邵思洋

shao0054@e.ntu.edu.sg | +86-15021988618 | github.com/SiyangShao

教育背景

南洋理工大学(NTUsg) 2021 年 8 月 – 2025 年 6 月
工学学士 (计算机工程) 新加坡

- 一等一荣誉学位; GPA: 4.60 / 5.0
- 院长名单 (2022-23 学年)

技能

- **编程语言**: Golang, C++, Python, CUDA, Java, etc.
- **工具**: Docker, Linux, Kubernetes, Knative, vLLM, bosun, kafka, clickhouse, etc.

工作经历

TikTok Pte. Ltd. (字节跳动) 新加坡
后端开发实习生 (视频架构) 2024 年 1 月 – 2024 年 5 月

- 利用 clickhouse 和消息队列实现可持久化 bosun 数据落盘, 提升 Tiktok 全链路稳定性
- 参与开发观测中心, 维护视频架构相关 SLI 和 SLA 指标, 并利用指标配置事故预设, 辅助计算事故等级

学术项目

NTU ICPC 团队(算法竞赛) https://icpc.global/ICPCID/B15T259WIX3C
2021 年 12 月 – 2024 年 3 月

- 代表学校多次参与算法竞赛, 使用 C++ 解决复杂算法问题
- 最高曾获马尼拉区域赛亚军, 并在 2023-24 ICPC 亚太总决赛 (包含来自日本, 韩国, 台湾, 越南, 新加坡等地高校) 中排名第 22.

vHive 社区 https://github.com/vhive-serverless/vHive
由 Dmitrii Ustiugov 教授指导 2023 年 5 月 – 至今

- NVIDIA MIG 对 GPU 推理性能的研究
  - 在 A100 GPU 上利用 NVIDIA MIG 技术, 将 GPU 划分为多个实例, 评估单卡多模型情况下如何优化 GPU 推理性能
- 集群级别的 GPU 推理性能优化
  - 基于 vLLM, 利用 grpc 实现一个 worker, 用于生成推理结果并支持集群级别的 GPU 推理调度
  - 利用 vLLM 提供的 Paged Attention 和 Continuous Batching, 动态调整推理批处理大小, 优化 GPU 计算能力和内存利用率

URECA 项目 - Deoxys https://github.com/SiyangShao/Deoxys
对新密码学加密工具的快速实现 2023 年 3 月 – 2023 年 7 月

- 利用 Intel AES 指令集, 实现一个能够优化 Deoxys 算法系统
- 分析生成的汇编代码, 以优化指令集流水线, 达到延迟和 CPI 的平衡

奖项

- 2022-23 ICPC 亚洲马尼拉区域赛 亚军 2022 年 12 月
- 2023-24 ICPC 亚洲雅加达区域赛 13 名 2023 年 12 月
- 2023-24 ICPC 亚洲太平洋区域总决赛 22 名 2024 年 3 月