MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

Performance of MVAPICH2-GDR on DGX A100

Chen-Chun Chen

chen.10252@osu.edu

MVAPICH User Group Conference (MUG) 2021

Network Based Computing Laboratory (NBCL)

The Ohio State University

Introduction

- With the advent of the NVIDIA DGX A100 system, MVAPICH2-GDR 2.3.6 includes optimized support for pt2pt and collective communication
- MVAPICH2-GDR performance on NVIDIA DGX A100 system
- NVIDIA DGX A100 Topology

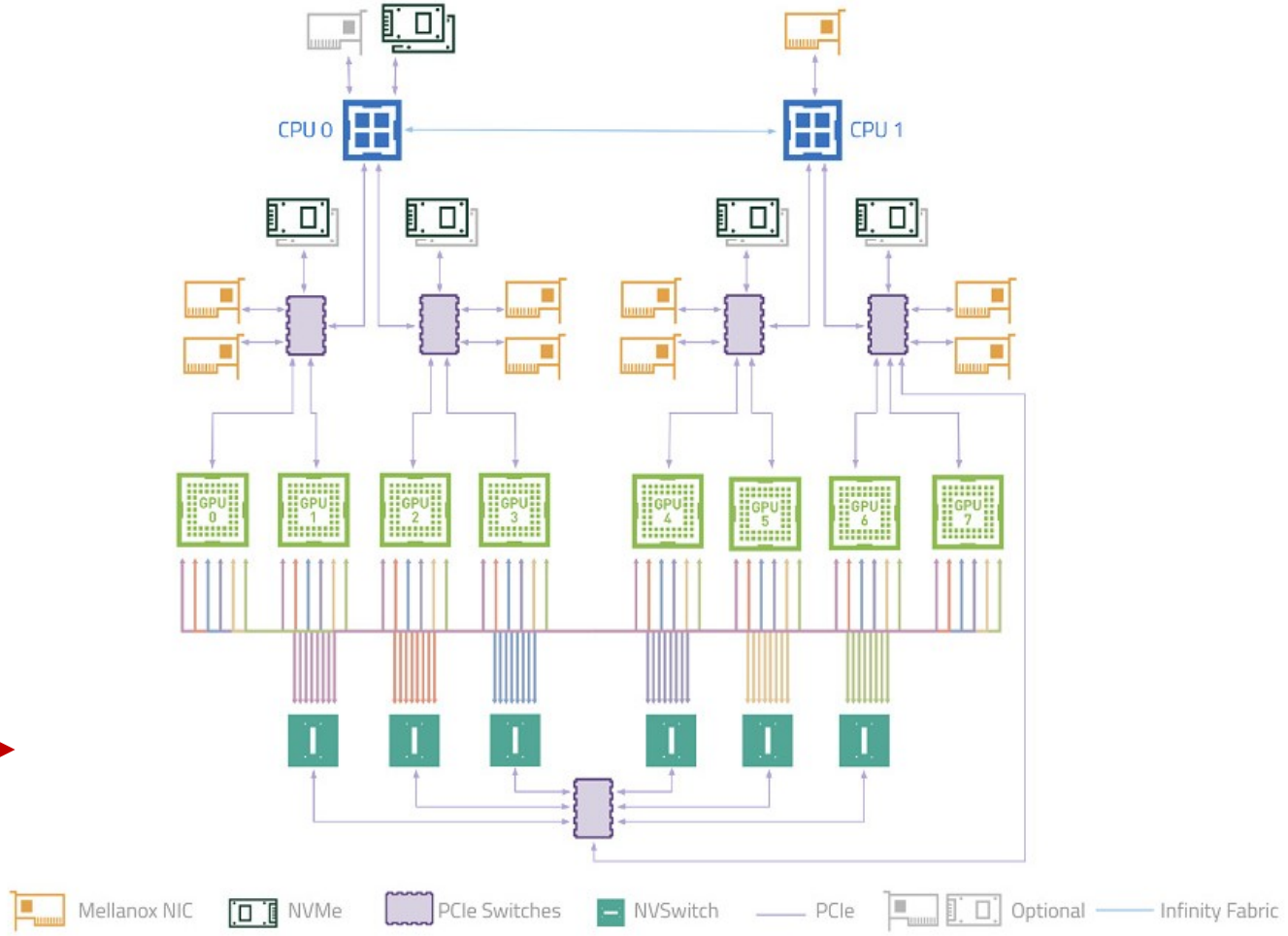


Image source: https://docs.nvidia.com/dgx/pdf/dgxa100-user-guide.pdf

Experimental Setup

ThetaGPU cluster at Argonne Leadership Computing Facility

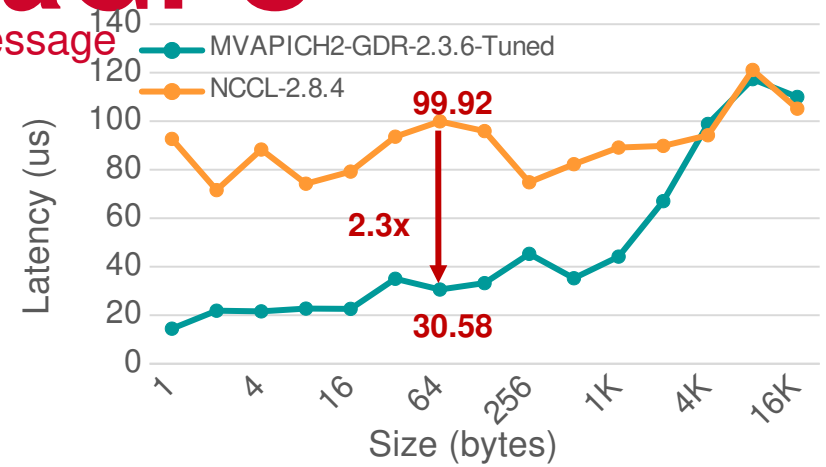
- 24 NVIDIA DGX A100 nodes
  - 8 NVIDIA A100 Tensor Core GPUs
  - 2 AMD Rome CPUs
  - 1 TB memory
- Mellanox ConnectX-6, 20 Mellanox QM9700 HDR200 40-port switches
- OS: Ubuntu 20.04.2
- CUDA Version 11.0.221

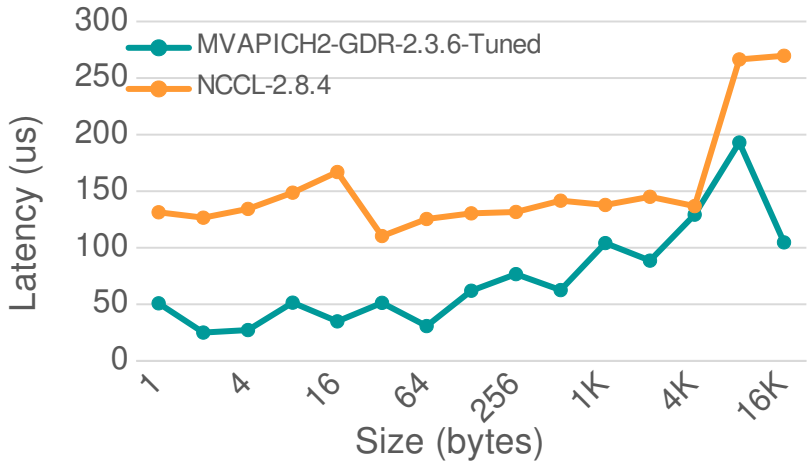
Experimental Setup

- NCCL Version 2.8.4
  - Link: https://developer.nvidia.com/nccl
- MVAPICH2-GDR Version 2.3.6
  - MVAPICH2-GDR + NCCL Hybrid
  - Link: http://mvapich.cse.ohio-state.edu/features/#mv2gdr
- OSU Micro-Benchmarks (OMB) + NCCL Benchmark Support:
  - osu_nccl_[allgather, allreduce, bcast, reduce, reduce_scatter]
  - Available from OSU Micro-Benchmarks (OMB) 5.8
  - Link: https://mvapich.cse.ohio-state.edu/benchmarks/

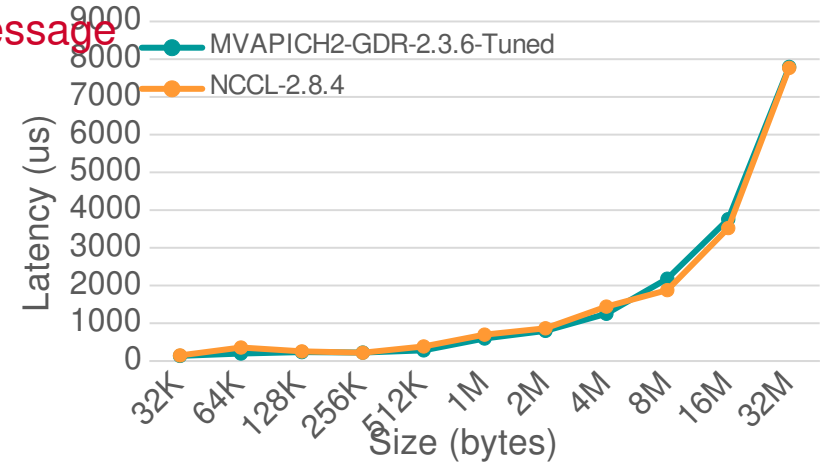
MPI_Allgather Performance on ThetaGPU

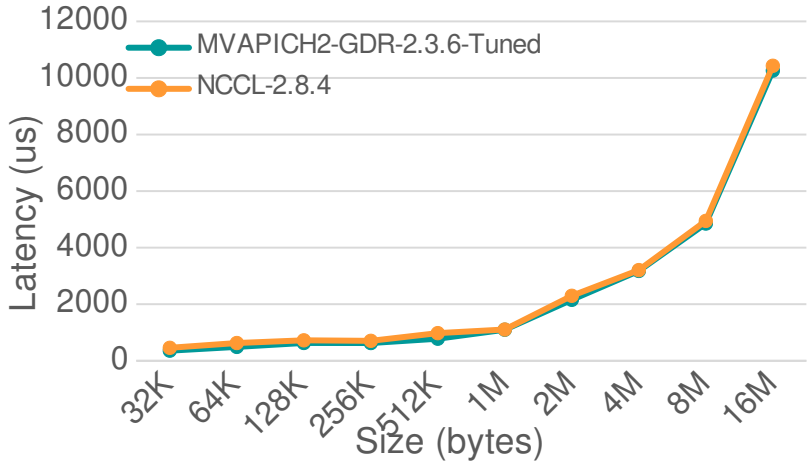
Small Message





Large Message



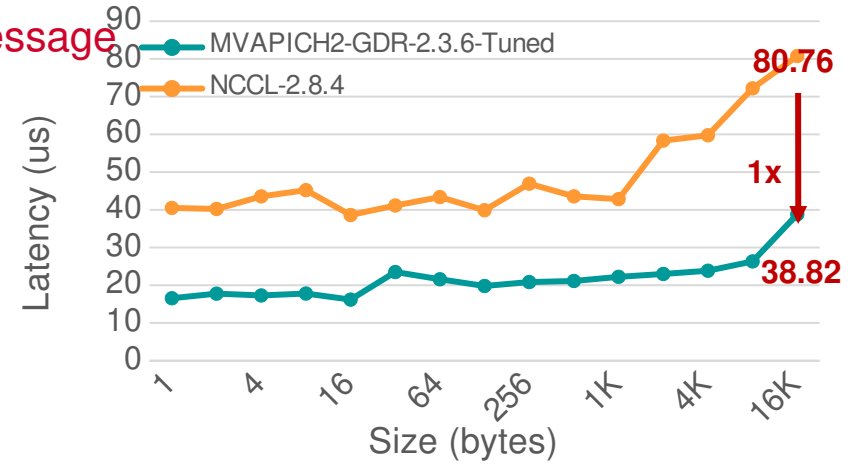


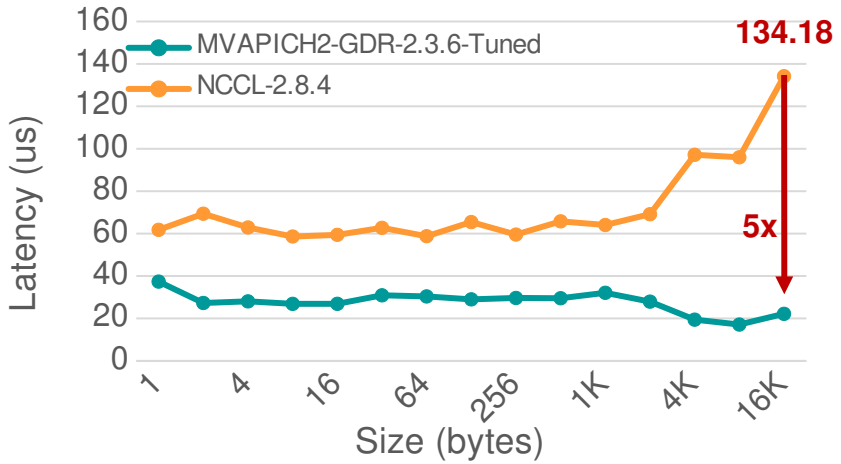
2-node 16-

4-node 32-

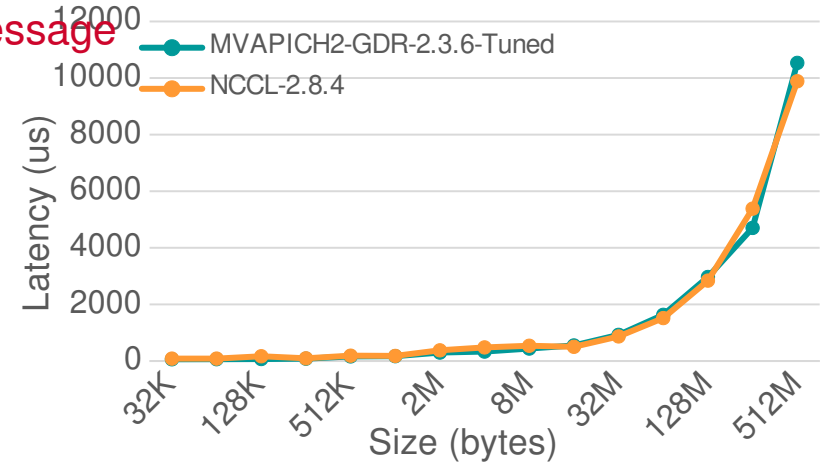
MPI_Bcast Performance on ThetaGPU

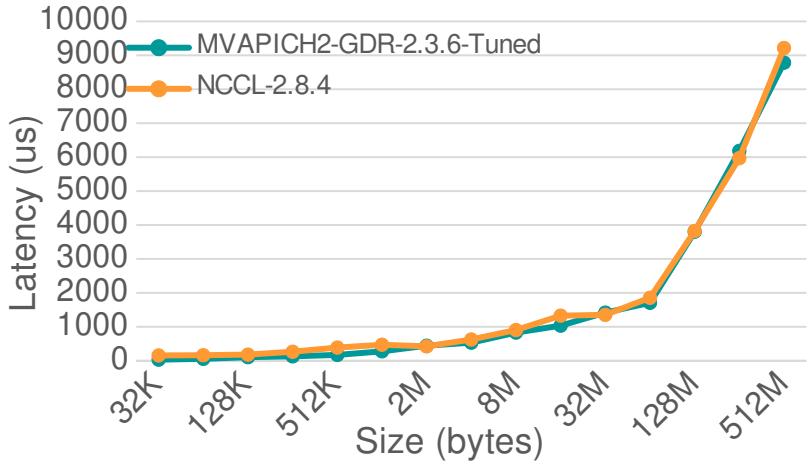
Small Message





Large Message



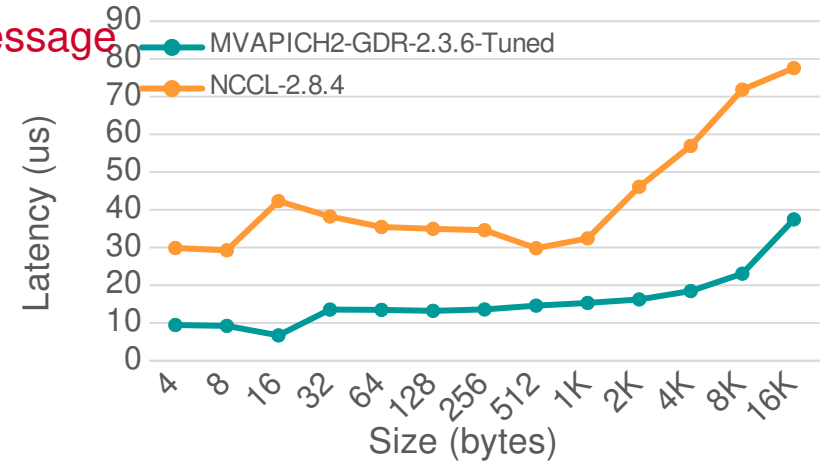


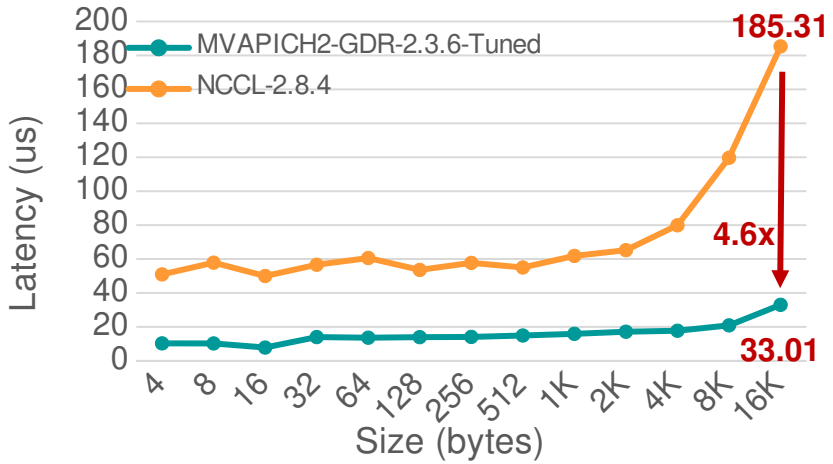
2-node 16-

4-node 32-

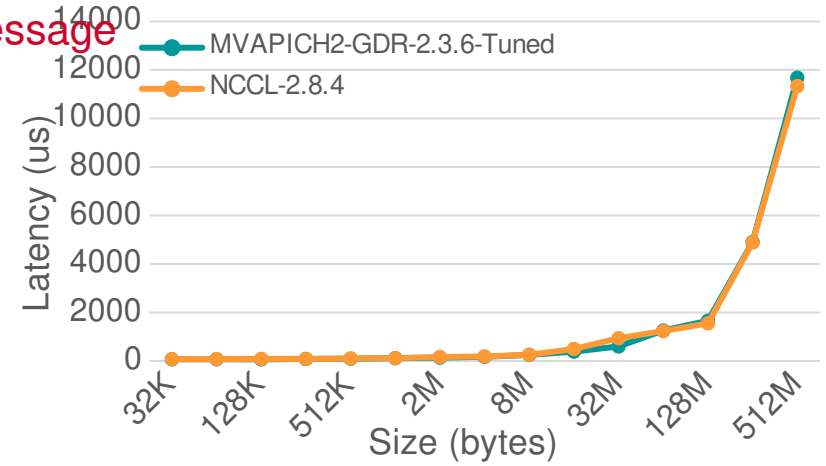
MPI_Reduce Performance on ThetaGPU

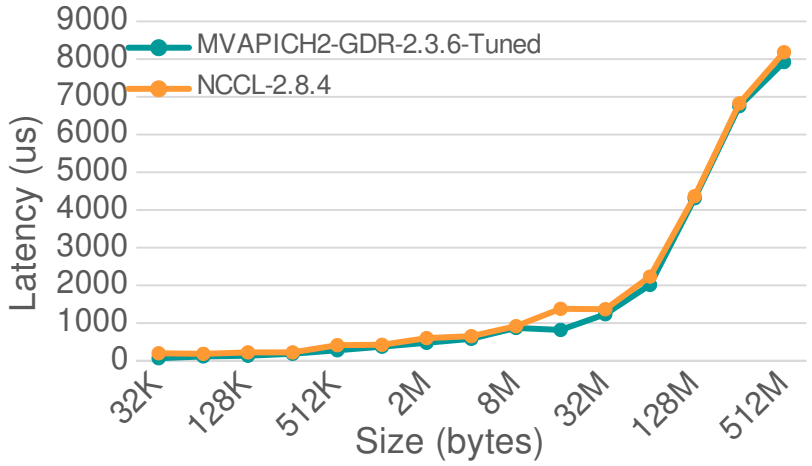
Small Message





Large Message



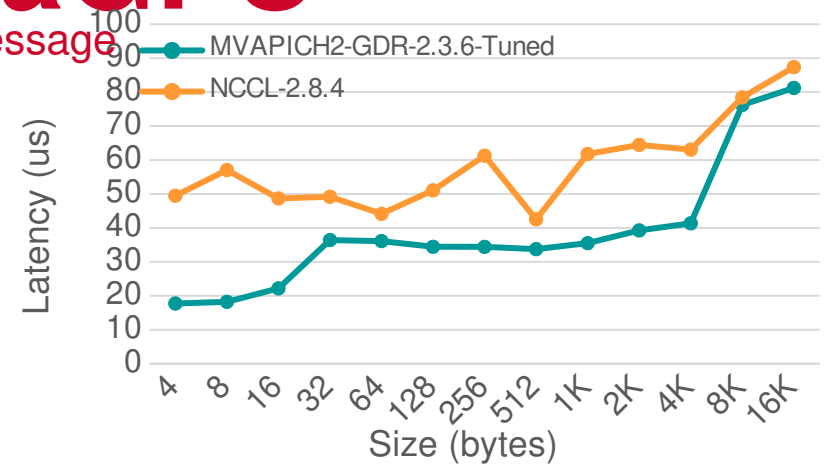


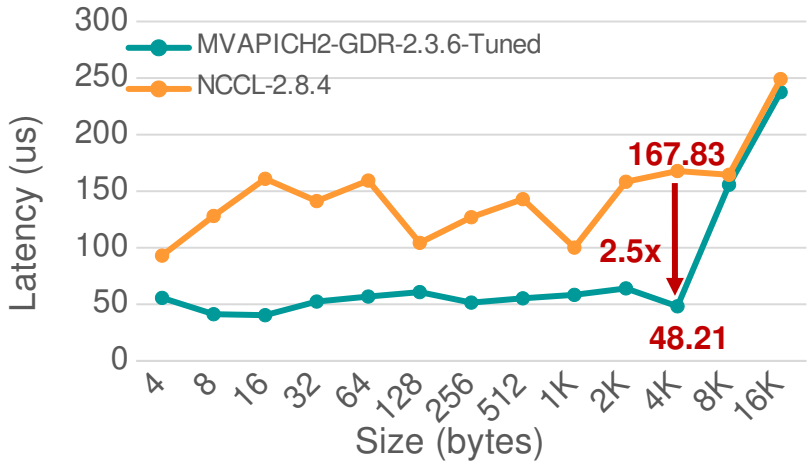
2-node 16-

4-node 32-

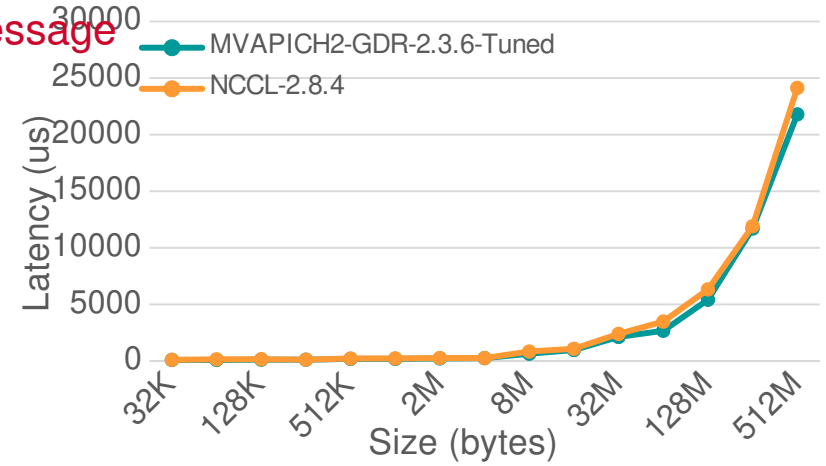
MPI_Allreduce Performance on ThetaGPU

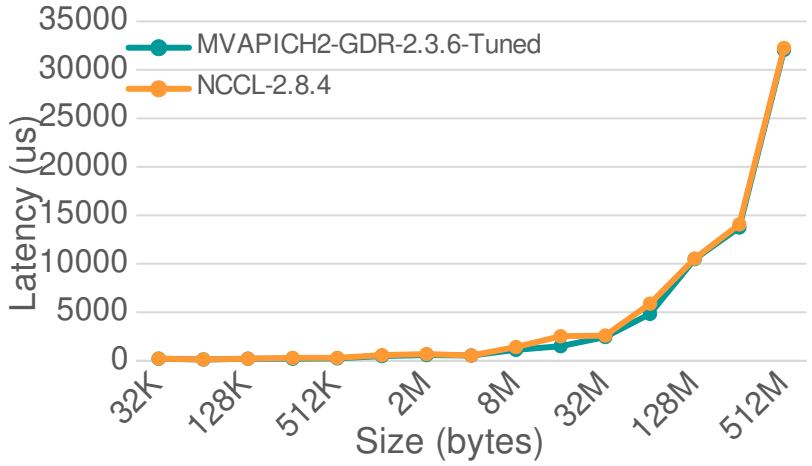
Small Message





Large Message





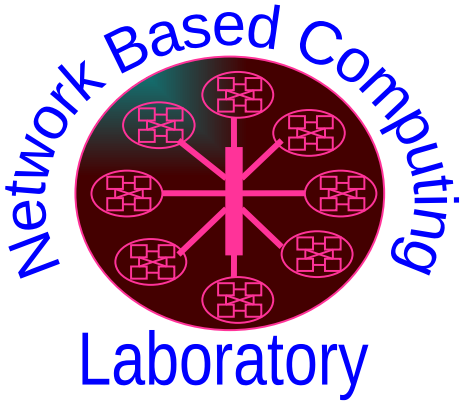
2-node 16-

4-node 32-

Conclusion

- Optimize the performance of GPU based Collective Communication
- Support for NVIDIA DGX A-100 systems
- Enhance collective tuning for allgather, bcast, reduce, and allreduce on ThetaGPU @ ALCF
- For small-message, MVAPICH2-GDR performs well (~5x); for large-message, MVAPICH2-GDR is competitive with NCCL
- Tuning for DGX A100 system is available through releases MVAPICH2-GDR 2.3.6+ and optimizations expected in future releases
- Future work: tuning for other collectives, e.g.: gather, scatter, alltoall

Thank You!



Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

Chen-Chun Chen
chen.10252@osu.edu

The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/