SHAO Siyang

siyang.shao@outlook.com | +86-15021988618 | github.com/SiyangShao

Education

Georgia Institute of Technology — Aug 2025 – Current
PhD student in Computer Science — Atlanta, Georgia, USA
- Research Area: Large Language Model Inference, Distributed System

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

Research Experience

Cluster Level Scheduing for LLM Inference
Supervised by Prof. Dmitrii Ustiugov — Mar 2024 – May 2025
- Investigated cluster-level scheduling for large language model inference in serverless systems
- Explored optimal scaling policies and mechanisms for serverless LLM environments

Open Source Projects

ServerlessLLM — https://github.com/ServerlessLLM/ServerlessLLM
Developer — Jun 2024 – May 2025
- Supported ROCm for `sllm-store`, the internal library of ServerlessLLM which provides high-performance model loading
- Integrated vLLM backend, enabling ServerlessLLM project to perform inference through vLLM
- Explored methods to enable vLLM backend to benefit from high-performance model loading via `sllm-store`
- Maintained the controller of the ServerlessLLM project, which manages the lifecycle of the inference backends

Work Experience

Jane Street Asia Limited — Hong Kong SAR
Software Engineer Intern — May 2025 – Jul 2025
- Designed a library to support callee version conversion for JSON-RPC, enabling seamless alignment between JSON-RPC and async-RPC. This allows both RPC types to be registered using a unified declaration and automates version conversion, simplifying integration and maintainance.
- Designed a database prototye that incrementally synchronizes an internal database (defined by a custom DSL) with a new SQL database. This approach simplifies usage for users and achieves approximately 5x faster query performance.

TikTok Pte. Ltd. — Singapore
Backend Engineer Intern (Video Infrastructure) — Jan 2024 – May 2024
- Co-Designed and implemented metrics metadata discover and manage system, bridged the gap between development teams and SRE teams concerning the monitoring metrics
- Implemented persistent global SLI monitor and manage system, contributing to improvements in full-link stability

Co-Curricular Activities

NTU ICPC Team — https://icpc.global/ICPCID/B15T259WIX3C
Team Member — Dec 2021 – Mar 2025
- Represented the school in ICPC (International Collegiate Programming Contest) and solved complex algorithm problems

Awards

- 2022 ICPC Asia Manila Regional Ranked 2 — Dec 2022
- 2023 ICPC Asia Jakarta Regional Ranked 13 — Dec 2023
- 2024 ICPC Asia Pacific Championship Ranked 22 — Mar 2024
- 2025 ICPC Asia Jakarta Regional Ranked 11 — Dec 2024
- 2025 ICPC Asia Pacific Championship Ranked 24 — Mar 2025
- Dean's List in Academic Year 2022-23 (Top 5% of cohort) — Aug 2023
- NTU President Research Scholar in Academic Year 2023-24 — Aug 2024