Ruibo Fan

Email: ruibo.fan@connect.hkust-gz.edu.cn Tel: +86-22-16631928798

Addr: HKUST(GZ), 1st Duxue Road, Nansha District, Guangzhou

EDUCATION

Hong Kong University of Science and Technology (Guangzhou) **Guangzhou, China**

Data Science and Analytics Thrust ***Sept. 2022 – present***

Major: High-Performance Computing & Computer System and Architecture

Peking University (PKU|985•211) --M.S. **Beijing, China**

Academy for Advanced Interdisciplinary Studies (AAIS) ***Sept. 2019 – June. 2022***

Major: Data Science & High-Performance Computing & Parallel Computing

Huazhong University of Science and Technology (HUST|985•211) --B.S. **Wuhan, China**

School of Artificial Intelligence and Automation ***Sept. 2015 – June. 2019***

Major: Automation & Pattern Recognition GPA: 3.91/4 (Top 5%)

Dissertation: Research on parallel optimization of SVM algorithm based on CUDA

PUBLICATION

Conference

- [1] **Ruibo Fan**, et al. "ZipServ: Fast and Memory-Efficient LLM Inference with Hardware-Aware Lossless Compression." *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. 2026. (**CCF-A**)
- [2] **Ruibo Fan**, et al. "SpInfer: Leveraging Low-Level Sparsity for Efficient Large Language Model Inference on GPUs." *Proceedings of the Twentieth European Conference on Computer Systems* (EuroSys). 2025. (**CCF-A, Best Paper Award**)
- [3] **Ruibo Fan**, Wei Wang, and Xiaowen Chu. "DTC-SpMM: Bridging the Gap in Accelerating General Sparse Matrix Multiplication with Tensor Cores." *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. 2024. (**CCF-A**)
- [4] **Ruibo Fan**, Wei Wang, and Xiaowen Chu. "Fast Sparse GPU Kernels for Accelerated Training of Graph Neural Networks." *International Parallel and Distributed Processing Symposium (IPDPS)*, 2023. (CCF-B)
- [5] Luo W, Chen Y, Yu X, Wang Q, **Fan R**, Liu H, et al. "ROME: Maximizing GPU Efficiency for All-Pairs Shortest Path via Taming Fine-Grained Irregularities." *Proceedings of the 31st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. ACM, 2026. (CCF-A)
- [6] Luo W, **Fan R**, Li Z, et al. "Benchmarking and Dissecting the Nvidia Hopper GPU Architecture." *International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2024. (CCF-B)
- [7] Dong P, Li L, Zhong Y, Du D, **Fan R**, Chen Y, et al. "STBLLM: Breaking the 1-Bit Barrier with Structured Binary LLMs. " *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Journal

- [1] **Ruibo Fan**, et al. "Exploiting Low-Level Sparsity for Efficient Large Language Model Inference on GPUs with SpInfer." ACM Transactions on Computer Systems (TOCS). (**CCF-A**, Invited, Under Review)
- [2] Luo W, **Fan R,** Li Z, et al. "Dissecting the NVIDIA Hopper Architecture through Micro-benchmarking and Multiple Level Analysis." ACM Transactions on Computer Systems (TOCS). (**CCF-A**, Under Review)

Other

- [1] IEEE Std 2941.1.-2022, "Standard for Operator Interfaces of Artificial Intelligence," Dec. 2022.

TEACHING

Teaching Assistant – Mathematics for Data Science, HKUST(GZ) ***Fall 2025***

- Led 50-minute weekly lab sessions and hands-on tutorials, reinforcing core mathematical foundations (linear algebra, probability, optimization) for data science.

Teaching Assistant – Introduction to Computer Science, HKUST(GZ) ***Fall 2024; Summer 2025***

- Led 50-minute weekly lab sessions and practical exercises, helping students master fundamental computer science concepts (data structures, algorithms, basic systems) through programming.

Teaching Assistant – Parallel Computing II, Peking University ***Spring 2021***

- Designed three programming assignments and led recitation/Q&A sessions on multithreading, message passing, and distributed memory, guiding students in parallel algorithm design and performance tuning.

INTERNSHIP

Research Intern – Technology Risk and Efficiency (TRE), Alibaba Group ***Oct. 2025 – Present***

- Working on performance optimization and reliability analysis for large-scale AI and data systems.

Research Intern – Research Center for Artificial Intelligence, Peng Cheng Laboratory (PCL) ***2019;2021***

- Developed GPU-accelerated machine learning and graph workloads, including optimized SpMM/SDDMM kernels for GNNs and a high-performance AI operator library focused on GPU implementations.

RESEARCH PROJECT

Key Standards and Verification Chips for Neural Network Processors, Subtopic 4: Neural Network Processor Standards (2018AAA0103304) ***Dec.*** **2019 –** ***Dec.*** **2022**

- Contributed to drafting national and international standards for AI chip platforms, operator interfaces, and neural network compression, and developed corresponding verification and high-performance implementations.

MindSpore Operator and Network Model Development, Huawei Tech Co., Ltd ***Jun.*** **2020 –** ***Jun.*** **2021**

- Profiled and optimized GPU compute and memory bottlenecks in the MindSpore AI framework, improving operator-level performance via kernel fusion, memory coalescing, and launch-parameter tuning.
- Implemented and upstreamed classic neural network models (e.g., U-Net, FastText) to the MindSpore model zoo with GPU-optimized training and inference pipelines.

Research and development of core technology for AI supercomputer prototype, Key-Area R&D Program of Guangdong Province (2019B121204008) ***Sept.*** **2019 –** ***Sept.*** **2021**

- Co-developed a high-performance AI operator library based on tensor data abstractions, supporting multiple hardware backends and AI frameworks and optimized key sparse and dense operators using auto-tuning, data-layout reordering, and kernel specialization to improve throughput and scalability.

AWARDS

- Best Paper Award, EuroSys 2025 *(Top 2 of 85 accepted papers; overall acceptance rate 12.2%)*
- Runner-up, First DSA Excellent Research Award Scheme, 2023
- Outstanding Contribution to IEEE Standard 2941.1-2022
- Excellent Graduate, Huazhong University of Science and Technology, 2019
- Academic Excellence Scholarship, Huazhong University of Science and Technology, 2017

SERVICE

- Served as a Shadow PC Member for EuroSys 2026, supporting program committee operations including paper review coordination, session scheduling, and technical discussion facilitation.