Letian Ruan

ruanlt@umich.edu | github.com/Risc-lt | Risc-lt.github.io

Education

University of Michigan, Ann Arbor 2025/08 – Present
B.S.E in Data Science, Computer Science and Engineering Department Michigan, USA

Shanghai Jiao Tong University 2022/08 – Present
B.S. in Mechanical Engineering, UM-SJTU Joint Institute (Duel Degree) Shanghai, China

Interest

My research interests mainly lie in Machine Learning Systems, particularly in designing efficient and scalable systems for GenAI. I am particularly interested in hardware-alignment optimization, large-scale distributed execution and system support for emerging AI paradigms like AIGC and RLHF. I also have a background in machine learning, GPU architectures, serverless data centers, operating systems and databases.

Experience

Research Intern, *SymbioticLab* 2025/08 - Present
advised by Mosharaf Chowdhury, UMich CSE department Ann Arbor, Michigan

- Serving Any-to-Any Multimodal LLMs with high throughput
  - Designed an offline planner and online distributed runtime that automatically solve the optimal resource allocations for heterogeneous multimodal requests. Achieved up to 3.8× throughput and 9.5× lower P99 latency versus state-of-the-art baselines in experiments. Under review of ***NSDI2026***.
  - Role: developer and maintainer of the open-source system project

Research Intern, *kvcache-ai org* 2025/05 - Present
advised by Teng Ma, Alibaba Cloud & MADSys Group at Tsinghua University Remote

- Promoting KVCache-centric disaggregated architecture for LLM Serving
  - Optimized the KVCache transfer pipeline under prefill-decode disaggregation in LMDeploy by integrating Mooncake transfer engine as the RDMA-based migration backend, achieving a 10% reduction in TTFT for disaggregated prefill-decode inference. Best Paper of ***FAST2025***, 4k ***Star***.
  - Role: team member of core developers

Research Assitant, *EPCC Lab* 2024/10 - 2025/10
advised by Shixuan Sun, School of Computer Science at SJTU Shanghai, China

- Reducing Latency for Multi-tenant concurrent LoRA serving
  - Decoupled the storage and computation of LoRA adaptors from the base LLM with diaggregated LoRA Store Server to solve insufficient GPU memory for LoRA cache, reducing the P95 TTFT by 50%, while maintaining an SLO attainment above 90% for more than 99% of LoRA adapters. In submission to ***MLSys2026***.
  - Role: project co-leader; designed and implemented the system.
- Building disaggregated architecture for serverless graph processing
  - Introduced disaggregatd architecture to enable multi-tier data communication by using container images as cache, serverless containers as intermediaries and autonomous-elastic functions as workers. The system delivers up to 3.8× higher compute performance and reduces monetary cost by up to 63.7%. Under review of ***SIGMOD2026***
  - Role: project core contributor; implemented the system; conducted experiments.

Publication

HLSS: Improving Multi-LoRA inference performenace with LoRA Store Server
Hongyu Chen*, Letian Ruan*, Shixuan Sun [Under review]***MLSys2026***

FaaSBoard: Efficient Graph Processing with a Disaggregated Architecture on Serverless Services
Yushi Liu, Yikang Ruan, Letian Ruan, Shixuan Sun, Bingsheng He, Minyi Guo [Under review]***SIGMOD2026***

Bridging the GPU Utilization Gap: Predictive Multi-Dimensional Resource Scheduling for AI Workloads
Yilei Lu, Dongbiao He, Teng Ma, Zhe Liu, Letian Ruan, Jinlei Jiang, Yongwei Wu [Under review]***EuroSys2026***

Projects

BusTub *A complete row-store relational database system* 2024/07 – 2024/09

- Implemented a buffer pool to improve memory utilization, developed Extensible HashTable for indexing and used Crabbing Protocol to add read-write locks.
- Completed basic SQL operators and optimizers on the volcano model and built a centralized transaction manager based on the MVCC protocol, supporting snapshot isolation.

xv6-riscv *A basic operating system based on the RISC-V architecture* 2024/02 – 2024/04

- Implemented Lazy Allocation and Copy-On-Write mechanisms in xv6, optimizing memory management and improving process creation efficiency.
- Added thread libraries and synchronization barriers for each CPU in kernel modules and optimized buffer pool replacement management using finer-grained hash tables.

LSM-KV *A key-value storage engine based on LSM-tree* 2024/04 – 2024/05

- Implemented LSM-Tree and skiplist-based MemTable and introduced vLog with reference to Wisckey LSM, achieving key-value separation.
- Implemented incremental merging, range queries, and garbage collection, and improved storage engine performance with asynchronous I/O.

Teaching

Shanghai Jiao Tong University Shanghai, China

FA.2024 Accelerated Introduction to Computer and Programming(ENGR1510J) **Teaching Assistant**

Services

jCourse *The largest unofficial course review platform of SJTU*

- As of November 2024, more than 28,000 users have posted over 33,000 reviews for 4,500+ courses.
- Serve as the member of main developing group for the latest version of the platform.

SKills

- **Programming languages:** C/C++, Python, CUDA, Golang and RISC-V Assembly.
- **Frameworks:** PyTorch, vLLM, SGLang, LMDeploy, slime, cutlass/cuBLAS, kubernetes and AWS.