Boyuan Zhang

Ph.D. Candidate

Indiana University
+1 (509) 715-8390
bozhan@iu.edu
boyuanzhang62.github.io

EDUCATION

Indiana University Bloomington Aug. 2022 – Present

Ph.D. in Intelligent Systems Engineering

- Advisor: Prof. Fengguang Song (since May 2024), succeeding Prof. Dingwen Tao (Aug. 2021 – May 2024) following Prof. Tao's transfer to the Chinese Academy of Sciences (CAS).

Washington State University Aug. 2021 – May 2022

Ph.D. in Computer Science

- Advisor: Prof. Dingwen Tao; Transferred to IU in 2022 together with Prof. Tao's research group.

University of Southern California Jan. 2019 – Dec. 2020

M.S. in Electrical Engineering

Shanghai Jiao Tong University Sep. 2014 – Jun. 2018

B.Eng. in Information Engineering

RESEARCH INTERESTS

- **High-Performance Computing (HPC):** Design and optimization of scalable systems for scientific and engineering workloads, with a focus on efficiency and parallel performance.
- **Data Compression:** Development of GPU-based lossy and lossless compression algorithms for scientific data and machine learning workloads, emphasizing fidelity, throughput, and scalability.
- **Quantum Computing:** Exploration of high-performance methods for large-scale quantum circuit simulation and quantum system modeling, addressing memory and scalability challenges.
- **Machine Learning Systems:** System-level optimizations for training and inference, including model compression, memory footprint reduction, and GPU-accelerated data pipelines.

ACADEMIC EXPERIENCE

Indiana University *Aug. 2022 – Present*

Research Assistant, advised by Dr. Dingwen Tao and Dr. Fengguang Song

- Exploring high-performance GPU kernel designs for both memory-intensive and computation-intensive workloads in scientific computing.
- Investigating hybrid classical–quantum solutions for accelerating large-scale HPC problems.
- Designing GPU-based lossy and lossless compression frameworks for scientific data and AI workloads to improve throughput, compression ratio, and fidelity.

Argonne National Laboratory *Aug. 2022 – Present*

Research Assistant, advised by Dr. Sheng Di and Dr. Franck Cappello

- Developing extreme high-throughput, error-bounded, GPU-only lossy compression techniques for large-scale scientific workloads.
- Analyzing data patterns in scientific applications such as climate modeling and cosmology simulations to guide adaptive compression strategies.

Pacific Northwest National Laboratory *May 2023 – Present*

Research Assistant, advised by Dr. Nathan R. Tallent

- Developed a compression-integrated quantum state-vector simulation workflow to reduce the memory overhead of quantum simulations.
- Exploring AI-driven lossy compression methods for microscopy image datasets to enhance the preservation of fine-grained data textures.

Meta *Aug. 2022 – Aug. 2024*

Research Assistant (Meta Research Award), advised by Dr. Min Si

- Collaborated with Meta researchers on efficient GPU-based compression strategies to reduce communication overhead in large-scale distributed recommendation model training.

WORK EXPERIENCE

ByteDance May 2025 – Aug. 2025

Research Scientist Intern (Seed — Machine Learning System), in Xin Liu's team

- Developed Vayne, a unified GPU kernel for quantization and compression of large language model (LLM) KV caches, enabling high-throughput inference with reduced memory footprint.

Pacific Northwest National Laboratory May 2023 – Aug. 2023

Ph.D. Intern in High-Performance Computing, advised by Dr. Nathan R. Tallent

- Evaluated lossy compression for microscopy images to assess efficiency and effectiveness. Built a GPU workflow for AI-based compression targeting high ratio, quality, and throughput.

PUBLICATIONS

[IPDPS'26] **Boyuan Zhang**, Ding Zhou, Yafan Huang, Shihui Song, Hao Feng, Jinda Jia, Chengming Zhang, and Zhi Zhang. *Near-Zero Cost KV Cache Compression for Large Language Model Inference.* In *Proceedings of the 39th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2026.

[IPDPS'26] **Boyuan Zhang**, Luanzheng Guo, Jiannan Tian, Jinyang Liu, Daoce Wang, Chengming Zhang, Bo Fang, Fengguang Song, Jan Strube, Nathan R. Tallent, and Dingwen Tao. *Accelerating AI Compression through Lightweight Lossless Encoding and Pipelined Workflows.* In *Proceedings of the 39th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2026.

[ICS'25] (Best Paper Runner-Up) **Boyuan Zhang**, Bo Fang, Fanjiang Ye, Luanzheng Guo, Fengguang Song, Nathan R. 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 (ICS)*, 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 (ICS)*, pp. 654–669, 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, and Dingwen Tao. *Accelerating Communication in Deep Learning Recommendation Model Training with Dual-Level Adaptive Lossy Compression.* In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2024. (*Co-first authors.)

[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 (HPDC)*, 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 ACM International Conference on Supercomputing (ICS)*, pp. 348–359, 2023.

[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 Symposium on Principles and Practice of Parallel Programming (PPoPP)*, pp. 557–559, 2025.

[SC'25] Daoce Wang, Pascal Grosset, Jesus Pulido, Jiannan Tian, Tushar M. Athawale, Jinda Jia, Baixi Sun, **Boyuan Zhang**, et al. *STZ: A High Quality and High Speed Streaming Lossy Compression Framework for Scientific Data.* In *SC25: International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE, 2025.

[ACM Computing Surveys'25] Sheng Di, Jinyang Liu, Kai Zhao, Xin Liang, Robert Underwood, Zhaorui Zhang, Milan Shah, **Boyuan Zhang**, et al. *A Survey on Error-Bounded Lossy Compression for Scientific Datasets. ACM Computing Surveys*, 57(11): 1–38, 2025.

[ICS'25] Wenqi Jia, Zhewen Hu, Youyuan Liu, **Boyuan Zhang**, Jinzhen Wang, Jinyang Liu, Wei Niu, et al. *NeurLZ: An Online Neural Learning-Based Method to Enhance Scientific Lossy Compression.* In *Proceedings of the 39th ACM International Conference on Supercomputing (ICS)*, pp. 26–42, 2025.

[PPoPP'25] Baixi Sun, Weijin Liu, J. Gregory Pauloski, Jiannan Tian, Jinda Jia, Daoce Wang, **Boyuan Zhang**, et al. *COMPSO: Optimizing Gradient Compression for Distributed Training with Second-Order Optimizers.* In *Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (PPoPP)*, pp. 212–224, 2025.

[FGCS'25] Franck Cappello, Mario Acosta, Emmanuel Agullo, Hartwig Anzt, Jon Calhoun, Sheng Di, Luc Giraud, **Boyuan Zhang**, et al. *Multifacets of Lossy Compression for Scientific Data in the Joint-Laboratory of Extreme Scale Computing. Future Generation Computer Systems*, 163: 107323, 2025.

[SC'24] Jinyang Liu, Jiannan Tian, Shixun Wu, Sheng Di, **Boyuan Zhang**, Robert Underwood, Yafan Huang, et al. *CUSZ-i: High-Ratio Scientific Lossy Compression on GPUs with Optimized Multi-Level Interpolation.* In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–15. IEEE, 2024.

[DAC'24] Chenxi Li, **Boyuan Zhang**, Yongqiang Duan, Yang Li, Zuochang Ye, Weifeng Liu, Dingwen Tao, and Zhou Jin. *MASC: A Memory-Efficient Adjoint Sensitivity Analysis through Compression Using Novel Spatiotemporal Prediction.* In *Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2024.

GRANTS & PROPOSALS

I contribute to the following grants and proposals by providing preliminary results, writing, and research.

[Grant Writing] Jan. 2023–Dec. 2027, "CAREER: A Highly Effective, Usable, Performant, and Scalable Data Reduction Framework for HPC Systems and Applications." NSF CAREER Award, funded, $450,000.

[Grant Writing] Jul. 2023–Jun. 2027, "Frameworks: FZ — A Fine-Tunable Cyberinfrastructure Framework to Streamline Specialized Lossy Compression Development." NSF Grant #2311876, funded, $580,000.

[Grant Writing] 2022, "Accelerating Communication in DLRM via Frequency-Aware Lossy Compression." Meta Research Award, funded, $50,000.

[Grant Writing] Oct. 2023–Sep. 2026, "SHF: Small — Reimagining Communication Bottlenecks in GNN Acceleration through Collaborative Locality Enhancement and Compression Co-Design." NSF Grant #2326495, funded, $300,000.

PRESENTATIONS & TALKS

- 06/2025, paper presentation, "BMQSim: Overcoming Memory Constraints in Quantum Circuit Simulation with a High-Fidelity Compression Framework." ICS'25, Salt Lake City, UT, USA.
- 06/2025, paper presentation, "Pushing the Limits of GPU Lossy Compression: A Hierarchical Delta Approach." ICS'25, Salt Lake City, UT, USA.
- 02/2025, poster presentation, "High-Performance Visual Semantics Compression for AI-Driven Science." PPoPP'25, Las Vegas, NV, USA.
- 11/2024, paper presentation, "Accelerating Communication in Deep Learning Recommendation Model Training with Dual-Level Adaptive Lossy Compression." SC'24, Atlanta, GA, USA.
- 06/2023, paper presentation, "FZ-GPU: A Fast and High-Ratio Lossy Compressor for Scientific Computing Applications on GPUs." HPDC'23, Orlando, FL, USA.
- 06/2023, paper presentation, "GPULZ: Optimizing LZSS Lossless Compression for Multi-Byte Data on Modern GPUs." ICS'23, Orlando, FL, USA.

TEACHING

- Spring 2026, Assisstant Instructor. ENGR-516: "Engineering Cloud Computing" Indiana University.
- Fall 2024, Assisstant Instructor. ENGR-516: "Engineering Cloud Computing" Indiana University.
- Spring 2022, Teaching Assistant. CPTS 360 "Systems Programming" Washington State University.
- Fall 2021, Teaching Assistant. CPTS 360 "Systems Programming" Washington State University.

AWARDS & HONORS

- ICS'25 Best Paper Runner-up (BMQSim), Top 3 of 320 submissions — Jun. 2025
- ICS'25 Best Paper Candidate (Aatrox), Top 6 of 320 submissions — Jun. 2025
- Reviewer for CCGRID'25 — 2025
- Reviewer for IEEE TPDS'24 and TPDS'25 — 2024–2025
- Reviewer for QCE'24 and QCE'25 — 2024–2025
- Web Chair of QCCC'24 — 2024
- Reviewer for CLOUD'23 — 2023
- Reviewer for ISSRE'23 — 2023
- Student Travel Grant ($1,500), HPDC'23 — Jun. 2023
- Graduate Conference Funding ($3,000), Indiana University Bloomington — Jun. 2022
- Program Committee (PC) Member, ISSRE'23 Artifact Evaluation (AE) — 2022
- Excellent Student Scholarship, Shanghai Jiao Tong University — Oct. 2015
- Second Prize, China Adolescents Science & Technology Innovation Contest — Aug. 2013

VOLUNTEER

- 11/2025. Student Volunteer. The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC). St. Louis, MO, USA.
- 11/2023. Student Volunteer. The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC). Denver, CO, USA.

SELECTED SOFTWARE

- **FZ-GPU**: A high-performance GPU-based error-bounded lossy compressor for scientific data.
- **GPULZ**: An optimized GPU implementation of the LZSS lossless compressor for multi-byte data.