Yangjia Hu

University of Science and Technology of China, 96 Jinzhai Road - 230026 Hefei - P.R.China
(+86) 13688355505 | huyangjia@mail.ustc.edu.cn |

EDUCATION

University of Science and Technology of China — HeFei, AnHui, China
B.Eng. in Computer Science and Technology (Senior) — Sep. 2022 – Present

- GPA: 3.91/4.3
- Arithmetic mean score: 90.62/100
- Rank: 6/167
- English: IELTS 6.5 (L6.0/R7.0/W7.0/S6.0)

RESEARCH EXPERIENCE

Optimization of FPGA-Based Control and Data Transmission for Piezoresistive Array Devices

Supervisor: Prof. Xiaohui Cai, USTC — Nov 2023 - Dec 2024

- Description: Optimized FPGA-based control systems and data transmission for piezoresistive array devices, focusing on improving data throughput, reducing latency, and enhancing real-time signal processing capabilities.
- Contribution: Analyzed system performance to identify bottlenecks, optimizing resource utilization and efficiency. Including increasing data acquisition frequency from 40Hz to 80Hz and implementing a command frame control system.
- Tech stack: SystemVerilog

8th National College Students Computer System Ability Competition, CPU Track

May 2024 - August 2024

- Description: Designed and implemented a in-order dual-issue 8-stage pipeline CPU, utilizing the AXI-4 bus protocol. Integrated components included ICache, DCache, a pre-decoder, and a branch predictor.
- Contribution: Led the implementation of CPU architecture and performance optimization for stages prior to the Execute stage, focusing on critical components like FIFO and the pre-decoder. Contributed to the overall CPU design and verified functionality using SystemVerilog.
- Tech stack: SystemVerilog

ZipperQuant: Bit-based 4-bit Quantization for Heterogeneous LLM Inference

Supervisor: Prof. Song Guo — July 2025 - September 2025

- Description: Researched AI systems and LLM quantization, focusing on accelerating heterogeneous GPU–CPU inference. Proposed ZipperQuant, a bit-sliced 4-bit quantization framework for efficient W4A4 LLM serving. Paper under review at ICLR 2026.
- Contribution: Developed the bit-level inlier–outlier separation algorithm and a LUT-based CPU execution engine to handle sparse high-order bits efficiently. Co-designed the hybrid GPU–CPU pipeline, enabling parallel INT4 GEMM on GPU and FP16 LUT aggregation on CPU. Achieved up to 3.01× inference speedup and 3.9× memory reduction on LLaMA3-8B and Qwen models.
- Outcome: Co-first author the paper "ZipperQuant: Bit-Based Inlier–Outlier Disaggregation for 4-Bit LLMs on GPU–CPU", under review at ICLR 2026 (OpenReview).
- Tech stack: PyTorch, CUDA, Python, GPU and CPU kernel optimization

ADDITIONAL EXPERIENCE

Rewriting Linux Kernel bpf_trace Module with Rust

Apr 2024 - Jul 2024

- Led the translation and debugging of the Linux kernel's bpf_trace module from C to Rust, enhancing safety and performance by refactoring data structures and optimizing code, leveraging Rust and Make.

Implemented a 2-issue out-of-order cpu(ongoing)

Supervisor: Prof. Weng, KAUST — Dec 2024 - Present

- Utilized a newly developed language from the lab to design and implement a high-performance CPU, exploring the potential and practical application value of this language in hardware design.

SKILLS

- Programming Languages: C/C++, Verilog, SystemVerilog, Python, Java
- Tools & Frameworks: Git, Maven, Make, Vivado, PyTorch, CUDA
- Hardware/Systems: CPU microarchitecture, RTL design, FPGA prototyping

Awards and Honors

Excellent Student Scholarship Silver Award	2022
National Encouragement Scholarship	2023
3rd Prize in the National Finals of the 8th National College Students Computer System Ability Competition (CPU Track) - Team Award	2024
National Scholarship(TOP 2%)	2024
National Scholarship(TOP 2%)	2025
Guo Moruo Scholarship(TOP 1%)	2025