邵思洋

siyangshao@gmail.com | (+86) 150-2198-8618 | github.com/SiyangShao | linkedin.com/in/shaosy/

教育经历

佐治亚理工学院 (Georgia Institute of Technology) 2025.08 – 2027.06
计算机科学硕士;GPA:4.0 / 4.0 美国 亚特兰大

南洋理工大学 (Nanyang Technological University) 2021.08 – 2025.06
计算机工程学士;最高荣誉学位 Honours (Highest Distinction);GPA:4.63 / 5.0 新加坡

实习经历

TikTok 美国 圣何塞
研发实习生 — 推荐架构 (Recommendation Infrastructure) 2026.05 - 2026.08
- 待补充 (TBD)

Jane Street 中国香港
软件研发实习生 2025.05 - 2025.07
- 搭建统一的 **JSON-RPC / Async-RPC** 声明库,支持自动版本转换,替换原有需双份手动维护的工作流,每个接口节省 **约 50 行重复代码**,上线至 **2 个线上服务**。
- 开发 SQL 兼容的数据库镜像层,替代复杂的存量 DSL 系统,新人上手成本从 **10+ 页内部文档** 降低为直接写 SQL;覆盖 **8 个核心 schema** 及其衍生 schema,通过增量同步流水线实现 **查询性能 5 倍提升**。

TikTok 新加坡
研发实习生 — 视频架构 (Video Infrastructure) 2024.01 - 2024.05
- 开发指标元数据服务,主动采集 metrics,覆盖 **1,000+ 微服务**,统一治理规范并打通跨区域 SRE 可观测视图。
- 搭建持久化 SLI 框架,支持 **20+ 可配置指标**,SRE 可通过预计算看板回溯与审计历史告警数据,替代临时查询。

项目经历

ServerlessLLM (**600+ ★**) — 面向 *LLM Serving* 的快速 *Checkpoint* 加载系统github.com/ServerlessLLM/ServerlessLLM
Core Maintainer 2024 – 2026
- 实现 **ROCm 支持**,使 AMD GPU 上支持高吞吐模型加载,冷启动延迟降低 **6-10 倍**。
- 开发 **系统控制器 (system controller)**,负责推理后端的生命周期管理(初始化、扩缩容),并完成与 **vLLM**、**Ray** 等框架的集成,保障多租户场景下的稳定性。
- 主导 Code Review、Issue 处理与文档维护,推动学术界与工业界贡献者共建的开源社区运转。

Liquid — 基于动态张量并行的自适应 *LLM* 推理系统
Core Contributor 2024 – 2025
- 发现 LLM Serving 场景下最优 Tensor Parallelism (TP) 配置随输入/输出序列长度变化;设计基于 **热迁移 (live migration)** 与 **动态张量并行** 的调度器,运行时为通用 LLM 服务动态调整 TP。
- 在 NVLink 互联集群上基于 **vLLM** 实现 **亚秒级 resharding**,在保持 P95 延迟 SLO 的前提下,相比 ServerlessLLM + vLLM 基线带来 **1.5-3.3 倍吞吐提升**。

获奖经历

- ICPC 国际大学生程序设计竞赛 亚洲马尼拉站:**银牌 (区域赛第 2 名)** 2022.12
- ICPC 亚太区总决赛:在各区域赛晋级队伍中分别取得 **第 22 名** (2024) 与 **第 24 名** (2025) 2024 - 2025
- Dean’s List 院长荣誉名单(年级前 5%) 2023.08

专业技能

- 编程语言:C++、Python、Rust、OCaml、Go
- 技术栈:CUDA、Triton、vLLM、SGLang、PyTorch、Ray、ZeroMQ、gRPC、Docker、Kubernetes