Accelerating Collective Communications with Lossy Compression on GPU

Jiajun Huang
jhuan380@ucr.edu
University of California, Riverside
Riverside, United States of America

Sheng Di(Advisor)
sdi1@anl.gov
Argonne National Laboratory
Lemont, United States of America

Xiaodong Yu(Advisor)
xyu38@stevens.edu
Stevens Institute of Technology
Hoboken, United States of America

Zizhong Chen(Advisor)
chen@cs.ucr.edu
University of California, Riverside
Riverside, United States of America

Franck Cappello(Advisor)
cappello@mcs.anl.gov
Argonne National Laboratory
Lemont, United States of America

Yanfei Guo(Advisor)
yguo@anl.gov
Argonne National Laboratory
Lemont, United States of America

Rajeev Thakur(Advisor)
thakur@anl.gov
Argonne National Laboratory
Lemont, United States of America

Abstract

GPU-aware collective communication has become a major bottleneck for modern computing platforms as GPU computing power rapidly rises. To address this issue, traditional approaches integrate lossy compression directly into GPU-aware collectives, which still suffer from serious issues such as underutilized GPU devices and uncontrolled data distortion. In this paper, we propose *GPU-LCC*, a general framework that designs and optimizes GPU-aware, compression-enabled collectives with well-controlled error propagation. To validate our framework, we evaluate the performance on up to 512 NVIDIA A100 GPUs with real-world applications and datasets. Experimental results demonstrate that our *GPU-LCC*-accelerated collective computation (Allreduce), can outperform NCCL as well as Cray MPI by up to 4.5× and 20.2×, respectively. Furthermore, our accuracy evaluation with an image-stacking application confirms the high reconstructed data quality of our accuracy-aware framework.

Keywords: GPU, Collective Communication, Compression

1 Introduction

For GPU-aware collective communication, numerous researchers are actively working on mitigating network congestion in large-message collectives. In fact, network saturation is often the major bottleneck because of limited network bandwidth. For example, even with advanced networks, such as HPE Slingshot 10, the network bandwidth is only about 100 Gbps. A straightforward solution is designing large-message algorithms that can minimize the transferred data volume instead of latency [1, 7, 9]. Another promising solution is shrinking the message size by error-bounded lossy compression techniques [2, 6, 8, 10], as it can significantly reduce the data volume and maintain the data quality.

Previous lossy-compression-integrated approaches can be divided into two categories. The first is *compression-enabled point-to-point communication* (namely CPRP2P) [11], which directly uses the 1D fixed-rate ZFP [6] to compress the data before it is sent and decompresses the received data after it is received. This method may cause significant overheads and unbounded errors in the collective communications as shown in [3]. The other category is to particularly optimize the *compression-enabled collectives*. Huang et al. designed an optimized general framework for compression-enabled collectives that can realize high performance for all MPI collectives with controlled errors. Nevertheless, this approach suffers from suboptimal performance on modern GPU clusters because of under-utilized GPU devices.

To address the aforementioned limitations, we design a generic framework for GPU-aware compression-accelerated collective communications that can realize both high performance and controlled error propagation.

2 *GPU-LCC* Design and Optimization

In this section, we present our design and optimization strategies as shown in Figure 1. To be specific, we analyze the problems of prior solutions and do a comprehensive performance breakdown to identify potential bottlenecks. Additionally, we also characterize the performance of the lossy compressor and find that the direct application of ring-based algorithms for collective computation with GPU compression may not always yield optimal results. It is hence vital to explore other algorithms that may offer superior performance. After that, we propose the *GPU-LCC* framework to address and overcome the performance issues noted in the previous

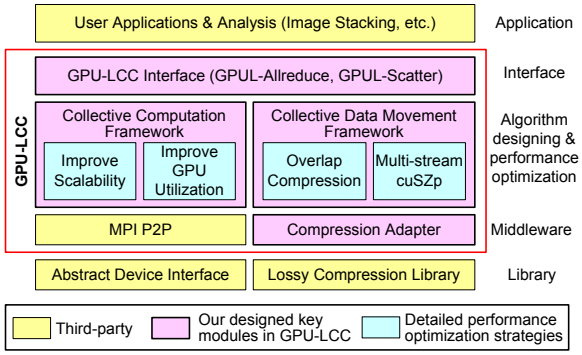

Figure 1. *GPU-LCC* design architecture.

GPU-aware MPI collective framework that incorporates compression, such that a superior performance can be reached. Our contributions are 5-fold: (1) To circumvent the high cost of device-to-host data transfer inherent in traditional CPU-centric designs, we implement a GPU-centric design. (2) To improve collective performance in compression-enabled collectives, we adapt the lossy compression to suit the requirements of collective communications. (3) We explore new metrics regarding GPU compression-enabled collective performance, focusing on minimizing total compression cost and accuracy loss. (4) We propose two algorithm design frameworks for both collective computation and collective data movement to increase device utilization, decrease times of compression/decompression, and maximize the performance. (5) Furthermore, we improve the error-bounded lossy compressor (cuSZp[5]) and develop a multi-stream version to suit the context of the two collective performance optimization frameworks. In our performance optimization frameworks, we try to let as many operations as possible overlap with each other, including kernel launching, compression/decompression operation, and data movement.

3 Experimental Evaluation

We present and discuss the evaluation results as follows.

3.1 Experimental Setup

We perform the evaluation on a GPU supercomputer that involves 512 NVIDIA A100 80G GPUs with 4 GPUs per node, interconnected with a bandwidth of 100 Gbps. Two distinct RTM datasets [4], originating from the real-world 3D SEG/EAGE Overthrust model, are generated under two different simulation settings. Table 1 exhibits the average compression ratio and PSNR that cuSZp can achieve for these datasets, where ABS denotes the absolute error bound.

Table 1. Compression ratio (CPR) and quality (PSNR)

Error Bound = 1E-4	Simulation Setting 1		Simulation Setting 2	
Dimensions	449X449X235		849X849X235	
ABS	CPR	PSNR	CPR	PSNR
1E-3	92.28	53.23	94.41	53.41
1E-4	73.35	65.67	63.94	70.38
1E-5	55.65	78.83	46.74	88.57

3.2 Comparisons of *GPU-LCC* with other collective communication libraries

In this section, we compare the performance of our *GPU-LCC* framework with other state-of-the-art GPU communication libraries, such as the widely-utilized NCCL and Cray MPI, using the prevalent Allreduce operation.

Evaluation with different message sizes. We evaluate the performance of our GPUL-Allreduce algorithm using various data sizes up to 600 MB on a configuration of 64 NVIDIA A100 GPUs across 16 nodes. As observed in Figure 2, our recursive doubling-based GPUL-Allreduce (ReDoub) consistently outperforms across all data sizes, achieving up to a speedup of 18.7× compared to Cray MPI and a 3.4× performance improvement over NCCL. Furthermore, with increasing data sizes, the speedup generally rises, demonstrating high scalability with respect to data size.

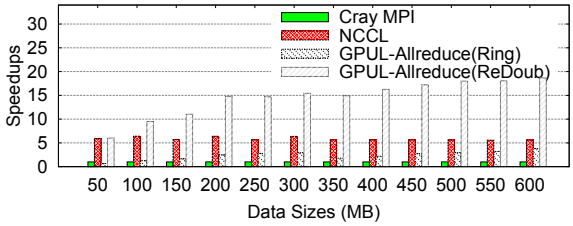

Figure 2. Performance evaluation of our GPUL-Allreduce with Cray MPI and NCCL in different data sizes.

Evaluation with different GPU counts. In this section, we assess the scalability of our GPUL-Allreduce algorithm with the complete RTM dataset of 646 MB data size, utilizing up to 512 NVIDIA A100 GPUs across 128 nodes. As depicted in Figure 3, our recursive doubling-based GPUL-Allreduce (ReDoub) consistently performs the best, achieving up to 20.2× and 4.5× speedups compared to Cray MPI and NCCL respectively, across varying GPU counts. This superior performance stems from the substantial reduction in message size with relatively low compression cost achieved by our *GPU-LCC* framework.

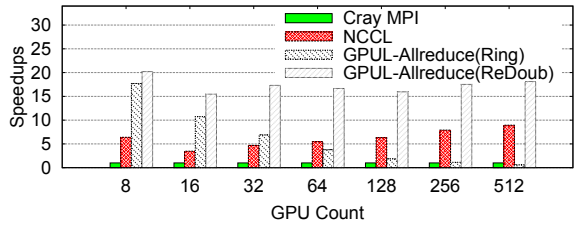

Figure 3. Scalability evaluation of our GPUL-Allreduce with Cray MPI and NCCL in different GPU counts.

4 Conclusion

This paper presents *GPU-LCC*, an innovative framework that optimizes GPU-aware collective communications, which can obtain 20.2× and 4.5× speedups over Cray MPI and NCCL on a testbed of up to 512 NVIDIA A100 GPUs.

Acknowledgment

This research was supported by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations – the Office of Science and the National Nuclear Security Administration, responsible for the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering and early testbed platforms, to support the nation's exascale computing imperative. The material was supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR), under contract DE-AC02-06CH11357, and supported by the National Science Foundation under Grant OAC-2003709 and OAC-2104023. We acknowledge the computing resources on Polaris (operated by Argonne Leadership Computing Facility).

References

[1] George Almási, Philip Heidelberger, Charles J. Archer, Xavier Martorell, C. Chris Erway, José E. Moreira, B. Steinmacher-Burow, and Yili Zheng. 2005. Optimization of MPI Collective Communication on BlueGene/L Systems. In Proceedings of the 19th Annual International Conference on Supercomputing (Cambridge, Massachusetts) (ICS '05). Association for Computing Machinery, New York, NY, USA, 253–262. https://doi.org/10.1145/1088149.1088183

[2] Sheng Di and Franck Cappello. 2016. Fast error-bounded lossy HPC data compression with SZ. In 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE, 730–739.

[3] Jiajun Huang, Sheng Di, Xiaodong Yu, Yujia Zhai, Jinyang Liu, Ken Raffenetti, Hui Zhou, Kai Zhao, Zizhong Chen, Franck Cappello, Yanfei Guo, and Rajeev Thakur. 2023. C-Coll: Introducing Error-bounded Lossy Compression into MPI Collectives. arXiv:2304.03890 [cs.DC]

[4] Suha Kayum et al. 2020. GeoDRIVE – A high performance computing flexible platform for seismic applications. First Break 38, 2 (2020), 97–100.

[5] Argonne National Laboratory. 2023. cuSZp-a lossy error-bounded compression library for compression of floating-point data in NVIDIA GPU. https://github.com/szcompressor/cuSZp.

[6] Peter Lindstrom. 2014. Fixed-Rate Compressed Floating-Point Arrays. IEEE Transactions on Visualization and Computer Graphics 20 (2014), 2674–2683.

[7] Pitch Patarasuk and Xin Yuan. 2009. Bandwidth optimal all-reduce algorithms for clusters of workstations. J. Parallel and Distrib. Comput. 69, 2 (2009), 117–124.

[8] Dingwen Tao, Sheng Di, and Franck Cappello. 2017. Significantly Improving Lossy Compression for Scientific Data Sets Based on Multidimensional Prediction and Error-Controlled Quantization. https://doi.org/10.1109/IPDPS.2017.115

[9] Rajeev Thakur, Rolf Rabenseifner, and William Gropp. 2005. Optimization of collective communication operations in MPICH. The International Journal of High Performance Computing Applications 19, 1 (2005), 49–66.

[10] Kai Zhao, Sheng Di, Xin Liang, Sihuan Li, Dingwen Tao, Zizhong Chen, and Franck Cappello. 2020. Significantly improving lossy compression for HPC datasets with second-order prediction and parameter optimization. In Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing. 89–100.

[11] Q. Zhou, C. Chu, N. S. Kumar, P. Kousha, S. M. Ghazimirsaeed, H. Subramoni, and D. K. Panda. 2021. Designing High-Performance MPI Libraries with On-the-fly Compression for Modern GPU Clusters. In 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 444–453. https://doi.org/10.1109/IPDPS49936.2021.00053