Ruoyu Qin — Curriculum Vitae

Email: qinry24@mails.tsinghua.edu.cn
Phone: (+86) 132-2390-6909
Github: Chestnut-Q
Homepage: https://qinruoyu.com

Education

Tsinghua University — Beijing, China
Ph.D. in Computer Science and Technology — 2029 *(expected)*
- Advisor: Prof. Mingxing Zhang

Tsinghua University — Beijing, China
B.E. in Computer Science and Technology — 2024
- GPA: 3.95/4.00, Rank: 10/182
- **Outstanding Graduate Award** (Top 5% of graduating class)

Research Interst

My research centers on efficient machine learning systems, with an emphasis on large-scale, distributed large language model (LLM) deployments. I'm driven by real-world industrial needs: I seek out practical challenges in areas like LLM serving and RL rollout, develop solutions, and push them all the way into production.

Publications

Mooncake: Trading More Storage for Less Computation — A KVCache-centric Architecture for Serving LLM Chatbot
Ruoyu Qin*, Zheming Li*, Weiran He, Jialei Cui, Feng Ren, Mingxing Zhang, Yongwei Wu, Weimin Zheng, Xinran Xu. (* indicates co-first author)
Erik Riedel Best Paper Award!
FAST 2025 – *The 23rd USENIX Conference on File and Storage Technologies*.

Projects

Mooncake Store

Mooncake Store aims to enhance the inference efficiency of large language models (LLMs), especially in slow object storage environments, by constructing a multi-level caching pool on high-speed interconnected DRAM/SSD resources. Compared to traditional caching, Mooncake utilizes (GPUDirect) RDMA technology to transfer data directly from the initiator's DRAM/VRAM to the target's DRAM/SSD in a zero-copy manner, while maximizing the use of multi-NIC resources on a single machine. Mooncake has received 3.5k stars and has been integrated into several LLM inference frameworks (vLLM, SGLang).

FlashAttention from Scratch

FlashAttention from Scratch (FAS) is a learning project based on FlashAttention-2, featuring from-scratch implementations using both GPU and CPU.

Experience

Moonshot AI Infra Team | Beijing, China — 2023.09 - Present
- Research Intern
- **Mentor:** Xinran Xu and Weiran He

THUMT | Beijing, China — 2022.09 - 2023.06
- Research Assistant
- **Advisor:** Yang Liu

Awards and Honors

- **FAST'25 Best Paper Award**, USENIX Association 2025
- **Outstanding Graduate Award**, Tsinghua University 2023
- **Tang Lixin Outstanding Scholarship**, Tsinghua University 2020
- **Gold Medal, 32nd Chinese Chemistry Olympiad**, Chinese Chemical Society 2018