CHENGXIANG QI

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

Keywords: AI Infra, LLM Infra, DL Compiler, CUDA, Rust.

Education

University of Chinese Academy of Sciences, Hangzhou Institute for Advanced Study Sep 2023 -- Jun 2026 (Expected)
Master of Computer Technology Hangzhou | Beijing

College of Intelligence and Computing, Tianjin University Sep 2019 -- Jun 2023
Bachelor of Computer Science and Technology Tianjin

Professional Experience

WeChat, Tencent. Jun 2025 -- Present
Technical Architecture Team Beijing

- Long-context inference acceleration: Implemented DuoAttention with CuteDSL and integrated it into SGLang, achieving a 1.43x performance improvement for sequence length of 16K.
- NVSHMEM: Conducted research on NVSHMEM combined with DeepEP, and implemented NVSHMEM-Tutorial , including hybrid communication based on CUDA IPC/RDMA for internal team technical sharing.
- Distributed Attention: Implemented Ring Attention Forward with LCF template based on ThunderKittens, outperforming ring-flash-attention on short sequences; Implemented Flash Attention Backward with LCF and submitted PRs(#134, #135) to the open-source community; Conducted comprehensive performance analysis for MagiAttention, ZigZag Ring Attention and ZigZag Flex Attention.
- Other: Investigated the performance and compatibility of mainstream DSLs on NVIDIA Hopper architecture.

Microsoft Research Asia (MSRA) Feb 2024 -- May 2025
System Research Group, Mentor: Ying Cao Beijing

- Based on the FractalTensor programming model, implemented and optimized algorithms including **GEMM, Back-to-Back GEMMs, Stacked/Dilated LSTM, FlashAttention-2** with CUTLASS. Conducted performance evaluation on NVIDIA A100, achieving up to 5.45x speedup and an average 2.14x speedup compared with SOTA implementations.
- As the core designer and developer, designed and implemented TileFusion, an efficient **C++ macro kernel** template library to elevate the abstraction level of tile processing in CUDA C. Implemented various hardware-aware algorithms with NVIDIA hardware optimization techniques, and TileFusion currently achieves comparable performance with CUTLASS.

Tsinghua University May 2023 -- Aug 2023
Operating System Laboratory, Mentor: Yuekai Jia Beijing

- Conducted network performance benchmarking with Apache Http Server, iperf and other tools, and developed a benchmark tool to evaluate the raw socket transmission and reception capabilities of network interface cards (NICs). Modified the network protocol stack and its interface with Arceos to enhance network bandwidth. Achieved lower latency than Linux for short messages (higher for long messages), and outperformed Unikraft in overall performance metrics.
- Implemented the Intel 82599 NIC driver in Rust, optimized performance with DPDK reference, and integrated it into Arceos as a crate. Successfully deployed real-world applications (httpserver, iperf, Redis) on AMD platforms.
- Developed a Type-2 Hypervisor based on Arceos that is capable of booting mainline Linux.

Publications

- 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. In Proceedings of the 32nd ACM Symposium on Operating Systems Principles(SOSP 2024) (* Equal Contribution)

Projects

TileFusion: An Experimental C++ Macro Kernel Template Library for Elevating Tile Processing Abstraction in CUDA C (★ 275 Stars) May 2024 -- Present

· Core designer and developer, built the library from scratch with BaseTile as the fundamental building block (bottom-up design). Implemented optimizations including **global memory coalescing** and **Swizzle**, encapsulated PTX instructions without external library dependencies; Enabled easy implementation of various hardware-aware algorithms (e.g., **FlashAttention/FlashDecoding**) with TileFusion, achieving performance comparable to CUTLASS.

FractalTensor: An Optimization Framework for Novel Data Organization in Deep Neural Networks Jan 2024 -- Jul 2024

· Based on the FractalTensor upper-layer architecture and programming model, implemented and optimized multiple models and algorithms (GEMM, Back-to-Back GEMMs, RNN, FlashAttention-2) with **CUTLASS 3.0 (CuTe)**, achieving an average 2.14x speedup over SOTA. The project's paper was accepted by SOSP 2024 (CCF-A).

Unikernel Virtualization Support, Network Driver Development & Network Performance Optimization (★ 653 Stars) May 2023 -- Aug 2023

· Integrated **hypercraft** into Arceos to enable booting as a **Type-2 VMM**; Added interrupt support to Arceos and implemented IO interrupts based on **virtio-net** and **virtio-blk**; Developed ixgbe NIC driver for Arceos with performance optimizations at both driver and network stack layers.

Type-1 & Type-2 Hypervisor Implementation Based on RISC-V (★ 155 Stars) Jan 2023 -- May 2023

· Implemented three RISC-V based hypervisors (hypocaust, hypocaust-2, hypercraft) in Rust: hypocaust adopted **S-mode trap emulation** with shadow page table and PLIC emulation, supporting rCore-Tutorial-v3 boot; hypocaust-2 leveraged **H-extension** with two-stage page table mapping and interrupt injection, compatible with rCore-Tutorial-v3, RT-Thread and mainline Linux; hypercraft was extended as a reusable OS component (Type-2 hypervisor) referring to KVM and Zicron designs, supporting mainline Linux boot.

Unix-like Operating System Implemented in Rust (★ 341 Stars) Mar 2021 -- Jan 2022

· Reimplemented MIT xv6-riscv in Rust; Replaced the original memory allocator with a buddy memory allocator and redesigned/optimized the file system for superior performance over the official xv6-riscv; Redesigned SpinLock/SleepLock as smart pointers with RAII features.