沈一芃

github | mysite | 1669335639@qq.com | +86 18310211513

教育经历

浙江大学 学士 – 计算机科学与技术 Sept 2022 - June 2026

- **GPA:** 3.99/4.3 或 88.38/100
- **相关课程:** 高性能计算，计算机系统概论，计算机组成，计算机体系结构，操作系统，计算机网络，数字逻辑设计，机器学习与数据分析，自然语言处理，编译原理，人工智能，微积分，线性代数，概率论与数理统计。
- **研究方向**: **机器学习系统**, **LLM 高效推理**, **Multi-Agent-System**.

发表论文

- Zaifeng Pan, Ajjkumar Patel, **Yipeng Shen**, Zhengding Hu, Yue Guan, Wan-Lu Li, Lianhui Qin, Yida Wang, Yufei Ding. "KVFlow: Efficient Prefix Caching for Accelerating LLM-Based Multi-Agent Workflows." *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- Anonymous. "ScaleSim: Scaling Multi-Agent Simulation with Invocation Distance-Based Memory Management." *Submitted to The Forty-Third International Conference on Machine Learning (ICML)*, 2026. *Under review.* **(共一)**
- Anonymous. "LIF Recurrent Memory Enables Long-Horizon Spiking Computation." *Submitted to The Forty-Third International Conference on Machine Learning (ICML)*, 2026. *Under review.* **(二作)**

研究经历

UCSD Picasso Lab *Advised by Prof. Yufei Ding* Mar 2025 - Present

- 基于 SGLang 框架，优化 multi-agent 场景下的 GPU Memory。主要负责修改 SGLang 后端引擎，实现新的内存结构、读写队列、interrupt 机制与 LoRA 调度器。
- 对应 KVFlow 论文工作，涉及 Multi-Agent 工作流的 Prefix Caching 优化。
- 对应 ScaleSim 项目，涉及 Multi-Agent Simulation 的内存管理优化。

浙江大学 ZIP Lab *Advised by Prof. Bohan Zhuang* June 2025 - Present

- 探索基于自回归和 Diffusion LLM 的新型文本生成范式。基于 KVCache Reuse 与 Sparse Attention 稀疏注意力机制进行加速，实现高效、高保真的文本生成。

浙江大学 CCNT Lab *Advised by Prof. Peng Lin* June 2024 - Nov 2025

- 基于类脑神经元件，设计和实现高准确率、低能耗的脉冲神经网络。结合 LIF 神经元的脉冲特性与 LSTM 的记忆门单元，实现 LRMM 结构，保存梯度和长期记忆的同时保证移步稀疏脉冲。
- 对应 SNN 论文工作，使用 Recurrent LIF Memory Module 处理长程依赖。

浙江大学 RC4ML Lab *Advised by Prof. Zeke Wang* Nov 2024 - Mar 2025

– 探索并实现基于 DSP48 的大数乘法加速算法与卷积计算加速算法。

– 设计针对 AMD VERSAL Network on Chips 传输带宽的 benchmark。

项目经历

NUS Cloud Computing Summer Workshop Github Page

– 领导 *Cloud on Cloud* 项目, 搭建实时天气预报系统与天气交流社区，获得云计算赛道第二名。

– 使用 Kafka 进行流数据处理, 使用 Spark 进行时序预测, 使用 Kubernetes 进行云资源运维与动态扩缩容。作为队长，主要负责整体系统架构设计，API 接口定义、调度算法、模块耦合与压力测试部分。

MiniOS Github Page

– 实现一个基于 RISC-V 64 的微型操作系统内核。从基础框架之上实现异常触发、线程调度、三级页表、ELF 程序加载、用户/内核态切换与文件系统嵌入 (VFS 及 FAT32)。

SysY Compiler ZJU Git

– 实现 SysY 语言全栈编译器，覆盖词法语法分析、抽象语法树生成、语义分析、代码生成。

– 成功将 SysY 源码转换为优化后的 RISC-V 32 汇编语言。

技术栈

工具使用: Git, Docker, Kubernetes, Vivado, Matlab

语言技能: Chinese(Native); English(TOEFL=R30 L30 S21 W24)

编程技能: C/C++, Python, Pytorch, Verilog, Risc-V, Cuda, OpenMP, MPI

个人荣誉

奖学金 2023, 2024, 2025

– 浙江大学三等奖学金

社会工作 2024, 2025

– 浙江大学学生吉他协会副会长

志愿工作 2023

– 第 19 届杭州亚运会、第 4 届杭州亚残运会主媒体中心公共工作间志愿者、负责人。

– 本科期间超 300 小时实地志愿服务时长。