Siyang Shao

siyangshao@gmail.com | (470) 378-9780 | github.com/SiyangShao | linkedin.com/in/shaosy/

Education

Georgia Institute of Technology Aug 2025 – Jun 2027
Master of Science in Computer Science; GPA: 4.0 / 4.0 Atlanta, Georgia, USA

Nanyang Technological University Aug 2021 – Jun 2025
Bachelor of Engineering (Computer Engineering); Honours (Highest Distinction); GPA: 4.63 / 5.0 Singapore

Work Experience

Tiktok San Jose, California, USA
Software Engineer Intern, Recommendation Infrastructure May 2026 - Aug 2026

- Designed and implemented an autonomous, end-to-end AI agent to optimize recommendation model inference pipelines.
- Enpowred the agent with capabilities to analyze model graphs and determine optimal subgraph partitions for kernel fusion.
- Implemented a closed-loop self-evolution harness that empowers the agent to autonomously generate, profile, and refine CUDA/Triton kernels based on real-time performance feedback.

Jane Street Hong Kong
Software Engineer Intern May 2025 - Jul 2025

- Built a unified **JSON-RPC / Async-RPC** declaration library with automatic version conversion, replacing a manual dual-maintenance workflow and eliminating **~50 lines of duplicate code** per endpoint; deployed across **2 services**.
- Developed a SQL-compatible database mirror to replace a complex legacy DSL system, reducing onboarding from a **10+ page internal guide** to familiar SQL; covered **8 core schemas** and their derived schemas with an incremental sync pipeline achieving a **5x query speedup**.

TikTok Singapore
Software Engineer Intern, Video Infrastructure Jan 2024 - May 2024

- Developed a metadata service managing **1,000+ microservices** by actively capturing metrics, which standardized governance and unified regional visibility for SREs.
- Built a persistent SLI framework supporting **20+ configurable indicators**, enabling SREs to trace and audit historical alert metrics with pre-computed dashboards instead of ad-hoc queries.

Projects

ServerlessLLM (**600+ ⋆**) — *Fast Checkpoint Loading for LLM Serving* github.com/ServerlessLLM/ServerlessLLM
Core Maintainer 2024 – 2026

- Engineered **ROCm support** for high-throughput model loading on AMD GPUs, achieving **6–10x faster cold-start** latency.
- Developed the **system controller** to coordinate backend lifecycles (init, scaling), integrating the system with **vLLM**, **ray**, etc., ensuring reliability under multi-tenant workloads.
- Led code reviews, issue triage, and documentation for a community-driven project with contributors across academia and industry.

Liquid — *Adaptive LLM Inference System with Dynamic Tensor Parallelism*
Core Contributor 2024 – 2025

- Discovered that optimal tensor parallelism (TP) level in LLM serving varies with input/output sequence lengths; designed a scheduler leveraging **live migration** and **dynamic tensor parallelism** to adjust TP levels at runtime for general LLM serving.
- Achieved **sub-1s resharding** on NVLink-connected clusters built on **vLLM**, delivering **1.5x–3.3x throughput improvement** over ServerlessLLM + vLLM baselines while maintaining P95 latency SLO.

Selected Awards

- ICPC Asia Pacific Manila Regional: **Ranked 2nd (Silver Medal)** Dec 2022
- ICPC Asia Pacific Championship: **Ranked 22** (‘24) & **24** (‘25) out of top regional qualifiers 2024 - 2025
- Dean’s List (Top 5% of cohort) Aug 2023

Skills

- Programming Languages: C++, Python, Rust, OCaml, Go
- Tech: CUDA, Triton, vLLM, SGLang, PyTorch, Ray, ZeroMQ, gRPC, Docker, Kubernetes