

CHAOYI RUAN

Singapore • cfrancis@gmail.com

RESEARCH INTERESTS

My research focuses on infrastructure for large-scale machine learning, with an emphasis on cross-stack co-design across algorithms, systems, and hardware. I specialize in:

- **Scalable ML Infrastructure:** Architecting scalable components to break down monolithic models.
- **System Efficiency:** Solving context and memory bottlenecks to make machine systems economically viable.
- **Agentic Data Foundations:** Designing high-performance storage/memory systems for stateful, multimodal agents.

EDUCATION

University of Science and Technology of China (USTC) <i>Ph.D. in Computer Science and Technology</i> Advisor: Prof. Cheng Li	Hefei, China Sep 2019 – Nov 2024
University of Science and Technology of China (USTC) <i>B.S. in Computer Science and Technology</i>	Hefei, China Sep 2015 – June 2019

PROFESSIONAL EXPERIENCE

National University of Singapore (NUS) <i>Research Fellow, advised by Prof. Jialin Li</i>	Singapore 2025 – Present
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) <i>Visiting Student, advised by Prof. Xiaosong Ma</i>	UAE 2024
Qatar Computing Research Institute (QCRI) <i>Research Assistant, advised by Prof. Xiaosong Ma</i>	Qatar 2023 – 2024
Alibaba Cloud <i>Research Assistant, advised by Yingqiang Zhang</i>	Hangzhou, China 2021 – 2023

PROFESSIONAL SERVICE

• Program Committee Member, ACM ATC	2026
--	------

SELECTED PUBLICATIONS

Conference Publications

1. **Hold On Loosely: Fast, Fault-Tolerant, and Cost-Efficient Reinforcement Learning over Decentralized Network**
Chaoyi Ruan, Geng Luo, Xinyi Wan, Long Zhao, Qinghe Wang, Jiaan Zhu, Duling Xu, Guanbin Xu, Dehui Wei, Xiang Liu, Cheng Li, Haifeng Sun, Liang Luo, Congcong Miao, Jialin Li.
Proc. of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2027.
2. **Accelerating Long-Tail Generation in Synchronous RLHF Training via Adaptive Tensor Parallelism**
Long Zhao, Qinghe Wang, Jiaan Zhu, Youhui Bai, Zewen Jin, **Chaoyi Ruan**, Shannon Wang, Cheng Li.
Proc. of the International Conference on Parallel Processing (ICPP), 2026.
3. **Libra: Flexible Request Partitioning and Scheduling for Serving Unbalanced and Dynamic LLM Workloads**
Chaoyi Ruan, Yinhe Chen, Dongqi Tian, Yandong Shi, Yongji Wu, Jialin Li, Cheng Li.
Proc. of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2026.
4. **Cortex: Achieving Low-Latency, Cost-Efficient Remote Data Access For LLM via Semantic-Aware Knowledge Caching**
Chaoyi Ruan, Chao Bi, Kaiwen Zheng, Ziji Shi, Xinyi Wan, Jialin Li.
Proc. of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2026.
5. **Revisiting Parameter Server in LLM Post-Training**
Xinyi Wan, Penghui Qi, Guangxing Huang, **Chaoyi Ruan**, Jialin Li, Min Lin.
Proc. of the International Conference on Learning Representations (ICLR), 2026.

6. **DHeLlam: General-Purpose, Automatic Micro-batch Co-execution for Distributed LLM Training**
 Haiquan Wang*, **Chaoyi Ruan*** (co-first author), Jia He, Jiaqi Ruan, Chengjie Tang, Xiaosong Ma, Cheng Li.
Proc. of the IEEE International Conference on Computer Design (ICCD), 2025.
Best Paper Award
7. **Adapting to Data Affinity Changes in Geo-Replicated Database via Row-Level Paxos-Group Affiliation Re-Assignment**
Chaoyi Ruan, Yingqiang Zhang, Juncheng Zhang, Cheng Li, Xiaosong Ma, Hao Chen, Jie Zhou, Feifei Li, Xinjun Yang.
Proc. of the VLDB Endowment (VLDB), 2025.
8. **Persistent Memory Disaggregation for Cloud-Native Relational Databases**
Chaoyi Ruan, Yingqiang Zhang, Chao Bi, Xiaosong Ma, Hao Chen, Feifei Li, Xinjun Yang, Cheng Li, Ashraf Abounnaga, Yinlong Xu.
Proc. of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2023.
9. **Towards Cost-Effective and Elastic Cloud Database Deployment via Memory Disaggregation**
 Yingqiang Zhang*, **Chaoyi Ruan*** (co-first author), Cheng Li, Xinjun Yang, Wei Cao, Feifei Li, Bo Wang, Jing Fang, Yuhui Wang, Jingze Huo, Chao Bi.
Proc. of the VLDB Endowment (VLDB), 2021.
10. **SpanDB: A Fast, Cost-Effective LSM-tree Based KV Store on Hybrid Storage**
 Hao Chen, **Chaoyi Ruan**, Cheng Li, Xiaosong Ma, Yinlong Xu.
Proc. of the USENIX Conference on File and Storage Technologies (FAST), 2021.
11. **Achieving Low Tail-latency and High Scalability for Serializable Transactions in Edge Computing**
 Xusheng Chen, Haoze Song, Jianyu Jiang, **Chaoyi Ruan**, Cheng Li, Sen Wang, Nicholas Zhang, Reynold Cheng, Heming Cui.
Proc. of the European Conference on Computer Systems (EuroSys), 2021.

Journal Publications

1. **GLPilot: Efficient Distributed GNN Training with Learnable Embeddings**
 Chengru Yang, **Chaoyi Ruan**[†], Chengjie Tang, Ping Gong, Shiyi Wang, Xiang Song, Cheng Li[†].
IEEE Transactions on Parallel and Distributed Systems (TPDS), 2025. ([†]Corresponding Authors)
2. **Leveraging NVMe SSDs for Building A Fast, Cost-Effective LSM-tree Based KV Store**
 Cheng Li, Hao Chen, **Chaoyi Ruan**, Xiaosong Ma, Yinlong Xu.
ACM Transactions on Storage (TOS), 2021.

Preprints & Workshops

1. **Tetris: Efficient and Predictive KV Cache Offloading for Agentic and Reasoning Workloads**
 Ziji Shi, **Chaoyi Ruan**, Penghui Qi, Guangxing Huang, Xinyi Wan, Min Lin, Jialin Li.
1st Workshop on Systems for Agentic AI (at SOSP), 2025.
2. **Fanar: An Arabic-Centric Multimodal Generative AI Platform**
 The Fanar Team (including **Chaoyi Ruan**).
Technical Report, 2025.

OPEN SOURCE & TECHNICAL SKILLS

- **DHeLlam** (github.com/DHeLlam-AI/dhellam): Automatic micro-batch co-execution for distributed LLM training.
- **SpanDB** (github.com/SpanDB/SpanDB): High-performance LSM-tree KV store for NVMe SSDs. Compatible with RocksDB.
- **Languages & Tools**: C/C++, Python, PyTorch, RocksDB, MySQL.
- **Systems Expertise**: Distributed Training, Memory Disaggregation, KV Stores, Geo-Replicated Databases.

HONORS & AWARDS

- | | |
|--|------|
| • Best Paper Award, ICCD | 2025 |
| • Outstanding Doctoral Dissertation Award, CCF Database Technical Committee | 2025 |
| • Suzhou Yucai Scholarship, USTC | 2021 |
| • Yuanqing Scholarship, USTC | 2018 |