林泽佳

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

关键词：**GPU**，**编译优化**，**推理优化**，**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。
- 独立完成和发表两篇 GPU 编译优化相关论文，熟悉 CUDA, LLVM, PyTorch, vLLM。

教育背景

中山大学, 国家超级计算广州中心, 广州 *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：卢宇彤、张献伟

西北工业大学, 西安 *2018 年 8 月 – 2022 年 6 月*
软件工程 本科 优秀毕业生

研究成果

- [LCTES'24] MixPert: Optimizing Mixed-Precision Floating-Point Emulation on GPU Integer Tensor Cores
  Zejia Lin, Aoyuan Sun, Xianwei Zhang and Yutong Lu. *The 25th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems, 2024.*
- [ICCD'23] KeSCo: Compiler-based Kernel Scheduling for Multi-task GPU Applications
  Zejia Lin, Zewei Mo, Xuanteng Huang, Xianwei Zhang and Yutong Lu. *The IEEE 41th International Conference on Computer Design, 2023.*
- [CF'22] moTuner: A Compiler-based Auto-tuning Approach for Mixed-precision Operators
  Zewei Mo, **Zejia Lin**, Xianwei Zhang, and Yutong Lu. *The 19th ACM International Conference on Computing Frontiers, 2022.*
- [CGO'25 under review] GoPTX: Fine-grained GPU Kernel Fusion by PTX-level Instruction Weaving
  Kan Wu, **Zejia Lin**, Mengyue Xi, Zhongchun Zheng, Wenxuan Pan, Xianwei Zhang and Yutong Lu. *IEEE/ACM International Symposium on Code Generation and Optimization, 2025.*

项目/实习/助教

细粒度资源共享的大模型推理优化 *2024 年 5 月 – 至今*
科研项目 导师：卢宇彤、张献伟
GPU 集群需要同时服务多个不同的大模型，i). 模型有不同的参数量，且热度动态变化，需要实时合理的并行划分与 GPU 资源分配；ii). 计算密集的 decode 阶段主导推理时间，使 GPU 利用率较低。我们允许多个模型共享同个 GPU，通过动态计算资源划分令 decode 与计算密集的 prefill 同时执行，并统一不同模型的 kv cache 管理与卸载策略，实现更高利用率的多大模型推理服务。

中山大学, 广州 *2023 年 3 月 – 2023 年 7 月*
编译原理课程助教 任课教师：张献伟
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，在实验课上现场教学。该编译器后续改进为 SYsU-Lang 实验框架，集成 CMake 和测评机方便学生进行实验构建与自动化测试，获得 2024 年教学案例一等奖。

腾讯, 深圳 *2021 年 7 月 – 2021 年 9 月*

业务连续性中心/研发效能组, 后端开发实习生 导师：张璇

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。设计和训练轻量的 NLP 模型，提高问题回复准确率。接入腾讯内部 TAPD 和七彩虹项目管理系统，完善问题流程追踪管理，自动更新语料库。

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
- 其它编程工具：Git, Docker, Vim, LaTeX
- CCF CSP 认证：Top 5.2%
- 英语六级：610 / 750

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 优秀本科毕业设计 *2022 年*

西北工业大学 一等奖学金 *2021, 2020, 2019 年*

美国大学生数学建模竞赛 M 奖 (6%) *2021 年 4 月*

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

- IEEE NAS'24, Sub-reviewer
- IEEE ICPADS'22, Sub-reviewer