BOYUAN ZHANG

+1(509) 715-8390 ◇ bozhan@iu.edu ◇ Bloomington, IN

EDUCATION

Indiana University (transfer from WSU) Aug. 2021 - Present
PhD in Intelligent System Engineering
University of Southern California Jan. 2019 - Dec. 2020
Master of Science in Electrical Engineering
Shanghai Jiao Tong University Sep. 2014 - Jun. 2018
Bachelor of Engineering in Information Engineering

EXPERIENCE

Pacific Northwest National Laboratory May. 2023 - Present
PhD Intern in High Performance Computing

- Conduct an evaluation of various lossy compression methods applied to microscopy images to assess their efficiency and effectiveness.
- Develop a new workflow on GPUs for AI-based compression techniques designed to simultaneously attain a high compression ratio, maintain quality, and ensure optimal performance.

PROJECTS

High-efficiency State Vector Quantum Circuit Simulation. Nov. 2022 - Jan. 2024

- Leverage the compression to mitigate the memory limitations in modern Quantum Simulation algorithms (e.g., the state vector and the density matrix), portable to simulators such as Qiskit-Aer, SV-Sim, cuQuantum and Cirq.
- Analyze the error propagation in Quantum Simulation when using lossy compression.
- Utilize GPU to accelerate both simulation and compression.

DLRM Communication Overhead Optimizations with Compression. Sep. 2022 - Present

- Investigate the effects of errors introduced by error-bounded lossy compression on the accuracy and training efficiency of the Deep Learning Recommendation Model (DLRM).
- Develop an efficient compression pipeline tailored for the specific data characteristics in the Deep Learning Recommendation Model (DLRM) to achieve a high compression ratio and increased throughput.

Fast and High-Ratio Lossy Compressor on GPUs. Oct. 2022 - Jan. 2023

- Propose a new pipeline to achieve both a high compression ratio and throughput.
- Optimize the dual-quantization in cuSZ to fit the new compression pipeline with significantly higher throughput.
- Design a new GPU bitshuffle kernel to fully leverage the GPU parallelism and exploit the potential spatial redundancy in datasets.
- Propose a new fast lossless encoder to reduce the redundancy brought by bitshuffle, at the same time fully paralleled implemented on GPU.

Optimizing LZSS Lossless Compressor on Modern GPUs. Aug. 2022 - Jan. 2023

- Improve the state-of-the-art lossless LZ compressor (i.e., CULZSS) with significantly higher compression throughput.
- Design a fully GPU-implemented LZ compressor that exploits the parallelism in the LZSS algorithm and uses algorithm-level optimizations (including two-pass prefix sum, kernel fusion, workflow redesign, and multi-byte matching) to accelerate GPU performance.
- Analyze the impact of the data chunk size, sliding window size, and multi-byte length in GPU LZSS compression and propose a lightweight configuration selection strategy.

SKILLS

Programming Language	C++, C, Python
Parallel Programming	CUDA, MPI, OpenMP

PUBLICATIONS

[**HPDC '23**] **Boyuan Zhang**, Jiannan Tian, Sheng Di, Xiaodong Yu, Yunhe Feng, Xin Liang, Dingwen Tao, Franck Cappello. "FZGPU: A Fast and High-Ratio Lossy Compressor for Scientific Computing Applications on GPUs." The 32nd ACM International Symposium on High-Performance Parallel and Distributed Computing, Orlando, FL, June 16–23, 2023. DOI: 10.1145/3588195.3592994.

[**ICS '23**] **Boyuan Zhang**, Jiannan Tian, Sheng Di, Xiaodong Yu, Martin Swany, Dingwen Tao, Franck Cappello. "GPULZ: Optimizing LZSS Lossless Compression for Multi-byte Data on Modern GPUs." The 37th ACM International Conference on Supercomputing, Orlando, FL, USA, June 21–23, 2023. DOI: 10.1145/3577193.3593706.