Accelerating Collective Communications with Lossy Compression on GPU

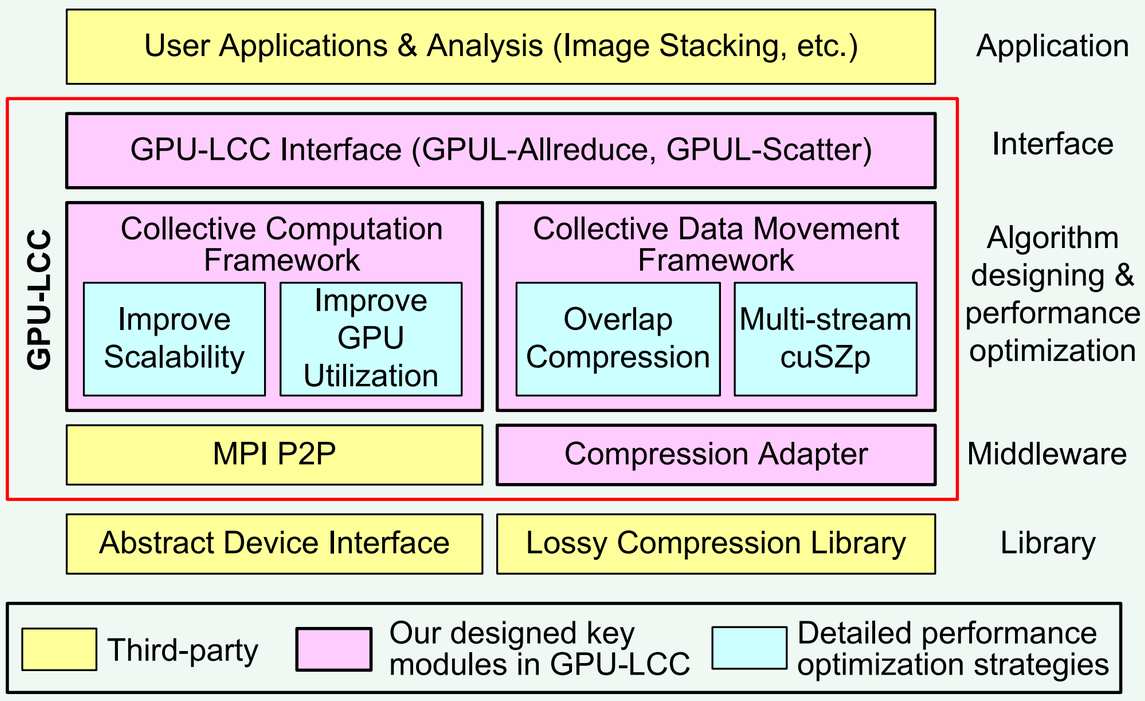
Jiajun Huang[1], Sheng Di(Advisor)[2], Xiaodong Yu(Advisor)[3], Zizhong Chen(Advisor)[1], Franck Cappello(Advisor)[2], Yanfei Guo(Advisor)[2], Rajeev Thakur(Advisor)[2]

1. University of California, Riverside
2. Argonne National Laboratory
3. Stevens Institute of Technology

Overview

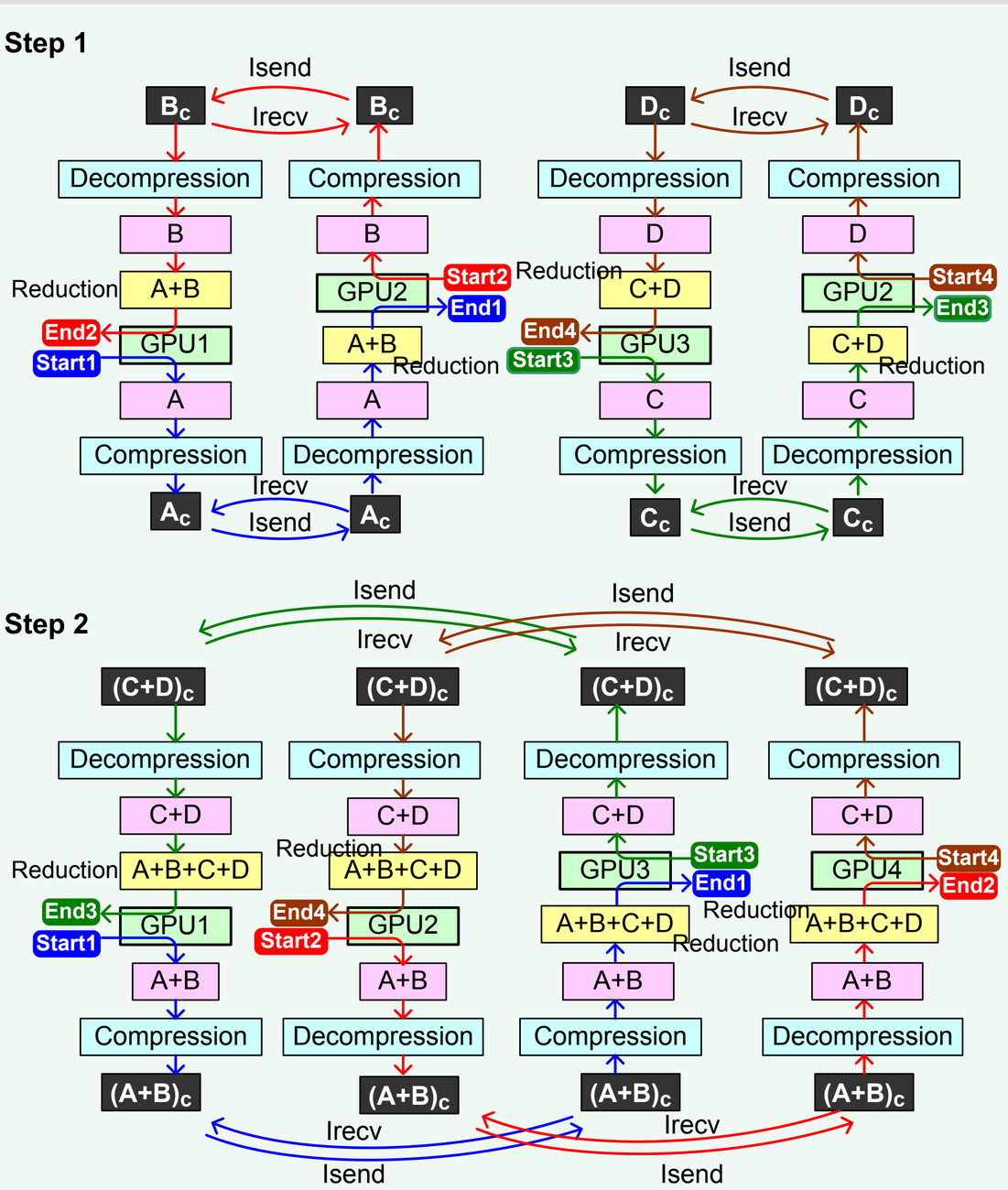
- **GPU-aware collective communication** has become a **major bottleneck** for modern computing platforms as GPU computing power rapidly rises.
- To address this issue, traditional approaches integrate lossy compression **directly** into GPU-aware collectives, which still **suffer** from serious issues 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 **4.5X** and **20.2X**, respectively. Furthermore, our accuracy evaluation with an image-stacking application confirms the high reconstructed data quality of our accuracy-aware framework.

System Design



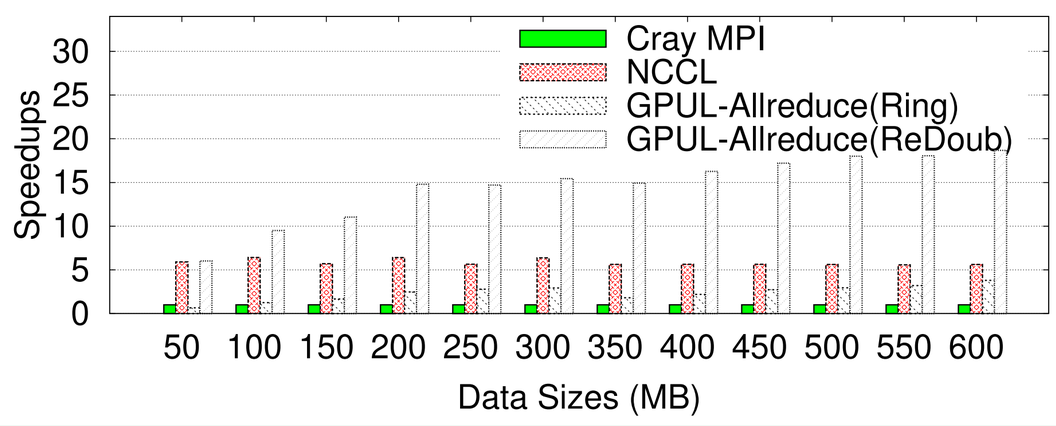
(a) Overall design architecture of our *GPU-LCC* framework.

Detailed Design

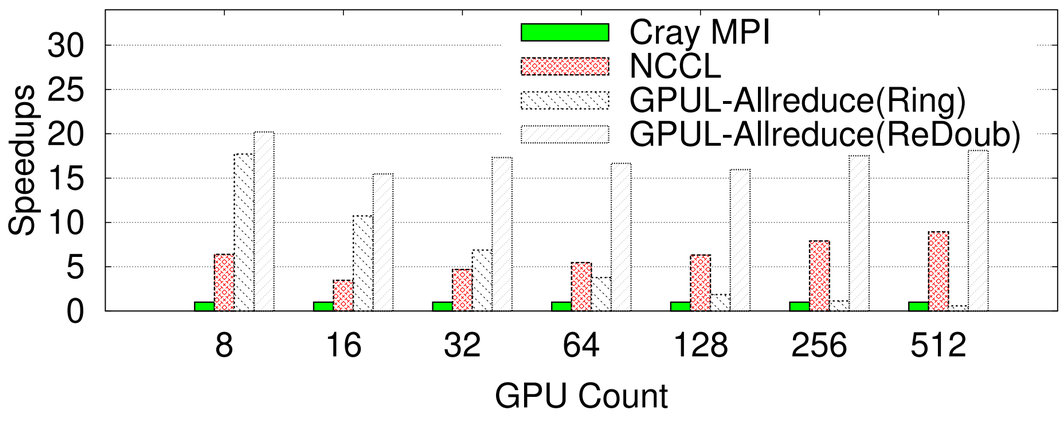


(b) Implementation of our *GPU-LCC* collective computation framework on compression-accelerated GPUL-Allreduce. This example uses four GPUs/Processes.

Experimental Results



(c) Our recursive doubling-based GPUL-Allreduce (ReDoub) consistently outperforms across all data sizes, achieving up to a speedup of **18.7X** compared to **Cray MPI** and a **3.4X** performance improvement over **NCCL**.



(d) Our recursive doubling-based GPUL-Allreduce (ReDoub) consistently performs the best, achieving up to **20.2X** and **4.5X** speedups compared to **Cray MPI** and **NCCL** respectively, across varying GPU counts.

Conclusion

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

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, OAC-2104023. We acknowledge the computing resources on Polaris (operated by Argonne Leadership Computing Facility).