BOYUAN ZHANG

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

EDUCATION

Indiana University Bloomington (Transferred 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

ByteDance May 2025 - Aug. 2025
Research Scientist Intern (Seed - Machine Learning System)

- Optimizations in LLM inference.

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

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

PROJECTS

Pushing the Limits of GPU-Based Lossy Compression Sep. 2024 - Jan. 2025

- Published in ICS '25. Pushing the Limits of GPU Lossy Compression: A Hierarchical Delta Approach.
- Designed and implemented a high-throughput, error-bounded lossy compression framework on GPUs, introducing a hierarchical data blocking strategy, large-block delta encoding, and dual-level delta decoding to balance compression ratio and speed.

DLRM Communication Optimizations with Compression Sep. 2022 - Feb. 2024

- Published in SC '24. Accelerating Communication in Deep Learning Recommendation Model Training with Dual-Level Adaptive Lossy Compression
- Developed multiple efficient compression schemes tailored to the specific data characteristics in DLRM, including a vector-based GPU LZ algorithm for embedding tables, to achieve a high compression ratio and better performance. The work also presents a dynamic error control scheme to manage error propagation, maintaining high accuracy, and a selection strategy for the best compression ratio among multiple schemes.

High-Performance AI-based Compression on GPUs Oct. 2023 - Jul. 2024

- Published in PPOPP '25 POSTER. High-performance Visual Semantics Compression for AI-Driven Science

- Extended research from the PNNL internship. With a series of GPU optimizations, this work achieves significantly higher image quality at similar compression ratios compared to non-AI-based compressors and achieves even higher end-to-end performance.
- Utilized kernel fusion, warp-level optimization, shared memory, Prefix-Sum via Decoupled Lookback, GPU direct storage, and GPU pipeline techniques to achieve high performance.

Memory Efficient State Vector Quantum Circuit Simulation Mar. 2023 - Jan. 2024

- Published in ICS '25. BMQSim: Overcoming Memory Constraints in Quantum Circuit Simulation with a High-Fidelity Compression Framework
- Utilized a unique characteristic of state vector simulation to divide the simulation process into separate jobs by partitioning the circuit. Leveraged compression to mitigate memory limitations in modern state vector quantum simulation algorithms, portable to simulators such as Qiskit-Aer, SV-Sim, cuQuantum, and Cirq. Achieved similar performance to state-of-the-art simulators with significantly less memory consumption.
- Employed OpenMP to manage multiple GPUs and GPU streams, implementing an efficient pipeline. Also, proposed the first GPU-based point-wise relative error control scheme to limit the error produced by compression.

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

- Published in HPDC '23. FZ-GPU: A Fast and High-Ratio Lossy Compressor for Scientific Computing Applications on GPUs.
- Proposed a new pipeline to achieve both a high compression ratio and throughput by optimizing the dual-quantization in cuSZ to fit the new compression pipeline with significantly higher throughput, a new GPU bitshuffle process to fully leverage GPU parallelism, and a new fast lossless encoder to reduce redundancy introduced by bitshuffle.
- Utilized shared memory and warp-level vote function to implement an efficient bit-level memory-intensive operation, carefully managing shared memory to avoid bank conflicts, with kernel fusion for multiple processes.

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

- Published in ICS '23. GPULZ: Optimizing LZSS Lossless Compression for Multi-byte Data on Modern GPUs
- Improved the state-of-the-art lossless LZ compressor (i.e., CULZSS) with significantly higher compression throughput. Designed a fully GPU-implemented LZ compressor that exploits the parallelism in the LZSS algorithm. Unlike state-of-the-art implementations, this work explores the advantages of utilizing multi-byte symbols in the sliding window lookup of the LZSS algorithm to achieve both higher compression ratio and performance.

PUBLICATIONS

[ICS'25] (**Best Paper Runner-Up**) **Boyuan Zhang**, Bo Fang, Fanjiang Ye, Luanzheng Guo, Fengguang Song, Nathan Tallent, and Dingwen Tao. "BMQSim: Overcoming Memory Constraints in Quantum Circuit Simulation with a High-Fidelity Compression Framework." In Proceedings of the 39th ACM International Conference on Supercomputing, pp. 689-704. 2025.

[**ICS'25**] (**Best Paper Candidate**) **Boyuan Zhang**, Yafan Huang, Sheng Di, Fengguang Song, Guanpeng Li, and Franck Cappello. "Pushing the Limits of GPU Lossy Compression: A Hierarchical Delta Approach." In Proceedings of the 39th ACM International Conference on Supercomputing, pp. 654-669. 2025.

[**PPoPP'25 POSTER**] **Boyuan Zhang**, Luanzheng Guo, Jiannan Tian, Jinyang Liu, Daoce Wang, Fanjiang Ye, Chengming Zhang, Jan Strube, Nathan R. Tallent, and Dingwen Tao. "High-performance Visual Semantics Compression for AI-Driven Science." In Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming, pp. 557-559. 2025.

[**SC'24**] Hao Feng*, **Boyuan Zhang***, Fanjiang Ye, Min Si, Ching-Hsiang Chu, Jiannan Tian, Chunxing Yin, Summer Deng, Yuchen Hao, Pavan Balaji, Tong Geng, Dingwen Tao. "Accelerating Communication in Deep Learning Recommendation Model Training with Dual-Level Adaptive Lossy Compression." The International Conference for High-Performance Computing, Networking, Storage, and Analysis, Atlanta, Georgia, USA, November 17-22, 2024. (*Equal contribution.)

[**HPDC'23**] **Boyuan Zhang**, Jiannan Tian, Sheng Di, Xiaodong Yu, Yunhe Feng, Xin Liang, Dingwen Tao, and Franck Cappello. "Fz-gpu: A fast and high-ratio lossy compressor for scientific computing applications on gpus." In Proceedings of the 32nd International Symposium on High-Performance Parallel and Distributed Computing, pp. 129-142. 2023.

[**ICS'23**] **Boyuan Zhang**, Jiannan Tian, Sheng Di, Xiaodong Yu, Martin Swany, Dingwen Tao, and Franck Cappello. "Gpulz: Optimizing lzss lossless compression for multi-byte data on modern gpus." In Proceedings of the 37th International Conference on Supercomputing, pp. 348-359. 2023.