OPENFABRICS ALLIANCE

2024 OFA Virtual Workshop

SCALING LARGE LANGUAGE MODEL TRAINING USING HYBRID GPU-BASED COMPRESSION IN MVAPICH

Aamir Shafi, Research Scientist

Lang Xu, Ph.D. Student

Network Based Computing Laboratory
The Ohio State University
http://nowlab.cse.ohio-state.edu/

Follow us on X
https://x.com/mvapich

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiDL
High-Performance Deep Learning

Presentation Outline

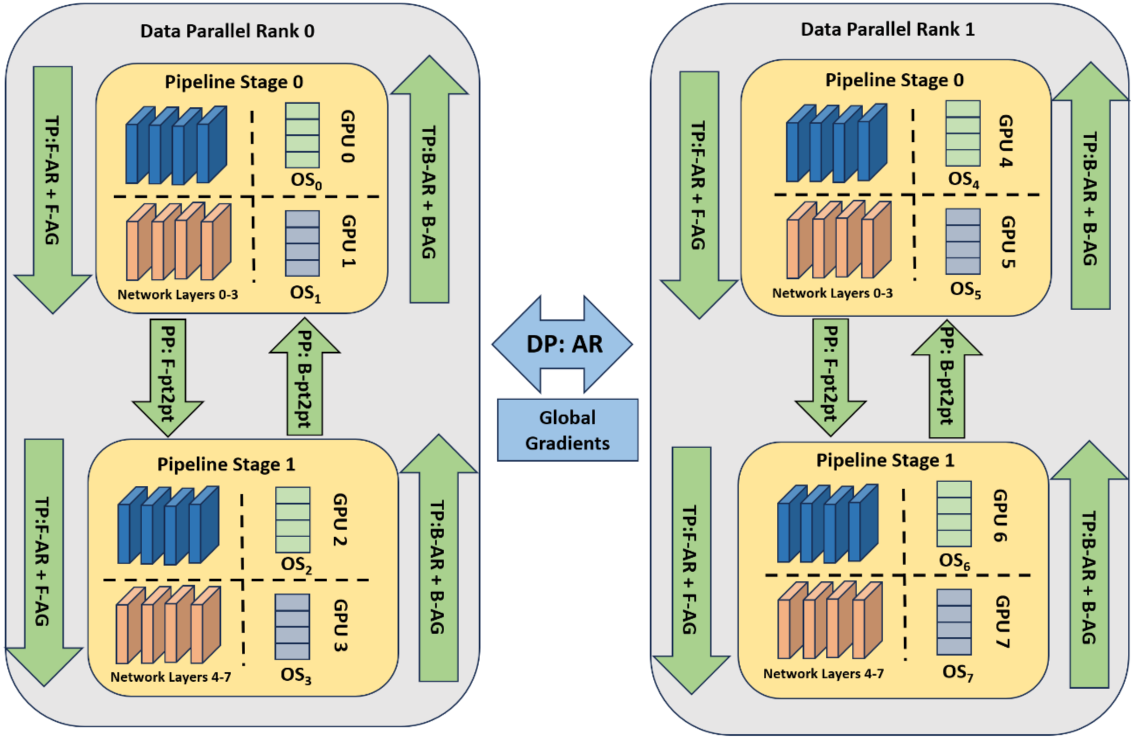
- Introduction & Background
- Motivation & Challenges
- Hybrid Compression Design
- Performance Evaluation
- Conclusion

Training Large Language Model

- Large Language Models (LLaMA2, GPT4, Claude3 …) are powerful in various areas (dialogue systems, knowledge base, …)
- Model capability scales with number of parameters (100 Million [BERT] to 500 Billion [Megatron-Turing NLG])
- Training Billion parameter models requires:
  - Parallelism strategies (scaling up to thousands of GPUs)
  - Memory optimization (fitting models within GPUs)
  - Efficient communication (reducing interconnect bandwidth pressure)

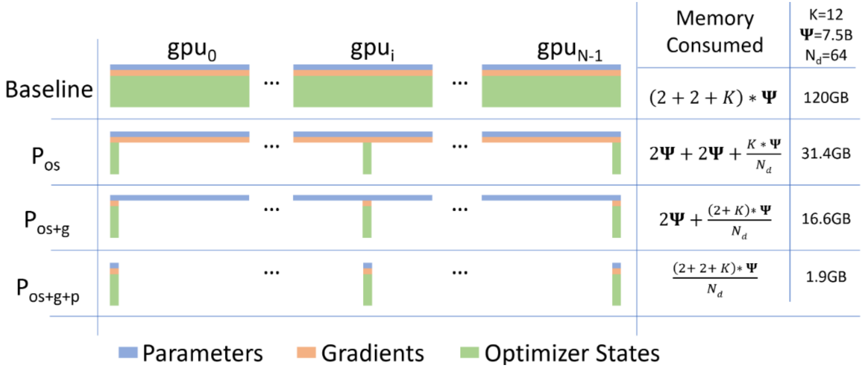
Parallelism Strategies

- Data Parallelism (DP):
  - Maintains full model replica on each DP rank and takes mini-batch as input
  - Data-intensive gradient synchronization using Allreduce
- Pipeline Parallelism (PP):
  - Shards model layers across devices and executes in a pipeline order
  - Point-to-point communication passing activations and gradients
- Tensor Parallelism (TP):
  - Distributes Matrix Multiplication over different devices
  - Frequent Allreduce and Allgather communication ensuring correctness
- 3D Parallelism combines DP+PP+TP (Megatron-LM)

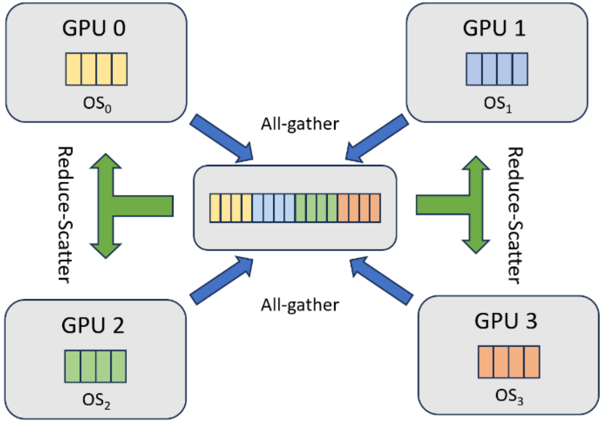


Memory Optimization

- DeepSpeed ZeRO Optimizer:
  - A novel memory optimization technology for large-scale distributed deep learning
  - Enables training models with billions of parameter among GPU
  - Each GPU only updates its portion of data (optimizer states, gradients, model parameters)
    - Reduces memory footprint
  - Requires Allgather and Reduce-Scatter to synchronize between processes
  - ZeRO-1: Partitions optimizer states (momentum & variances) across GPUs
  - ZeRO-2: Further partitions gradients
  - ZeRO-3: Further partitions model parameters

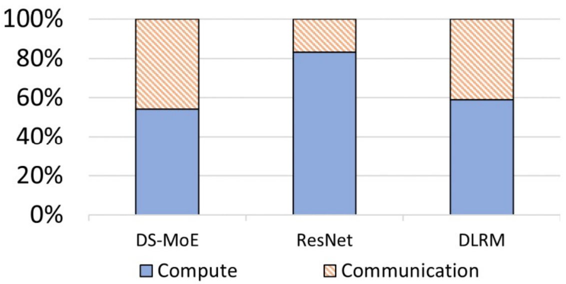


[Deepspeed Zero: https://arxiv.org/abs/1910.02054v3](https://arxiv.org/abs/1910.02054v3)

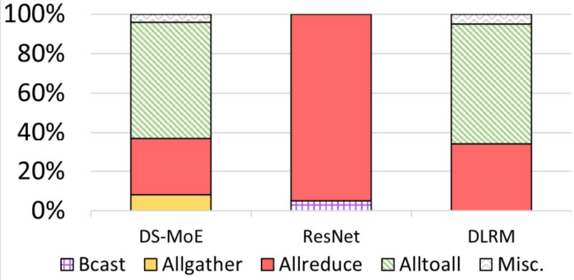


Profiling and Optimizing Communication

- LLM Training requires data-intensive collective communication using 3D parallelism + ZeRO-1
  - Large communication overhead [1]
  - Saturates interconnect bandwidth
- Different sparsity across data structure [2]
  - Gradients are generally sparse (mostly zeros)
  - Activations and optimizer states are dense
- Co-designing MPI with GPU-based Compression has proved to greatly leverage bandwidth and throughput! [3][4]



(a) Proportion of computation to communication for distributed DL training



(b) Breakdown of individual communication operations for distributed DL training

[1] Q. Anthony, et al., "MCR-DL: Mix-and-Match Communication Runtime for Deep Learning," in 2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS), St. Petersburg, FL, USA, 2023

[2] S. Bian et al “Does compression activations help model parallel training?” https://arxiv.org/abs/2301.02654

[3] Q. Zhou *et al.*, "Designing High-Performance MPI Libraries with On-the-fly Compression for Modern GPU Clusters," *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Portland, OR, USA, 2021, pp. 444-453, doi: 10.1109/IPDPS49936.2021.00053.

[4] Q. Zhou *et al.*, "Accelerating Distributed Deep Learning Training with Compression Assisted Allgather and Reduce-Scatter Communication," *2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, St. Petersburg, FL, USA, 2023, pp. 134-144, doi: 10.1109/IPDPS54959.2023.00023.

Motivation

*Using compression-assisted MPI collectives (**Allgather**, **Reduce-scatter** & **point-to-point**) to accelerate large language model training (in a **3D parallelism**+ **ZeRO-1** setting)*

Challenges

- **What are the major communication routines involved in a typical 3D parallelism + ZeRO-1 training scenario?**
  - *Understanding different implementations on these parallelism strategies*
- **How to efficiently utilize the different sparsity inherent in the messages without compromising accuracy?**
  - *Determine message types being transferred in each parallelism degree*
  - *Utilize lossless and lossy compression*
- **How to avoid over-compression in certain parallelism degree?**
  - *Different parallelism stage uses different compression ratio*

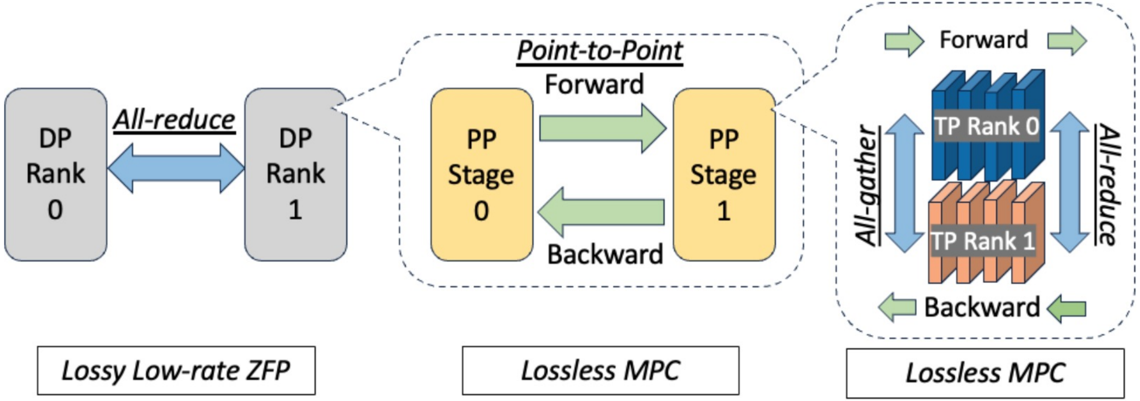
Presentation Outline

- Introduction & Background
- Motivation & Challenges
- Hybrid Compression Design
- Performance Evaluation
- Conclusion

MZHybrid: MPC for MP & ZFP for DP

- Utilize ***lossless MPC*** compression for model parallelism
  - Maintains activation accuracy
  - Applies to inter-layer gradients to avoid over-compression
  - Preserving accuracy
- Utilize ***lossy ZFP*** compression for Data-Parallel data-intensive gradient Allreduce
  - Compress sparse gradients
  - Providing speedups

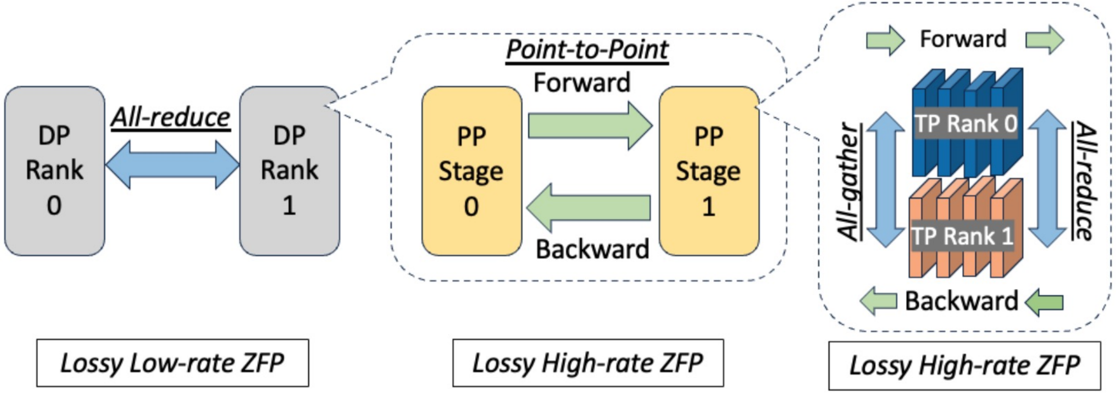
MZHybrid	MPI Collectives	Compression Schemes
DP	All-reduce	ZFP
PP	Point-to-point	MPC
TP	All-reduce	MPC
	All-gather	MPC
ZeRO stage 1	All-gather	MPC
	Reduce-Scatter	MPC



ZHybrid: high-rate ZFP for MP & low-rate ZFP for DP

- Utilize ***high-rate ZFP*** compression for model parallelism
  - Maintains activation accuracy
  - Applies to inter-layer gradients to avoid over-compression
  - Preserving accuracy
- Utilize ***low-rate ZFP*** compression for Data-Parallel data-intensive gradient Allreduce
  - Compress sparse gradients
  - Providing speedups
- More throughput oriented (no lossless components)

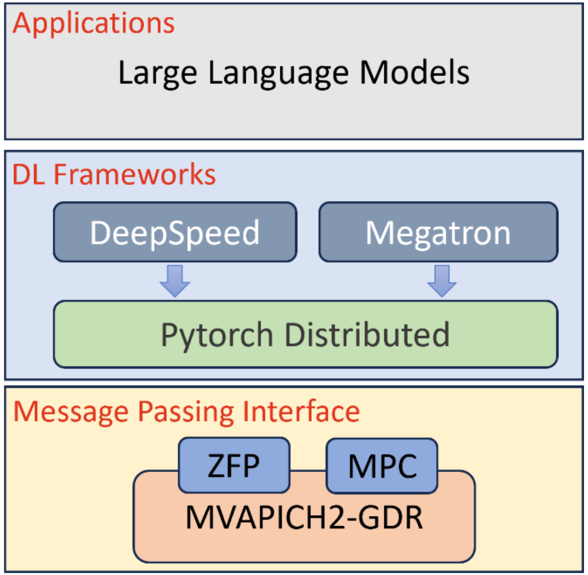
ZHybrid	MPI Collectives	Compression Schemes
DP	All-reduce	low-rate ZFP
PP	Point-to-point	high-rate ZFP
TP	All-reduce	high-rate ZFP
	All-gather	high-rate ZFP
ZeRO stage 1	All-gather	high-rate ZFP
	Reduce-Scatter	high-rate ZFP



Presentation Outline

- Introduction & Background
- Motivation & Challenges
- Hybrid Compression Design
- Performance Evaluation
- Conclusion

Experiment Setup



Model	GPT-NeoX-20B
Dataset	Books3
PP Degree	6
MP Degree	4
Grad Accumulation Step	1
Micro batch size per GPU	4

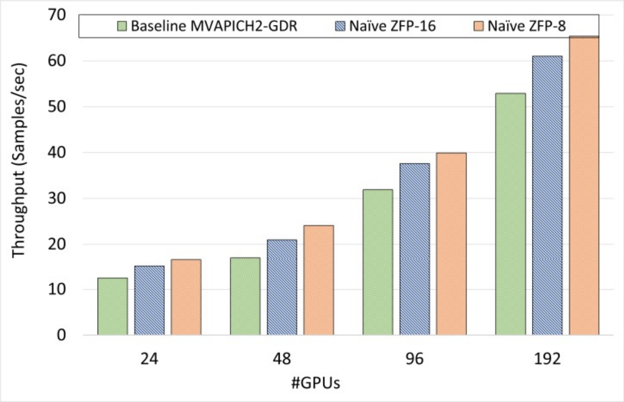
Lassen cluster configuration

CPU	IBM Power9 44 Cores/Node
Memory	256GB
GPU	NVIDIA Tesla V100 (32GB)
Interconnect	InfiniBand EDR 100GB/s

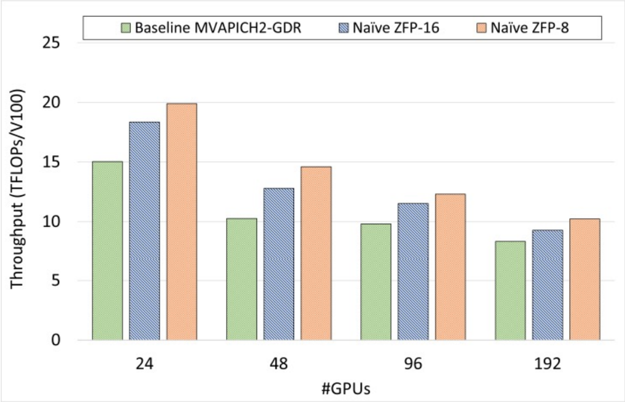
Starting from Naive Compression (ZFP)

- Enforce consistent ZFP compression across all parallelism and ZeRO-1
- ZFP-8 is more aggressive than ZFP-16 in compression (loses more info)
- ZFP-16:
  - **15.4%** increase in throughput (samples/sec)
  - **11.14%** increase in TFLOPS per GPU
- ZFP-8:
  - **23.6%** increase in throughput (samples/sec)
  - **22.5%** increase in TFLOPS per GPU

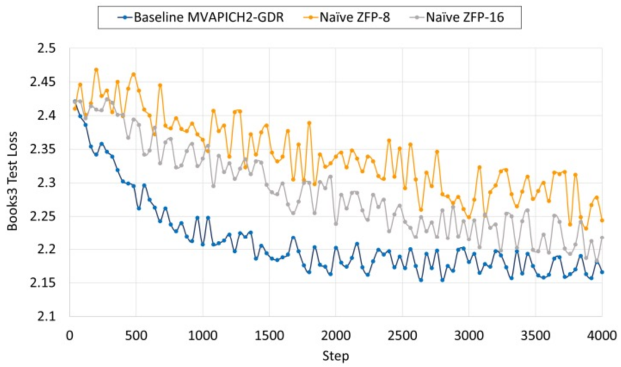
Aggressive lossy compression across all collective communication results in model performance degradation! (higher final test loss)



(a) Naive ZFP: Training samples per second



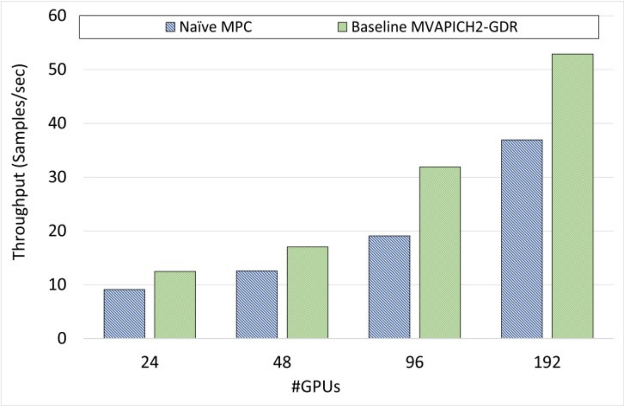
(b) Naïve ZFP: TFLOPS per GPU



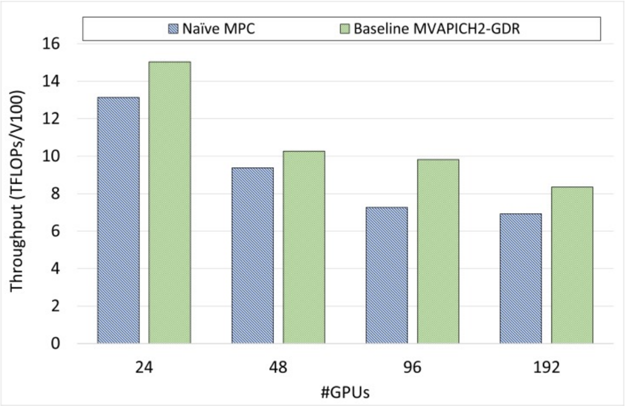
(c) Naïve ZFP: Books3 test loss

Starting from Naive Compression (MPC)

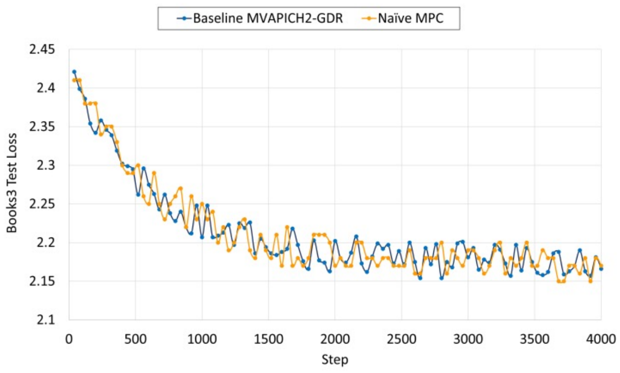
- Enforce lossless MPC for all collectives
- Close to baseline accuracy!
- However, we lose TFLOPS and throughput



(a) Naïve MPC: Training samples per second



(b) Naïve MPC: TFLOPS per GPU



(c) Naïve MPC: Books3 test loss

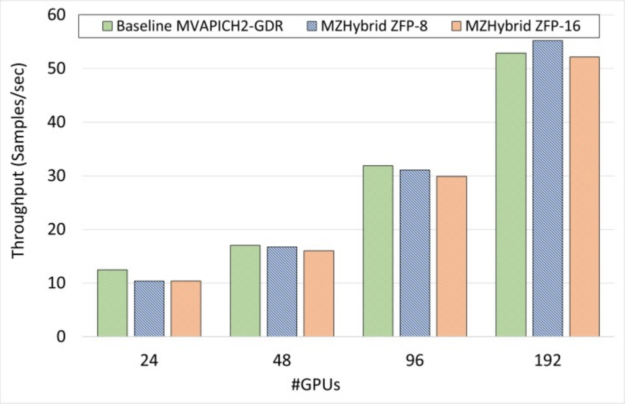
Hybrid Compression

- Naïve ZFP or MPC solution poses different pros and cons
  - ***Lossy ZFP*** provides speedups but degradation in accuracy
  - ***Lossless MPC*** maintains baseline accuracy but degradation in throughput

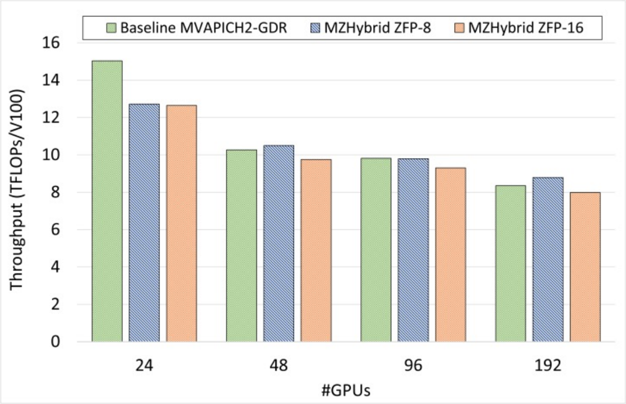
- DP Gradients are sparse, MP activations are dense
  - Possible Hybrid solution for according parallelism degree

MZHybrid

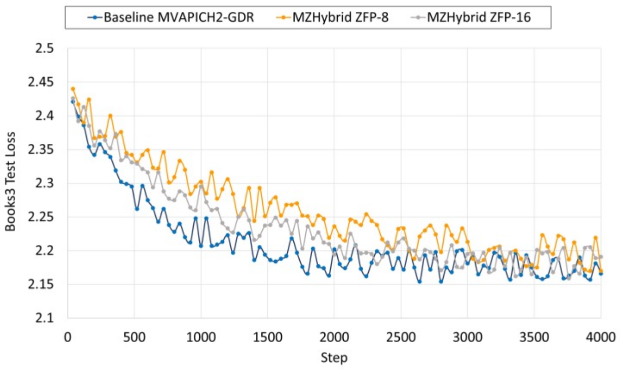
- lossy ZFP compression for Data Parallel gradient Allreduce + lossless MPC compression for Model Parallel (TP + PP) communication
- Good performance speedup (4.4% increase for samples/sec & 5.3% increase for TFLOPS), loss curves greatly improved



(a) MZHybrid: Training samples per second



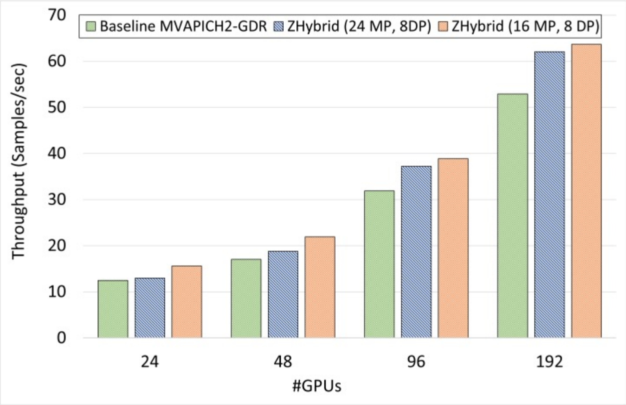
(b) MZHybrid: TFLOPS per GPU



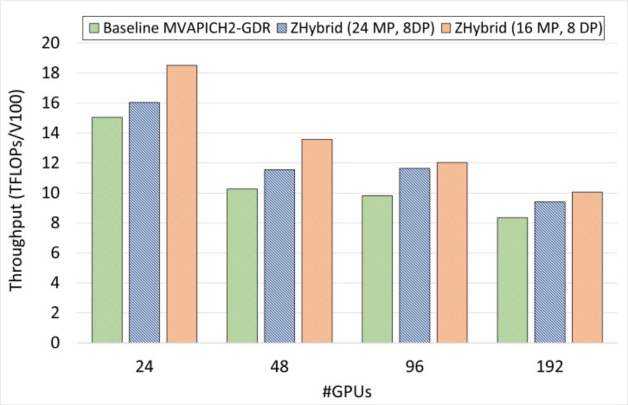
(c) MZHybrid: Books3 test loss

ZHybrid

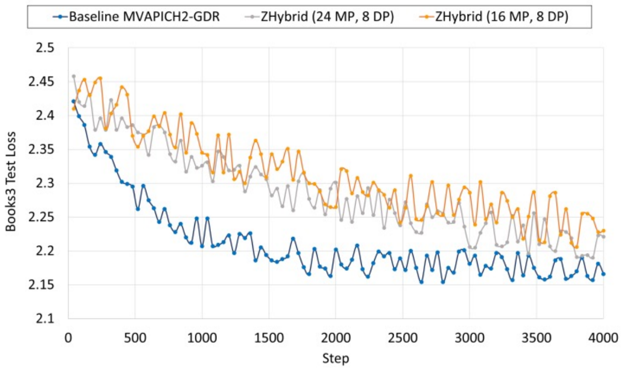
- Low-rate ZFP compression for Data Parallel gradient Allreduce + high-rate ZFP compression for Model Parallel (TP + PP) communication
- Even better performance speedup (17.3% increase for samples/sec & 12.7% increase for TFLOPS), loss curves still acceptable



(a) ZHybrid: Training samples per second



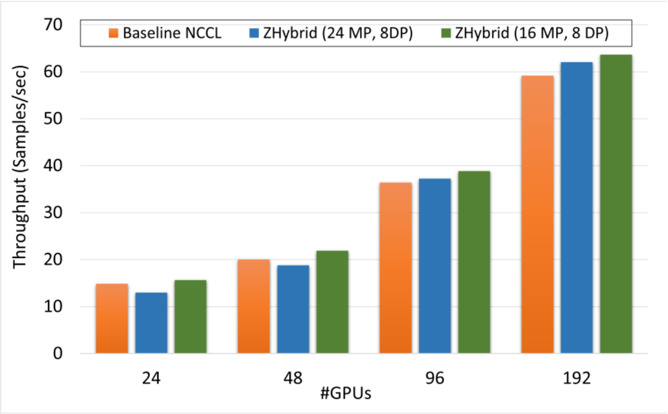
(b) ZHybrid: TFLOPS per GPU

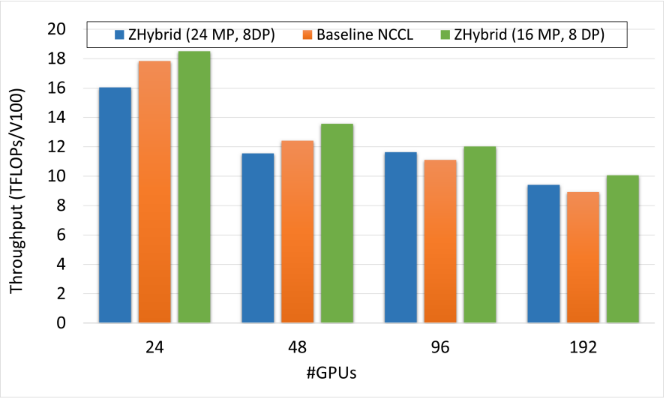


(c) ZHybrid: Books3 test loss

Discussion

- Comparing Zhybrid with NCCL:
  - Up to *7.6%* increase in samples/sec and *12.9%* in TFLOPS per GPU on 192 V100 GPUs
- Compression-assisted MPI collectives capable of reducing message size and mitigate bandwidth pressure as we scale up
- Higher ZFP rates lead to loss closer to baseline than lower ZFP rates
- For specific tradeoffs on accuracy and speedups, the users can select a proper ZFP rate.





Presentation Outline

- Introduction & Background
- Motivation & Challenges
- Hybrid Compression Design
- Performance Evaluation
- Conclusion

Conclusion

- Analyzed different communication routines under 3D parallelism and ZeRO stage 1 for a typical LLM training scenario
- Proposed *MZHybrid* and *ZHybrid*, two hybrid compression schemes that adopts GPU-based Compression MPI collectives on LLM training.
- The two proposed schemes consider data sparsity within communication and utilizes different compression library (MPC & ZFP) for different parallelism to provide training speedups and baseline-level model performance
- MZHybrid provides up to 4.4% increase in samples/sec and 5.3% increase in TFLOPS per GPU while maintaining baseline model accuracy
- ZHybrid provides up to 20.4% increase in samples/sec and 20.6% increase in TFLOPS per GPU

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/