SHAO Siyang

shao0054@e.ntu.edu.sg | +65-98602734 | github.com/SiyangShao

Education

Nanyang Technological University Aug 2021 – Jun 2025
Bachelor of Engineering (Computer Engineering) Singapore, Singapore
- Expected: Honours (Highest Distinction); GPA: 4.60 / 5.0
- Dean's List (Academic Year 2022-23)

Skills Summary

- **Languages**: Golang, C++, Python, CUDA, Java, etc.
- **Tools**: Docker, Linux, Kubernetes, Knative, vLLM, bosun, kafka, clickhouse, etc.

Work Experience

TikTok Pte. Ltd. Singapore
Backend Engineer Intern (Video Infrastructure) Jan 2024 – May 2024
- Implement persisting large volumes of monitoring data by using **message queues** for exporting the **bosun** data into **clickhouse**, contributing to improvements in full-link stability.
- Participating in the development of an observation center, maintained all relevant video architecture **SLI** (service level indicator) and **SLA** (service level agreement) metrics, and monitoring system alarms in real-time.

Academic Projects

NTU competitive Programming (ICPC) Team https://icpc.global/ICPCID/B15T259WIX3C
Team FailedSystemTest / CheesyLeopard / NTRLover Dec 2021 – Mar 2024
- **Impact**: Representing the school in **competitive programming contests**, using **C++** to solve complex algorithm questions.
- **Awards**: Ranked 2 and secured the **Silver medal** at the 2022-23 ICPC Asia Manila Regional, solved 6 problems and ranked 22 for 2023-24 **ICPC Asia Pacific Championship**.

vHive Community https://github.com/vhive-serverless/vHive
Supervised by Professor Dmitrii Ustiugov May 2023 – Current
- NVIDIA MIG on GPU Performance
  - Using **NVIDIA MIG** on A100, segmenting a single GPU into multiple MIG instances and Compute Instances to investigate their impact on LLM inference latency.
- Cluster level Large Language Model Inference
  - Based on **vLLM**, implement a worker with **grpc** functions to retrieve the model and generate the inference result.
  - Leveraged vLLM's **Paged Attention** and **Continuous Batching**, it could dynamically adjust inference batch size, optimize GPU computational power and memory utilization.

URECA Project - Deoxys https://github.com/SiyangShao/Deoxys
Fast Software Implementations of New Cryptographic Primitives Mar 2023 – Jul 2023
- Implement a system that optimizes calculations using Intel AES intrinsics on the x86-64 architecture, using **C** language
- Analyzing **assembly** code for instruction set pipelining, achieve a balance in the latency and CPI of the instructions

Co-Curricular Activities

NTU Open Source Society
HackOSS Technical Director Jun 2022 – Jun 2023
- Organized open-source community events in 'HackOSS Day'
- Lead team to complete projects, help team members learn and use open-source tools

Awards

- 2022 ICPC Asia Manila Regional Silver Medal (Ranked 2) Dec 2022
- 2023 ICPC Asia Jakarta Regional Ranked 13 Dec 2023
- 2024 ICPC Asia Pacific Championship Ranked 22 Mar 2024