# 齐呈祥

https://kuangjux.top/ · kuangjux@outlook.com · KuangjuX · KuangjuX

**技术方向：AI / 大模型基础设施、GPU Kernel、深度学习编译器、CUDA、Rust**

## 教育经历

**中国科学院大学 · 杭州高等研究院** 2023 年 9 月 - 2026 年 6 月
工学硕士 · 计算机技术 杭州 / 北京

**天津大学 · 智能与计算学部** 2019 年 9 月 - 2023 年 6 月
工学学士 · 计算机科学与技术 天津

## 工作与实习经历

**腾讯 · 微信事业群（WXG）** 2026 年 7 月 - 至今
WeLM 团队 · 机器学习系统工程师 北京

- **稀疏 Attention 训练优化：** 面向下一代大语言模型的长上下文训练，参与稀疏 Attention 计算模块的研发与性能优化，推进算子库建设、数值验证与性能评测，为后续训练系统集成提供支持。

**腾讯 · 微信事业群（WXG）** 2025 年 6 月 - 2026 年 6 月
WeLM 团队 · 机器学习系统实习生 北京

- **长上下文推理加速：**使用 CuTeDSL 实现 DuoAttention 并集成至 SGLang，在 16K 序列长度下取得 1.43 倍加速。
- **NVSHMEM：** 结合 DeepEP 调研 NVSHMEM，实现 NVSHMEM-Tutorial，涵盖基于 CUDA IPC/RDMA 的混合通信，并进行组内技术分享。
- **分布式 Attention：**基于 ThunderKittens 的 LCF 模板实现 Ring Attention 前向计算，在短序列场景下性能优于 ring-flash-attention；实现 Flash Attention 反向计算并向开源社区提交 PR（#134、#135）；对 MagiAttention、ZigZag Ring Attention 和 ZigZag Flex Attention 开展系统的性能分析。
- **GPU 编程工具调研：** 研究主流 DSL 在 NVIDIA Hopper 架构上的性能与兼容性。

**微软亚洲研究院（MSRA）** 2024 年 2 月 - 2025 年 5 月
系统与网络组 · 研究实习生；导师：曹莹 博士 北京

- 基于 FractalTensor 编程模型，使用 CUTLASS 实现并优化 **GEMM、Back-to-Back GEMMs、Stacked/Dilated LSTM、FlashAttention-2** 等算法。在 NVIDIA A100 上评测，相比 SOTA 实现最高取得 **5.45 倍加速**，平均加速 **2.14 倍**。
- 作为核心设计者与开发者，设计并实现高效的 **C++ 宏内核模板库 TileFusion**，提升 CUDA C 中 Tile 处理的抽象层次；结合 NVIDIA 硬件优化技术实现多种硬件感知算法，性能与 CUTLASS 相当。

**清华大学** 2023 年 5 月 - 2023 年 8 月
操作系统实验室 · 研究实习生；导师：贾越凯 博士 北京

- 使用 Apache HTTP Server、iperf 等工具开展网络性能基准测试，并开发专用工具评估网卡的 raw socket 发送与接收能力。修改网络协议栈及其与 Arceos 的接口以提升网络带宽；短消息延迟低于 Linux，长消息延迟较高，整体性能优于 Unikraft。
- 使用 **Rust** 实现 **Intel 82599 网卡驱动**，参考 DPDK 优化性能并以 crate 形式集成至 Arceos，在 AMD 平台上运行 httpserver、iperf、Redis 等应用。
- 基于 Arceos 开发可启动主线 Linux 的 **Type-2 Hypervisor**。

## 论文发表

- Siran Liu*, **Chengxiang Qi***, Ying Cao, Chao Yang, Weifang Hu, Xuanhua Shi, Fan Yang, Mao Yang. Uncovering Nested Data Parallelism and Data Reuse in DNN Computation with FractalTensor. ACM Symposium on Operating Systems Principles（SOSP 2024，*共同第一作者）

## 项目经历

**TileFusion：C++ 宏内核模板库（275 Stars）** 2024 年 5 月 - 至今

面向 CUDA C 中 Tile 处理抽象的实验性模板库。

- **架构设计：**作为核心设计者与开发者，从零构建模板库，采用自底向上的设计，以 BaseTile 为基础构建单元，提升 Tile 计算的抽象层次。
- **底层优化：**实现全局内存合并访问、Swizzle 等优化，封装 PTX 指令，无需依赖外部库。
- **算法实现：**支持便捷实现 FlashAttention、FlashDecoding 等硬件感知算法，取得与 CUTLASS 相当的性能。

**FractalTensor：深度神经网络数据组织与优化框架** 2024 年 1 月 - 2024 年 7 月

面向深度神经网络新型数据组织方式的优化框架。

- **计算实现：**基于 FractalTensor 的上层架构与编程模型，使用 CUTLASS 3.0（CuTe）实现并优化 GEMM、Back-to-Back GEMMs、RNN、FlashAttention-2 等模型与算法。
- **性能与论文：**相比 SOTA 实现平均取得 2.14 倍加速，项目论文发表于 SOSP 2024（CCF-A）。

**Unikernel 虚拟化支持与网络优化（653 Stars）** 2023 年 5 月 - 2023 年 8 月

围绕 Arceos 开展虚拟化支持、网卡驱动开发与网络性能优化。

- **虚拟化与中断：**将 hypercraft 集成至 Arceos，使其能够作为 Type-2 VMM 启动；为 Arceos 添加中断支持，实现基于 virtio-net、virtio-blk 的 I/O 中断。
- **驱动与网络：**开发 ixgbe 网卡驱动，同时在驱动层与网络协议栈层面开展性能优化。

**基于 RISC-V 的 Type-1 / Type-2 Hypervisor（155 Stars）** 2023 年 1 月 - 2023 年 5 月

使用 Rust 实现 hypocaust、hypocaust-2、hypercraft 三个 RISC-V Hypervisor。

- **hypocaust：**采用 S-mode 陷入模拟，实现影子页表与 PLIC 模拟，支持启动 rCore-Tutorial-v3。
- **hypocaust-2：**利用 RISC-V H 扩展实现硬件辅助虚拟化，支持两阶段页表映射与中断注入，可运行 rCore-Tutorial-v3、RT-Thread 和主线 Linux。
- **hypercraft：**参考 KVM 与 Zircon 的设计，将其扩展为可复用的操作系统组件（Type-2 Hypervisor），支持启动主线 Linux。

**使用 Rust 实现类 Unix 操作系统（341 Stars）** 2021 年 3 月 - 2022 年 1 月

使用 Rust 重新实现 MIT xv6-riscv。

- **内存与文件系统：**将原有内存分配器替换为伙伴分配器，重新设计并优化文件系统，性能优于官方 xv6-riscv。
- **同步机制：**将 SpinLock/SleepLock 重构为具有 RAII 特性的智能指针。