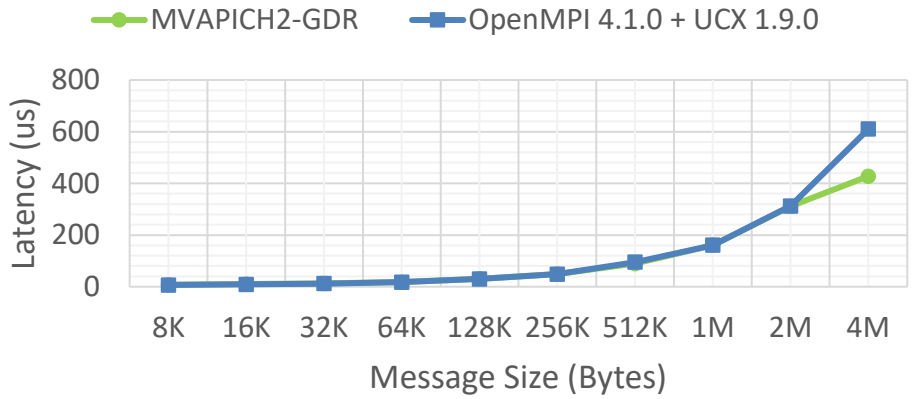
Corona Cluster – (mi50 GPUs) ROCm 3.10

MVAPICH2-GDR achieves a bandwidth of ~21.9 GB/s at 1MB utilizing **ROCm IPC**

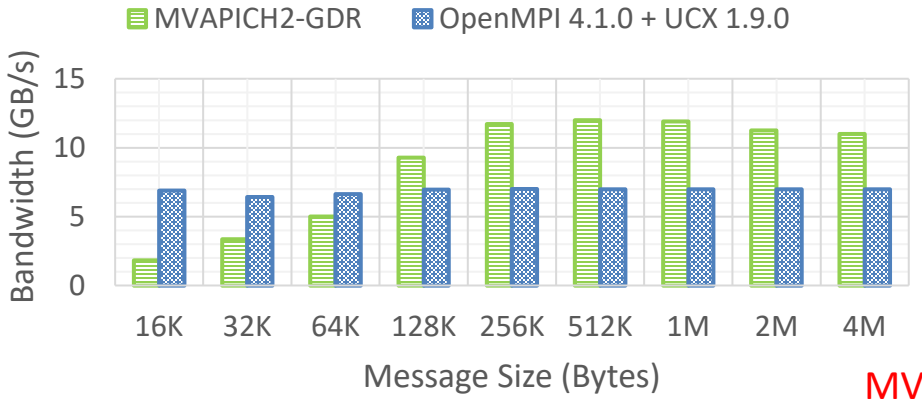
INTER-NODE – MVAPICH2-GDR & OPENMPI + UCX

Latency:

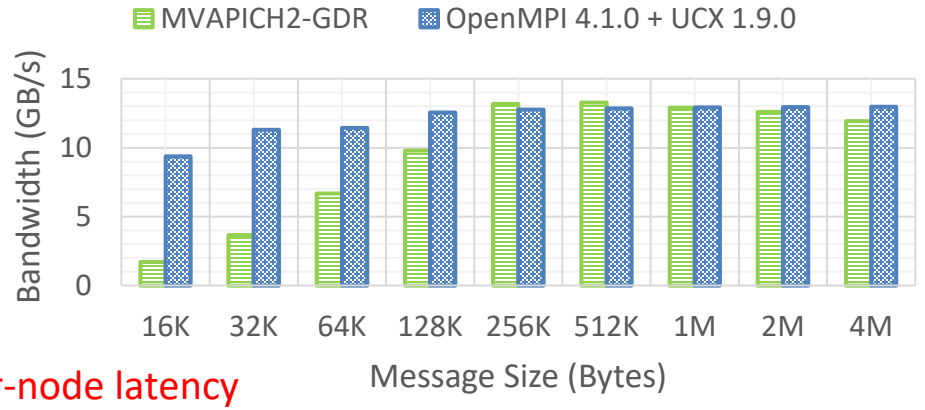




Bandwidth:



Bi-Directional Bandwidth:



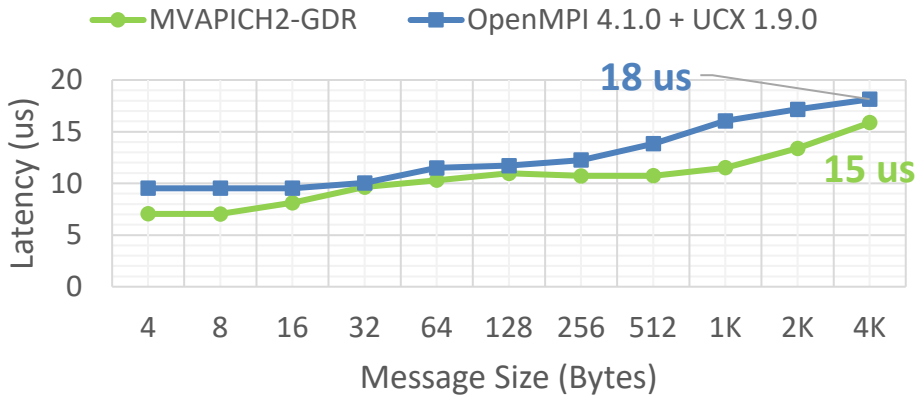
Corona Cluster – (mi50 GPUs) ROCm 3.10

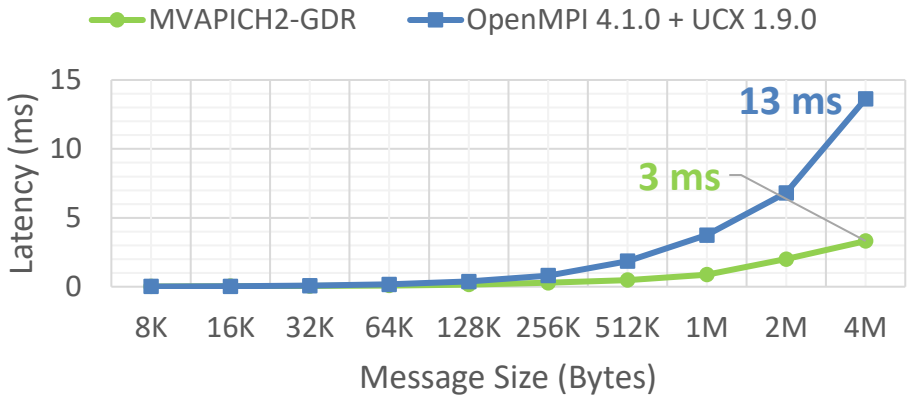
MVAPICH2-GDR inter-node latency at 1 Byte is 3.45 us and achieves a bandwidth of ~12 GB/s at 512 KB

COLLECTIVES – MVAPICH2-GDR & OPENMPI + UCX

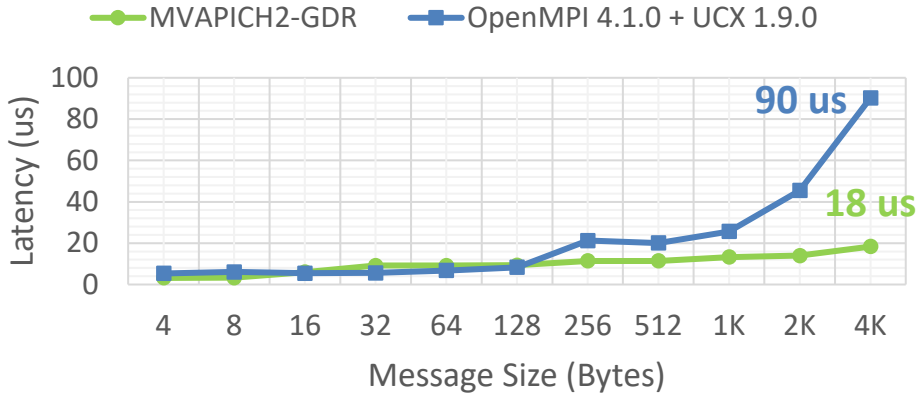
BROADCAST:

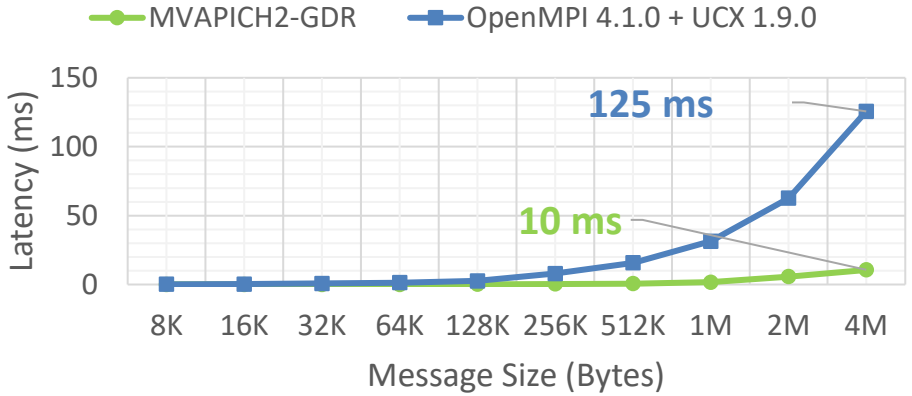
BCAST latency of 7.06 us at 4B and 3 ms latency at 4MB





REDUCE:





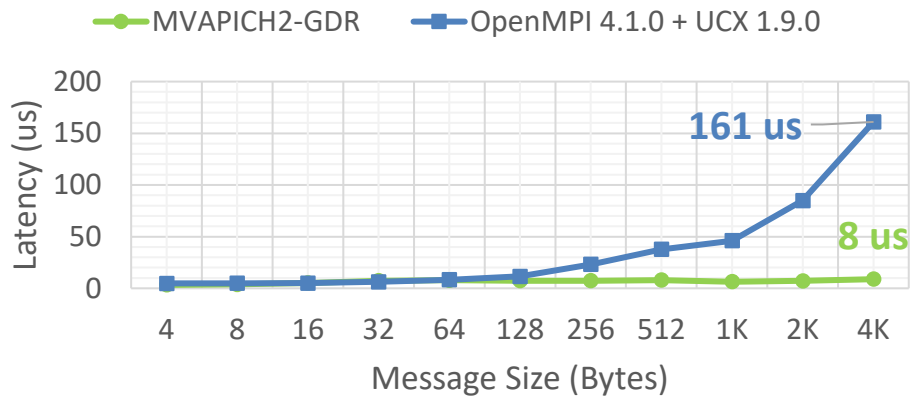
REDUCE latency of 3.21 us at 4B and 10 ms latency at 4MB

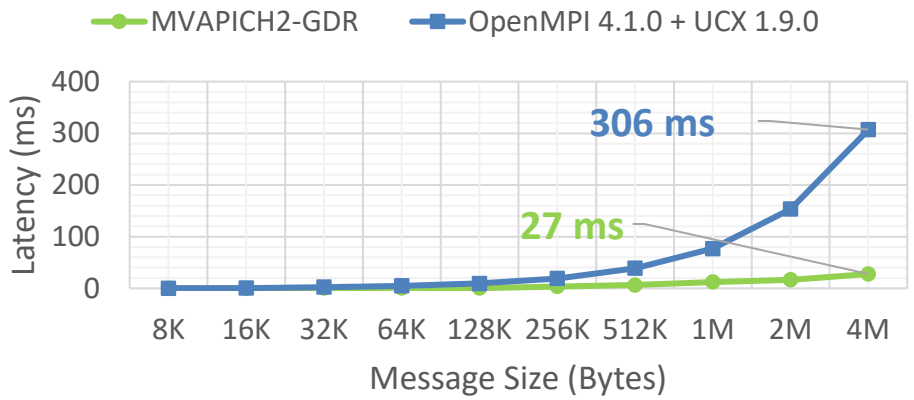
Corona Cluster – (mi50 GPUs) ROCm 3.10 **– 64 NP (8 Nodes, 8 GPUs per Node)**

COLLECTIVES – MVAPICH2-GDR & OPENMPI + UCX

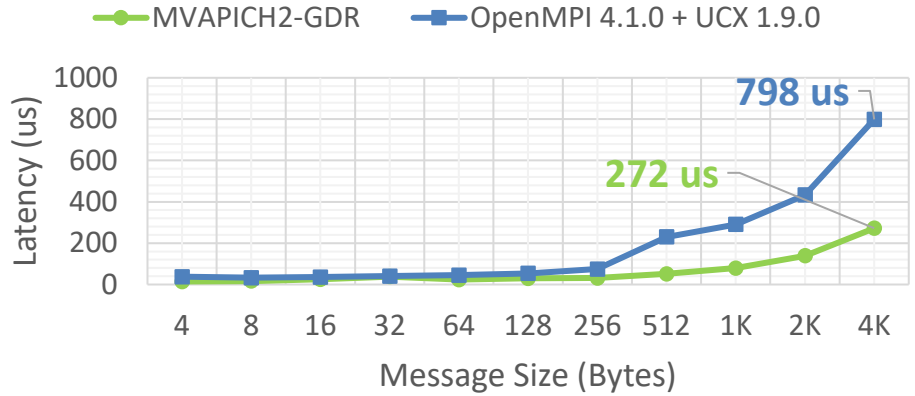
GATHER:

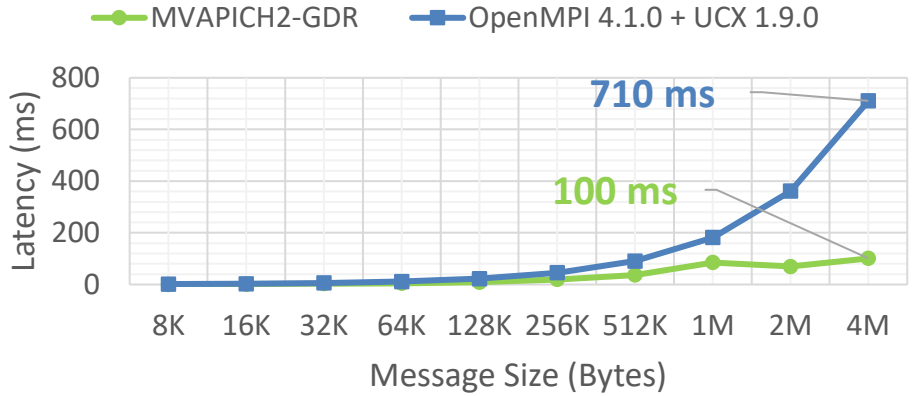
Gather latency of 27 ms at 4MB and **Allgather** latency of 100 ms at 4MB





ALLGATHER:



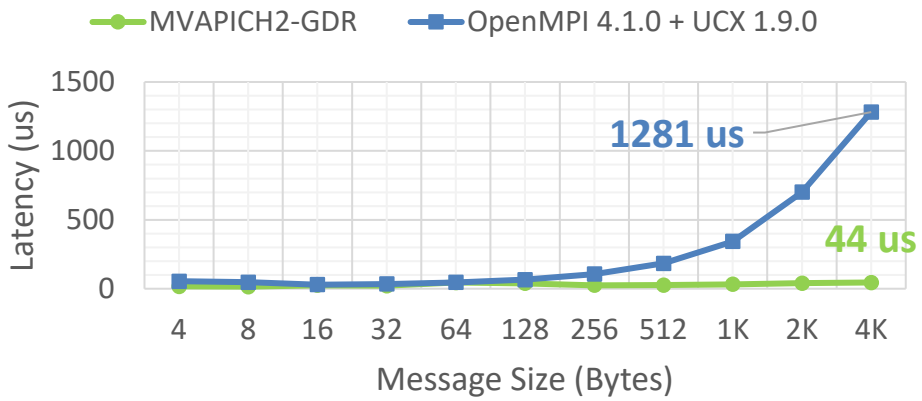


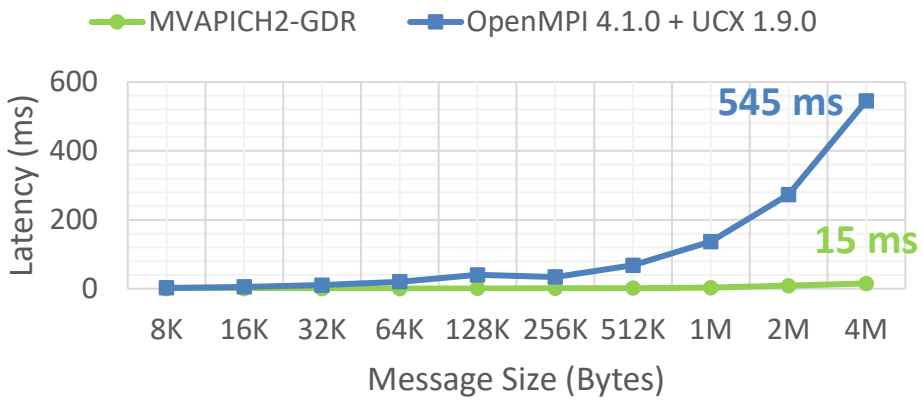
Corona Cluster – (mi50 GPUs) ROCm 3.10 **– 64 NP (8 Nodes, 8 GPUs per Node)**

COLLECTIVES – MVAPICH2-GDR & OPENMPI + UCX

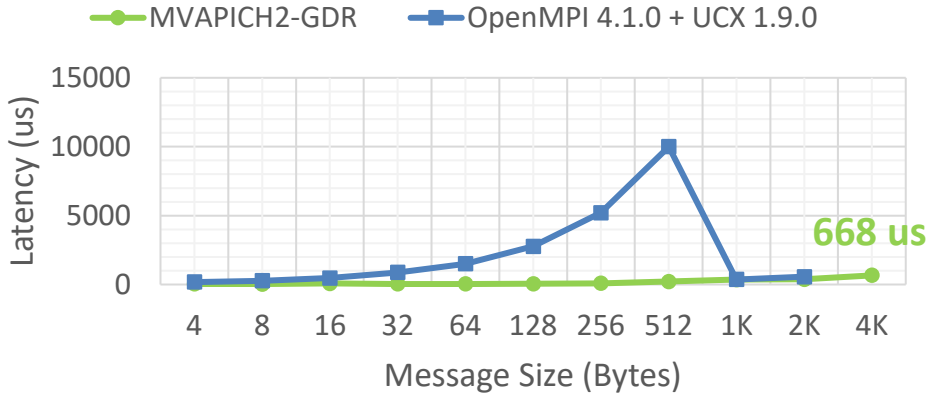
ALLREDUCE:

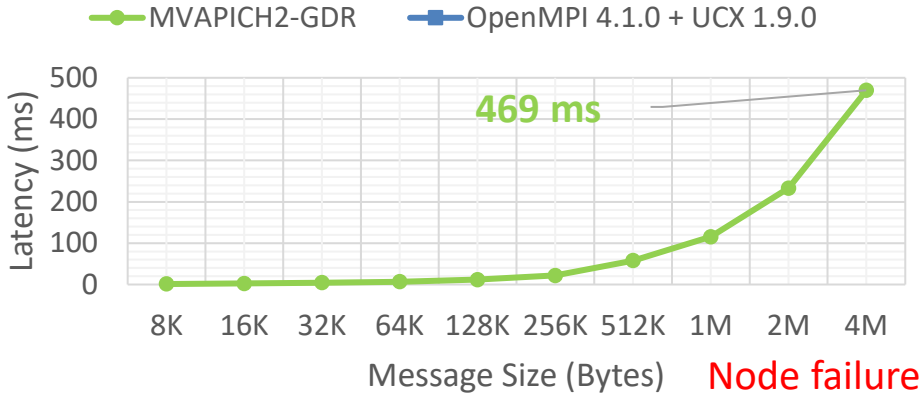
Tuned **Allreduce** in next release of MVAPICH2-GDR achieving **15 ms latency at 4MB**





ALLTOALL:





Node failure with Alltoall using OpenMPI for large message sizes

Corona Cluster – (mi50 GPUs) ROCm 3.10 **– 64 NP (8 Nodes, 8 GPUs per Node)**

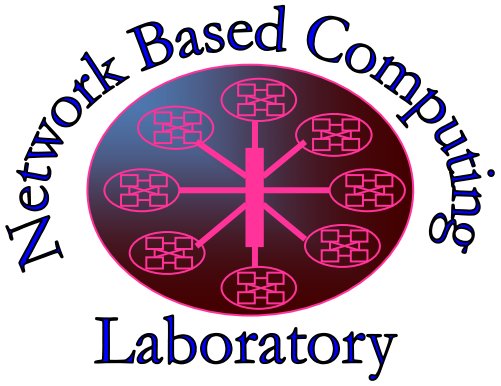
OUTLINE

- Introduction
- GPU-aware Communication Features
- ROCm-aware MPI
- Performance of ROCm-aware MPI
- **Conclusion and Future Work**

CONCLUSION AND FUTURE WORK

- ROCm-aware MPI through MVAPICH2-GDR to exploit AMD GPUs
- Utilize the communication features offered by the ROCm driver and run-time to integrate support for
  - ROCm RDMA, ROCm IPC, and Large BAR feature
- Optimized communication performance on AMD GPUs for point-to-point and collective MPI communication
- Demonstrate scalability and performance of the MVAPICH2-GDR ROCm-aware Feature
- Utilize MVAPICH2-GDR ROCm-aware Feature for Hipified Applications and Deep Learning Frameworks

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/