MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiDL
High-Performance Deep Learning

HiBD
High-Performance Big Data

Accelerating MPI All-to-All Communication with Online Compression on Modern GPU Clusters

Presentation at MUG '22

Qinghua Zhou

Network Based Computing Laboratory (NOWLAB)

Dept. of Computer Science and Engineering , The Ohio State University

{zhou.2595}@osu.edu

Follow us on

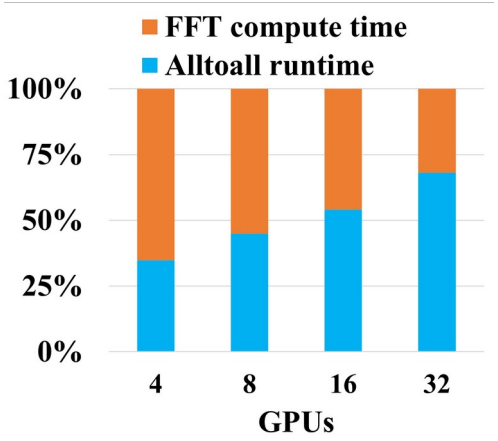
https://twitter.com/mvapich

Outline

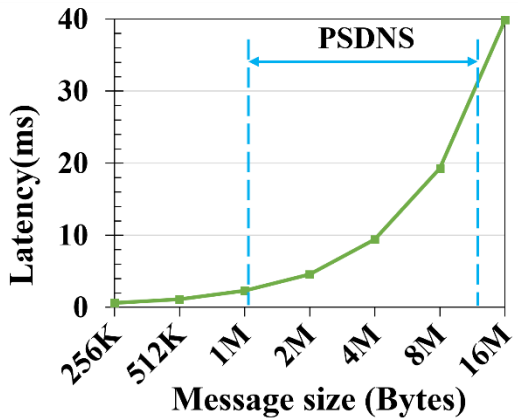
- Motivation
- Problem Statement
- Focus of the Work
- Online Compression Design
  - Overview of Host-Staging based compression design
  - Integration of compression library
- Performance Evaluation
  - Benchmark-level evaluations
  - Application-level evaluations
- Conclusions and Future Work

Motivation

- For HPC and Deep Learning applications on modern GPU clusters
  - With larger problem sizes, applications exchange **orders of magnitude more data** on the network
  - **AlltoAll** is one of the most communication-intensive MPI operations that become the bottleneck of efficiently scaling these applications(e.g, PSDNS) to larger dense GPU systems
  - Existing AlltoAll algorithms for transferring GPU data still suffer from poor performance due to the **saturated bandwidth** of commodity networks



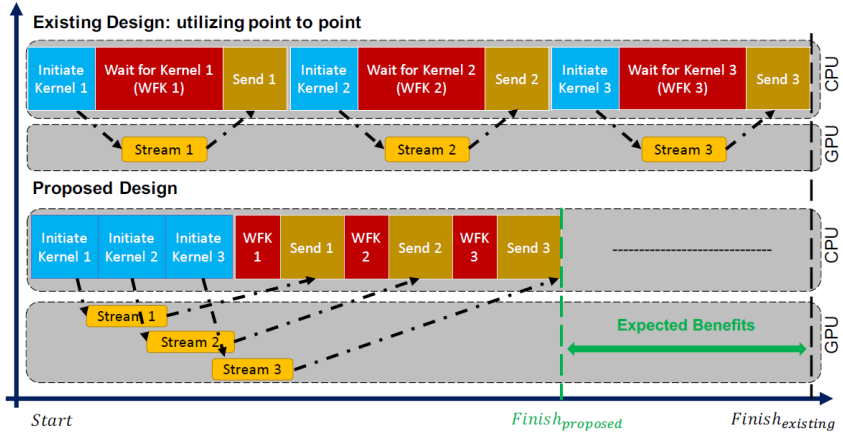
(a) PSDNS Time Breakdown



(b) AlltoAll Latency for 8 GPUs on 2 Longhorn nodes

Problem Statement

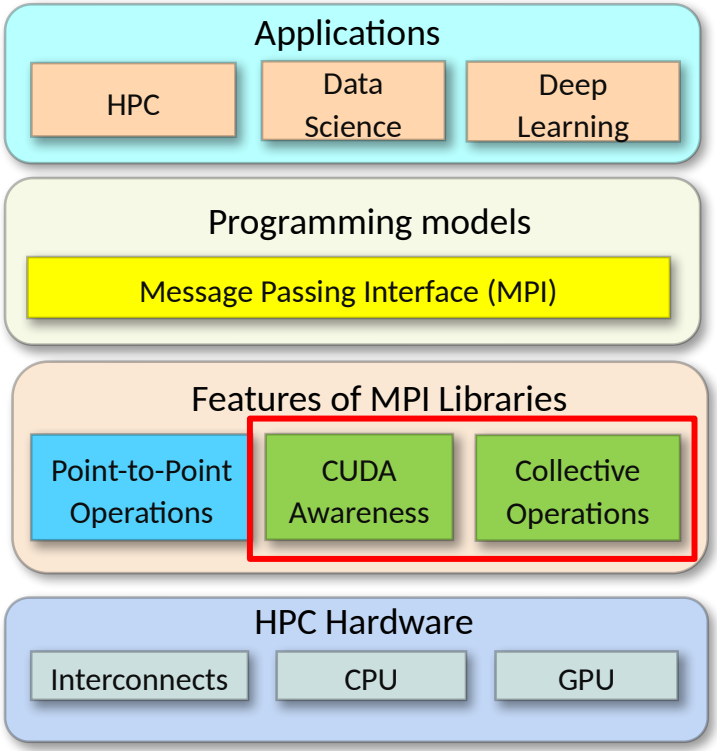
- Existing **Point-to-Point** based compression has limitation of overlapping compression/decompression kernels across send/receive operations.
- How can we overcome the limitation of Point-to-Point based compression schemes and revamp the data transfer pattern to accelerate the applications?
  - Move the point-to-point compression to the **collective-level**
  - Revamp and optimize GPU-based compression for the collective-level online compression



Comparison between point-to-point compression operations versus proposed design.

Focus of the Work

- Optimizing and co-designing the existing GPU based compression algorithms at the **collective** level
- Designing a Host-Staging based **Online compression** schemes for **AlltoAll** in an MPI library
- Accelerating AlltoAll communication performance of transferring large GPU-to-GPU data
- Demonstrating performance benefits for two categories of applications:
  - PSDNS (HPC) [1]
  - DeepSpeed (Deep Learning) [2]



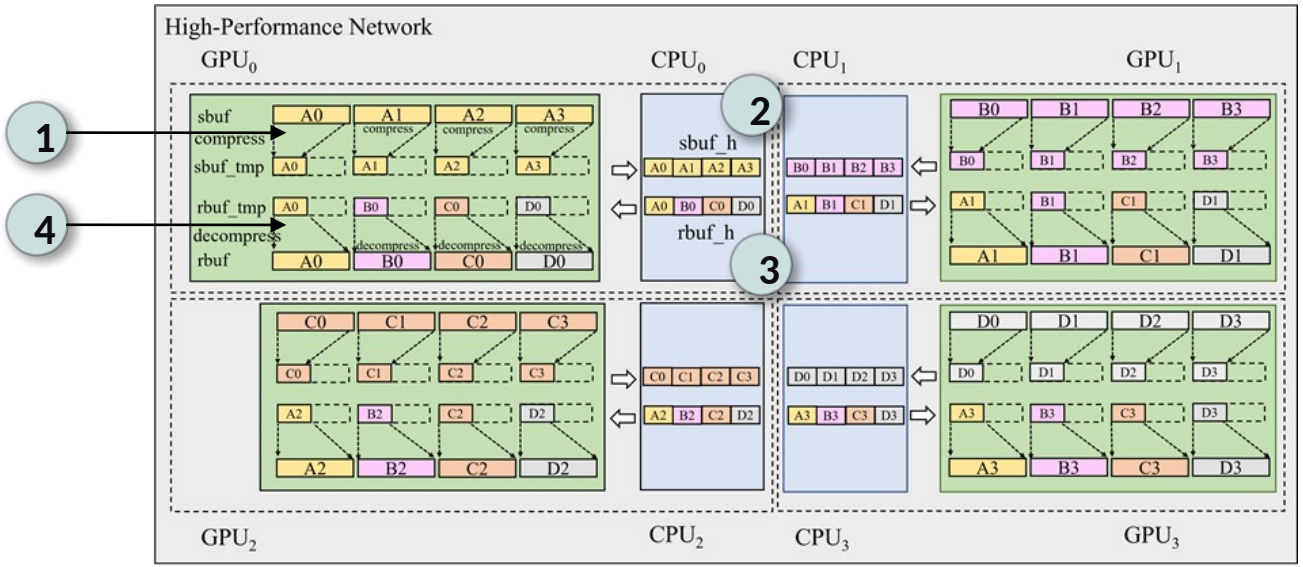
[1] Ravikumar, K., Appelhans, D., Yeung, P.K., Gpu acceleration of extreme scale pseudo-spectral simulations of turbulence using asynchronism. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. SC '19, Association for Computing Machinery, New York, NY, USA (2019).
[2] Rasley, J., Rajbhandari, S., Ruwase, O., He, Y., Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining. p. 3505–3506. KDD '20, Association for Computing Machinery, New York, NY, USA (2020).

Outline

- Motivation
- Problem Statement
- Focus of the Work
- Online Compression Design
  - Overview of Host-Staging based compression design
  - Integration of compression library
- Performance Evaluation
  - Benchmark-level evaluations
  - Application-level evaluations
- Conclusions and Future Work

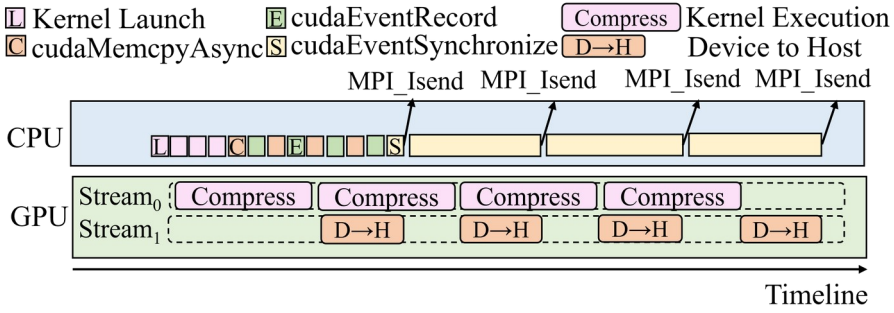
Host-Staging based Online Compression Design

- Data Flow of Host-Staging based Online Compression
  - 1. GPU data is compressed to the temporary device buffer and copied to the host buffer asynchronously
  - 2. MPI_Isend sends out the data in the host buffer to other CPUs
  - 3. MPI_Irecv receives the data to the host buffer from other CPUs
  - 4. Received data is copied to the temporary device buffer asynchronously and decompressed to the target buffer

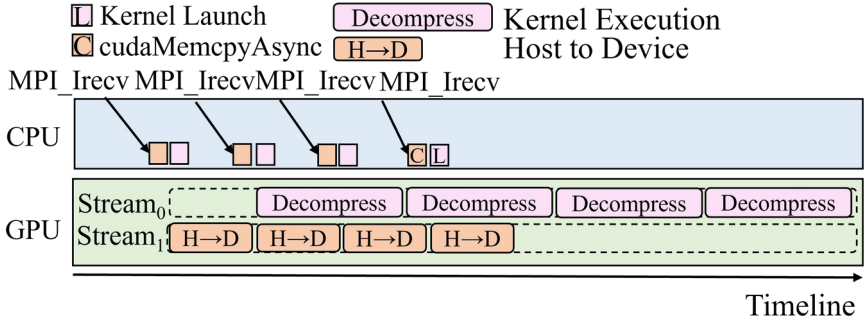


Integration of Compression Library

- Limitation of integrating exiting ZFP compression library
  - Compression/Decompression kernels run on default CUDA stream
  - Long waiting time for MPI_Isend operations due to serial compression kernels
  - Long operation time to restore compressed GPU data due to serial decompressions
- We optimized and co-designed the ZFP compression library at the collective level [3]



(a) Send operations



(b) Receive operations

[3] Q. Zhou, P. Kousha, Q. Anthony, K. Khorassani, A. Shafi, H. Subramoni, and D. K. Panda, Accelerating MPI All-to-All Communication with Online Compression on Modern GPU Clusters. ISC High Performance 2022, May 2022.

Outline

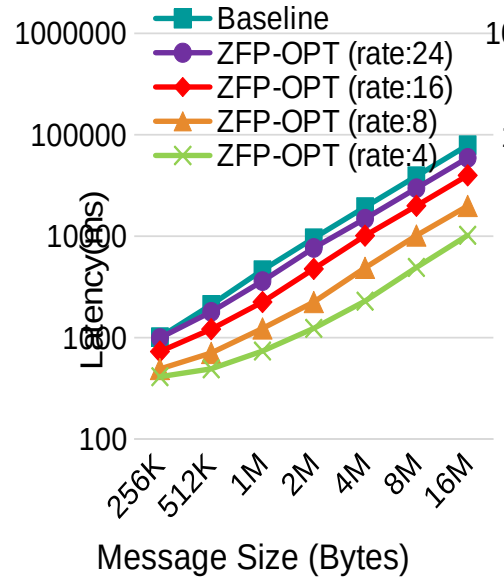
- Motivation
- Problem Statement
- Focus of the Work
- Online Compression Design
  - Overview of Host-Staging based compression design
  - Integration of compression library
- Performance Evaluation
  - Benchmark-level evaluations
  - Application-level evaluations
- Conclusions and Future Work

Experimental Environment

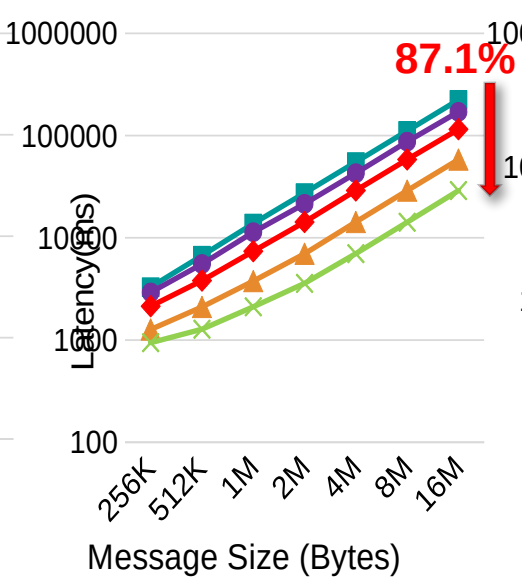
Cluster Specs	Frontera Longhorn	Frontera Liquid	Lassen
CPU Processor	Dual-socket IBM POWER9 AC922 2.3GHz, 20 Cores/socket	Dual-socket Intel Xeon E5-2620 2.10GHz, 8 Cores/socket	Dual-sock IBM POWER9 AC922 3.14GHz, 44 Cores/Socket
System Memory	256 GB	384 GB	256 GB
GPU Processor	4 NVIDIA Tesla V100	4 NVIDIA Quadro RTX 5000	4 NVIDIA Tesla V100
GPU Memory	4 x 16 GB	4 x 16 GB	4 x 16 GB
Interconnects between CPU and GPU	NVLink-2 (one-way 75 GB/s)	PCIe Gen3 x16 and x64 switches (one-way 16 GB/s)	NVLink-2 (one-way 75 GB/s)
Interconnects between GPUs	NVLink-2 (one-way 75 GB/s)	PCIe Gen3 x16 and x64 switches (one-way 16 GB/s)	NVLink-2 (one-way 75 GB/s)
Interconnects between nodes	Mellanox InfiniBand EDR (one-way 12.5 GB/s)	Mellanox InfiniBand FDR (one-way 7 GB/s)	Dual-rail Mellanox InfiniBand EDR (one-way 25 GB/s)
Operating System	RHEL 7.6 (4.14.0-115.10.1.1)	CentOS 7.6.1810 (3.10.0-957.27.2.el7)	RHEL 7.3 (4.14.0-115.10.1.1)
NVIDIA Driver Version	440.33.01	430.40	418.87.00
CUDA Toolkit Version	10.1.168	10.1.243	10.1.243

Benchmark-level evaluation

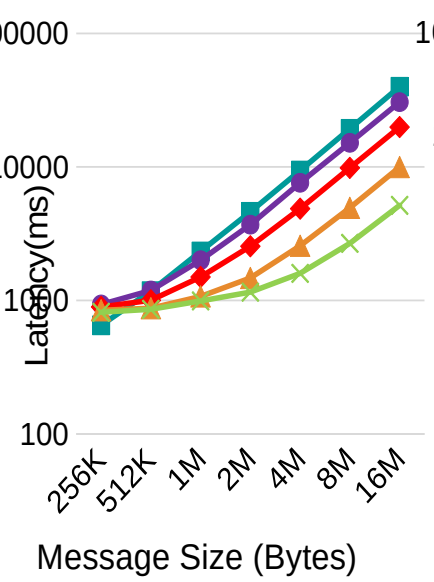
- MPI_AlltoAll Communication Latency with OSU Micro-Benchmark
  - Reduces the latency by up to **87.1%** for 16MB on 2nodes and 4nodes with ZFP-OPT (rate: 4)



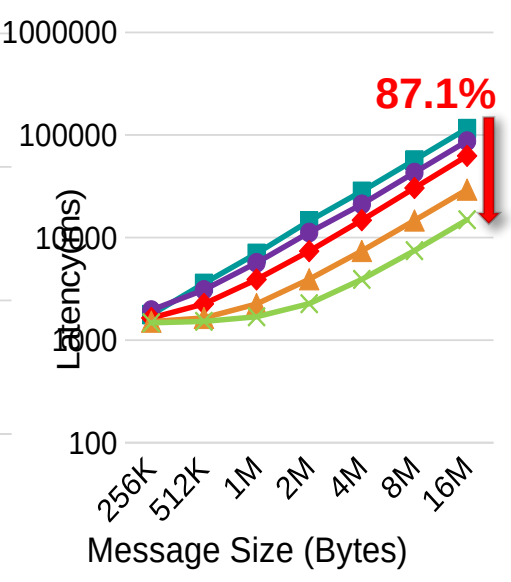
Frontera Liquid: 8GPUs
(2 nodes, 4 ppn)



Frontera Liquid: 16GPUs
(4 nodes, 4 ppn)



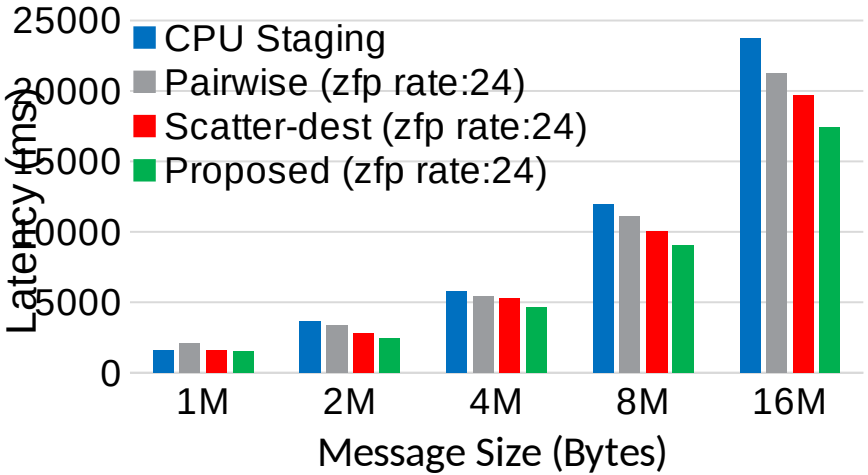
Longhorn: 8GPUs
(2 nodes, 4 ppn)



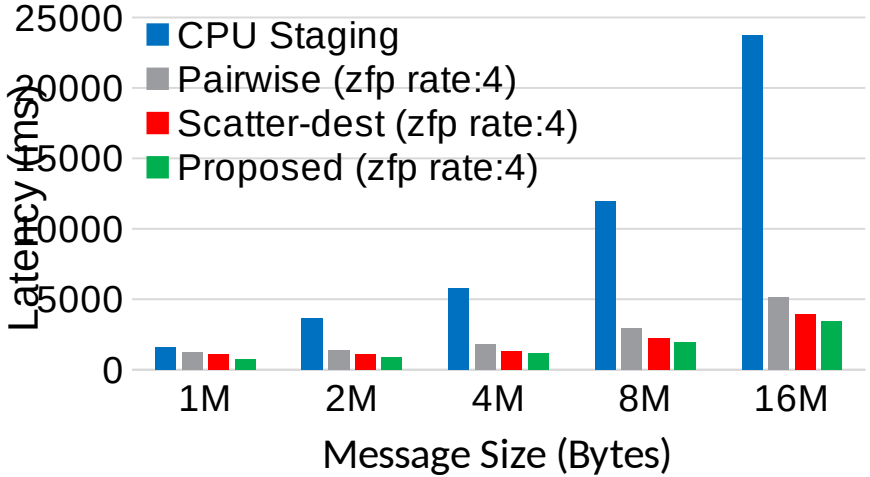
Longhorn: 16GPUs
(4 nodes, 4 ppn)

Comparison with Existing Point-to-Point Compression

- Compare with existing Alltoall algorithms with point-to-point compression in MVAPICH2-GDR-2.3.6
  - The proposed design reduces the Alltoall latency of 16MB by up to **11.2%**, **17.8%**, **26.6%** with ZFP (rate: 24) and **12.4%**, **32.3%**, **85.4%** with ZFP (rate: 4) respectively compared to the Scatter Destination, Pairwise Exchange, and CPU Staging (w/o compression)



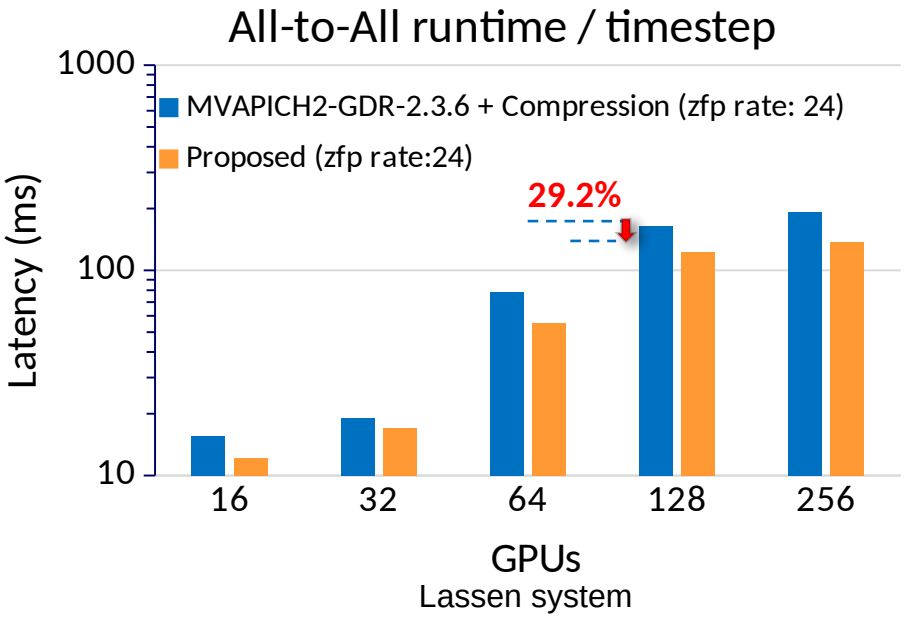
Comparison with zfp (rate:24) for 8 GPUs on 2 Lassen nodes

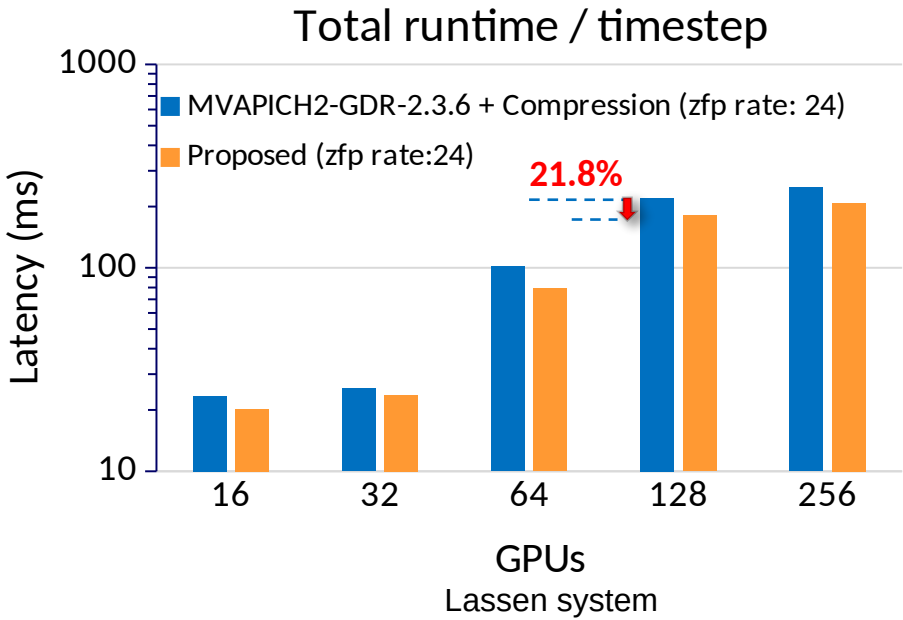


Comparison with zfp (rate:4) for 8 GPUs on 2 Lassen nodes

Application-Level Evaluations (3D-FFT Kernel of PSDNS)

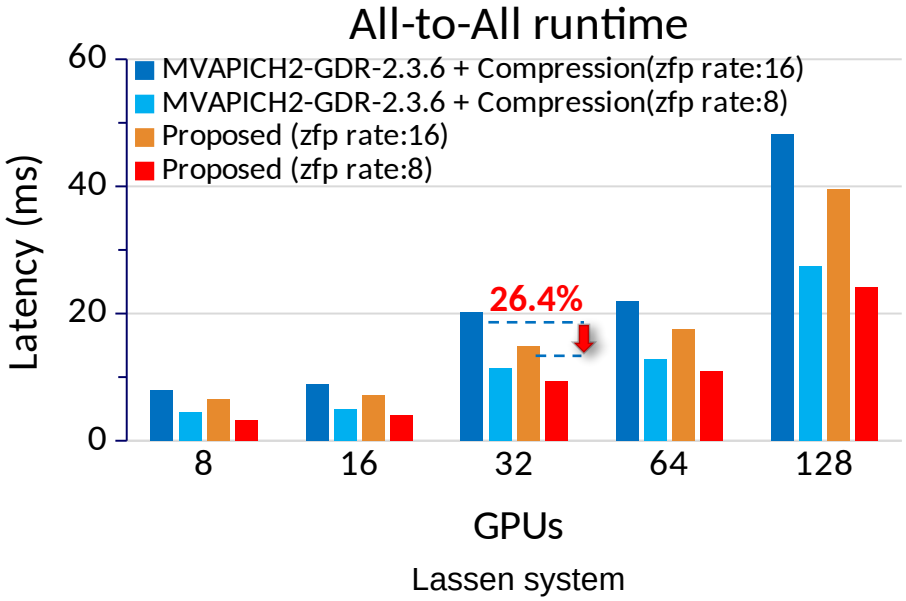
- Improvement compared to MVAPICH2-GDR-2.3.6 with Point-to-Point compression
  - Reduces All-to-All runtime by up to **29.2%** with ZFP(rate: 24) on 64 GPUs
  - Reduces total runtime by up to **21.8%** with ZFP(rate: 24) on 64 GPUs

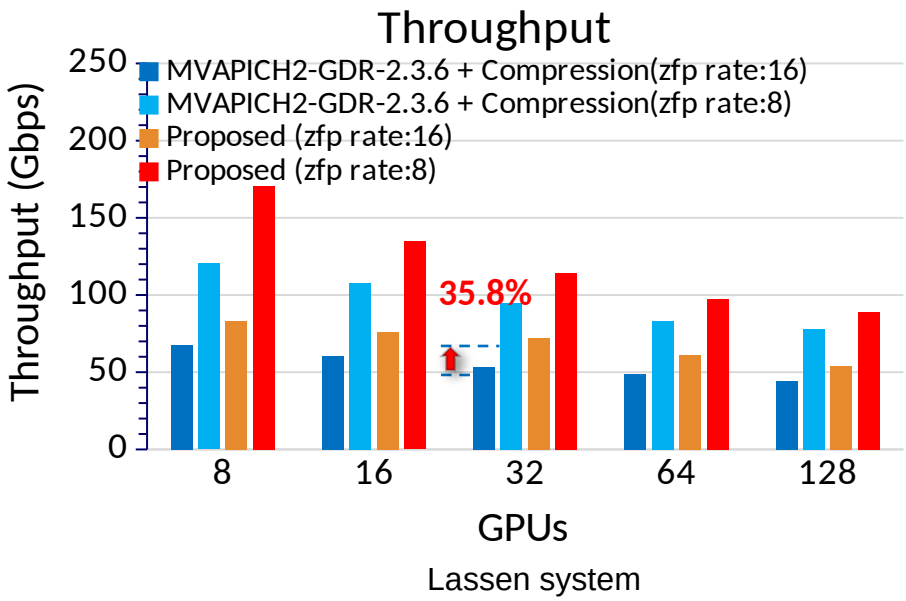




Application-Level Evaluations (DeepSpeed Benchmark)

- Improvement compared to MVAPICH2-GDR-2.3.6 with Point-to-Point compression
  - Reduces All-to-All runtime by up to **26.4%** with ZFP(rate: 16) on 32 GPUs
  - Improves the throughput by up to **35.8%** with ZFP(rate: 16) on 32 GPUs





Conclusions and Future Work

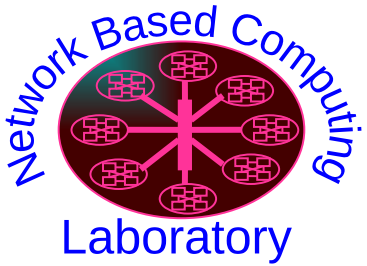
- This paper proposed a **Host-Staging** based **Online Compression** design for optimizing **All-to-All** communication in an MPI library
- We optimized and co-designed the existing GPU based compression algorithm at the **collective level** to tackle the limitation of Point-to-Point compression
- Benchmark-level benefits
  - The proposed design reduced the All-to-All communication latency by up to **87.1%**
  - Reduced the latency by up to **32.3%** compared to Pairwise Exchange algorithm and **12.4%** compared to Scatter-Destination algorithm with point-to-point compression

Conclusions and Future Work (Continued)

- Application-level Benefits
  - PSDNS achieved up to **29.2%** reduced Alltoall runtime and **21.8%** reduced total runtime with ZFP(rate:24) on 64 GPUs
  - DeepSpeed benchmark reduced Alltoall runtime by up to **26.4%** and improve the throughput by up to **35.8%** with ZFP(rate:16) on 32 GPUs
- Future work
  - Study and incorporate more GPU-based compression algorithms (e.g., NVIDIA nvCOMP, etc.)
  - Extend our designs to other common collectives, such as allgather, allreduce, etc.

Thank You!

zhou.2595 @osu.edu



Follow us on

https://twitter.com/mvapich

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/