林泽佳

linzj39@mail2.sysu.edu.cn zejia-lin.github.io 18022302700 Next-Ways

关键词：GPU，推理优化，编译优化，混合精度

- 熟悉计算机系统和体系结构，针对 GPU 的软硬件特性，从编译/运行时寻找优化机会，包括硬件资源共享、任务调度和混合精度计算。
- 微信核心部门实习一年以上，发表三篇 GPU 优化论文，熟悉 CUDA, LLVM, SGLang, veRL。

教育背景

中山大学, 广州（985, 双一流） *2022 年 9 月 – 2027 年 6 月 (预计)*
计算机科学与技术 博士研究生在读 导师：张献伟、卢宇彤

西北工业大学, 西安（985, 双一流） *2018 年 8 月 – 2022 年 6 月*
软件工程 本科 GPA 排名：22 / 259 优秀毕业生

研究成果

- **Zejia Lin**, Hongxin Xu, Guanyi Chen, Zhiguang Chen, Yutong Lu and Xianwei Zhang
[ASPLOS'26, CCF-A] Bullet: Boosting GPU Utilization for LLM Serving via Dynamic Spatial-Temporal Orchestration. *The 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2026.*

- **Zejia Lin**, Aoyuan Sun, Xianwei Zhang and Yutong Lu
[LCTES'24, CCF-B] MixPert: Optimizing Mixed-Precision Floating-Point Emulation on GPU Integer Tensor Cores. *The 25th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems, 2024.*

- **Zejia Lin**, Zewei Mo, Xuanteng Huang, Xianwei Zhang and Yutong Lu
[ICCD'23, CCF-B] KeSCo: Compiler-based Kernel Scheduling for Multi-task GPU Applications. *The IEEE 41th International Conference on Computer Design, 2023.*

- Wenxuan Pan, **Zejia Lin**, Jiangsu Du and Xianwei Zhang
[TACO, CCF-A] HuntKTm: Hybrid Scheduling and Automatic Management for Efficient Kernel Execution on Modern GPUs. *ACM Transactions on Architecture and Code Optimization, 2025.*

- Kan Wu, **Zejia Lin**, Mengyue Xi, Zhongchun Zheng, Wenxuan Pan, Xianwei Zhang and Yutong Lu
[DAC'25, CCF-A] GoPTX: Fine-grained GPU Kernel Fusion by PTX-level Instruction Weaving. *The 62nd ACM/IEEE Design Automation Conference, 2025.*

项目/实习/助教

腾讯, 深圳 *2024 年 12 月 – 现在*
微信事业群/搜索应用部, AI Infra 科研实习生
研究在 RAG 场景下提升 LLM 推理性能。提出通过 GPU 时空共享实现单卡 PD 分离的方法 Bullet 并成功落地，在镜像流量测试中验证延迟和吞吐均优于现有方法。该工作已被 ASPLOS'26 接收（接收率: 10.6%），并在 SGLang Beijing Meetup 中宣讲。目前正在研究强化学习训练的系统优化。

中山大学, 广州 *2023 年 3 月 – 2023 年 7 月*
编译原理课程助教
负责实验搭建和课堂教学，基于 LLVM 实现了一个 C 语言子集的编译器，并将任务拆解作为实验课作业。编写实验教程文档，在实验课上现场演示。该编译器后续改进为 SYsU-Lang 实验平台

(https://yatcc-ai.com/)，获得中国高校计算机教育大会 2024 年教学案例一等奖。

腾讯, 深圳 *2021 年 7 月 – 2021 年 9 月*

微信事业群/微信支付线, 后台开发实习生

负责研效提升工作，用 Django 搭建分布式的企业微信内部机器人，为开发人员自动回复与后台测试环境相关问题。接入腾讯内部 TAPD 和七彩石项目管理系统，完善问题流程追踪管理，自动更新缺陷库，用于持续训练轻量的 NLP 模型提升自动回复准确率。

专业技能

- 编程语言: C/C++, CUDA, Python
- 编译工具链：LLVM, CMake, Clang, GDB
- GPU 相关工具：NVIDIA nsys/ncu, CUTLASS, cuBLAS
- AI 生态：PyTorch, vLLM, SGLang, veRL
- 其它编程工具：Git, Docker, Vim, LaTeX
- CCF CSP 认证：Top 5.2%
- 英语六级：610 / 710

获奖情况

中山大学 一等奖学金 *2024, 2023, 2022 年*

中山大学计算机学院 腾讯奖学金 *2023 年*

西北工业大学 优秀毕业生 *2022 年*

西北工业大学 一等奖学金 *2021, 2020, 2019 年*

中国高校计算机大赛微信小程序应用开发赛 全国二等奖 *(5%, 参赛队长)* *2020 年 8 月*

专业服务

- ACM EuroSys 2026, Artifact Evaluation Committe
- ACM SOSP 2025, Artifact Evaluation Committe
- IEEE NAS 2024, Sub-reviewer
- IEEE ICPADS 2022, Sub-reviewer