Yifan Sui

suiyifan@sjtu.edu.cn

Personal Statement

I am a fifth-year Ph.D. candidate in the School of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University, advised by Prof. Jianxun Li. I collaborate closely with Prof. Hao Wang (Stevens Institute of Technology). I received my B.Eng. in Telecommunications Engineering from Beijing University of Posts and Telecommunications. I am now a Research Intern at Microsoft Research Asia (Systems Group) from Summer 2025.
My research interests include **serverless computing**, **efficient LLM systems**, and **cloud systems**. I received the **ACM SoCC 2024 Best Paper Award**.

Education

Shanghai Jiao Tong University *2021–Present*
Ph.D. Candidate, Control Science and Engineering

Beijing University of Posts and Telecommunications *2017–2021*
B.Eng. in Telecommunications Engineering

- Average score: 89/100; rank: top 8% (41/522).

Research Experience

Act While Thinking: Accelerating LLM Agent Serving via Speculative Tool Execution *Under submission*

- Conducted during my internship at Microsoft Research.
- Identified that tool execution dominates latency for tool-augmented LLM agents, and the agentic workflows can defeat serverless & MicroService schedulers.
- Found recurring short-horizon workflow "patterns" that are locally predictable.
- Proposed PASTE, a pattern-aware speculative tool execution framework that predicts near-future tool calls and derives arguments.
- Achive up to 48.5% lower task completion time and 1.8× higher throughput.

xLoRA: Faster and Cheaper LoRA LLM Serving with Serverless Computing *Under submission (arXiv:2505.14468)*

- Identified key inefficiencies in serverless LoRA serving due to redundant backbone loading, heavy cold-start overheads, and amplified GPU contention.
- *xLoRA* enables secure backbone sharing across isolated functions, opportunistic pre-loading of full LoRA artifacts, and contention-aware batching/offloading.
- *xLoRA* reduces time-to-first-token (TTFT) by up to 86% and cost by up to 89% compared to state-of-the-art solutions.

Pre-Warming Is Not Enough: Accelerating Serverless Inference With Opportunistic Pre-Loading *ACM SoCC 2024 (Best Paper Award)*

- Showed that, for ML inference functions, loading libraries and models can take several times longer than container cold start, yet is ignored by existing work.

- Proposed *InstaInfer*, a model-specific serverless platform that mitigates ML function loading delay.
- *InstaInfer* is compatible with state-of-the-art cold-start mitigation solutions; it reduces loading latency by up to 98% and achieves up to 6× end-to-end speedup.

Accelerating ML Inference via Opportunistic Pre-Loading on Serverless Clusters *IEEE TPDS 2025*

- Developed an efficient and global-optimal approach for opportunistic pre-loading in large-scale clusters.
- Proposed *Tyche*, which mitigates ML function loading delay with minimal overhead via cluster-wide scheduling and lightweight locality-aware load balancing.

Work Experience

Microsoft Research, Shanghai, China *Jun 2025–Present*
Research Intern, System Group (Mentor: Yuqing Yang)

Stevens Institute of Technology, Hoboken, USA *Aug 2022–Present*
Research Intern, IntelliSys Lab (Advisor: Prof. Hao Wang)

Shanghai Jiao Tong University, Shanghai, China *Feb 2022–Jul 2022*
Teaching Assistant (ICE2501: Signals and Systems)

Beihang University, Beijing, China *Jan 2020–Oct 2020*
Research Assistant (Advisor: Prof. Hui Zhang, IEEE Fellow)

Skills

Distributed systems	OpenWhisk, Kubernetes, Docker
LLM Systems	SGLang, Transformers
Language	TOEFL iBT: 101

Publications

1. Yifan Sui, Han Zhao, Rui Ma, Hao Wang, Zhiyuan He, Jianxun Li, Yuqing Yang. "Act While Thinking: Accelerating LLM Agent Serving via Speculative Tool Execution." Manuscript under submission, 2026.
2. Yifan Sui, Hao Wang, Hanfei Yu, Yitao Hu, Chen Chen, Jianxun Li. "xLoRA: Faster and Cheaper LoRA LLM Serving with Serverless Computing." Manuscript under submission, 2025.
3. Yifan Sui, Hanfei Yu, Yitao Hu, Jianxun Li, Hao Wang. "Accelerating ML Inference via Opportunistic Pre-Loading on Serverless Clusters." *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 2025.
4. Yifan Sui, Hanfei Yu, Yitao Hu, Jianxun Li, Hao Wang. "Pre-Warming Is Not Enough: Accelerating Serverless Inference With Opportunistic Pre-Loading." *ACM Symposium on Cloud Computing (SoCC)*, 2024.
5. Yifan Sui, Meng Cai, Jianxun Li. "GuardGrid: A High-Availability Cloud Platform for Deep Learning Applications." *Cluster Computing*, 2025.
6. Yifan Sui, Shaodong Zhou, Zhiyang Ju, Hui Zhang. "A Vision-Based System Design and Implementation for Accident Detection and Analysis via Traffic Surveillance Video." *IEEE Canadian Journal of Electrical and Computer Engineering*, 2022.