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

Skills Summary

- **Languages**: Golang, C++, Python, CUDA
- **Tools**: Docker, vLLM, ray, hipify, bosun, grpc, Kubernetes, Knative, clickhouse

Research Experience

LLM Inference in Serverless Systems
Supervised by Dmitrii Ustiugov Mar 2024 – Current
- Investigated cluster-level scheduling for large language model inference in serverless systems
- Explored optimal scaling policies and mechanisms for serverless LLM environments
- Utilized GPU memory usage for a memory-centric scheduling LLM inference system
- Optimized overall throughput and reduced request queueing latency

MIG-based GPU Partitioning and Performance Analysis
Supervised by Dmitrii Ustiugov Jun 2023 – Nov 2023
- Explored the use of MIG (Multi-Instance GPU) to physically partition a single GPU
- Investigated the performance overhead associated with achieving physical isolation across multiple instances
- Compared performance gains and trade-offs of using MIG for multi-model scenarios on a single GPU
- Analyzed memory and PCIe bandwidth utilization across multiple MIG instances

Open Source Projects

ServerlessLLM https://github.com/ServerlessLLM/ServerlessLLM
Contributor Jun 2024 – Current
- 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

Co-Curricular Activities

NTU ICPC Team https://icpc.global/ICPCID/B15T259WIX3C
Team Member Dec 2021 – Mar 2024
- 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
- 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

Work Experience

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