Ruoyu Qin Ph.D.

qinry24@mails.tsinghua.edu.cn
https://qinruoyu.com

Education

Tsinghua University — Beijing, China
Ph.D. in Computer Science and Technology — 2028 *(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.
Erik Riedel Best Paper Award!
USENIX Conference on File and Storage Technologies (FAST), Santa Clara, CA, USA, 2025.

Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving
Ruoyu Qin*, Zheming Li*, Weiran He, Jialei Cui, Heyi Tang, Feng Ren, Teng Ma, Shangming Cai, Yineng Zhang, Mingxing Zhang, Yongwei Wu, Weimin Zheng, Xinran Xu.
ACM Transactions on Storage, 2025.

Seer: Online Context Learning for Fast Synchronous LLM Reinforcement Learning
Ruoyu Qin, Weiran He, Weixiao Huang, Yangkun Zhang, Yikai Zhao, Bo Pang, Xinran Xu, Yingdi Shan, Yongwei Wu, Mingxing Zhang.
USENIX Symposium on Operating Systems Design and Implementation (OSDI), Seattle, WA, USA, 2026.

Prefill-as-a-Service: KVCache of Next-Generation Models Could Go Cross-Datacenter
Ruoyu Qin, Weiran He, Yaoyu Wang, Zheming Li, Xinran Xu, Yongwei Wu, Weimin Zheng, Mingxing Zhang.
arXiv:2604.15039.

TENT: A Declarative Slice Spraying Engine for Performant and Resilient Data Movement in Disaggregated LLM Serving
Feng Ren, **Ruoyu Qin**, Teng Ma, Shangming Cai, Zheng Liu, Chao Lei, Dejiang Zhu, Ke Yang, Jinyang Su, Weixiao Huang, Yikai Zhao, Yongwei Wu, Weimin Zheng, Mingxing Zhang.
Frontier AI Systems Workshop (FAISys), Hong Kong, China, 2025.

Projects

Mooncake

Mooncake is the KVCache-centric serving platform behind Kimi, Moonshot AI's flagship LLM service. It disaggregates the prefill and decoding stages and pools otherwise-underutilized CPU, DRAM, and SSD resources into a multi-tier KVCache store, trading cheap storage for expensive computation to boost long-context throughput. At its core, a high-performance Transfer Engine moves data in a zero-copy manner across heterogeneous transports (TCP, RDMA/GPUDirect RDMA, NVMe-oF), saturating multi-NIC bandwidth to stream KVCache directly between the initiator's DRAM/VRAM and the target's DRAM/SSD. Recognized with the Best Paper Award at USENIX FAST 2025, Mooncake has gathered 5.6k GitHub stars and been integrated into mainstream LLM inference frameworks including vLLM, SGLang, TensorRT-LLM, and LMDeploy.

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, Thirty-Second Chinese Chemistry Olympiad**, Chinese Chemical Society 2018

Services

- **Artifact Evaluation Committee**, OSDI'26 2026
- **Teaching Assistant**, Algorithms and Its Complexity Theory 2026