Zhanqiu(Summer) Hu

607-216-7811 | zh338@cornell.edu | New York, NY | zhanqiuhu.github.io | She/Her/Hers

Education

Cornell University, MS in Computer Engineering Aug 2023 - Dec 2025
Cornell University, BS in Computer Science & Electrical and Computer Engineering Aug 2019 - May 2023
GPA: 4.05/4.0; *summa cum laude*

Experience

Cornell Tech Graduate Research Assistant | New York, NY Aug 2023 – Present

Test-Time Optimization for Diffusion LLMs

- Achieved up to **34**× inference speedup by developing *FreeCache* (KV cache optimization) and *Guided Diffusion* (cross-model token unmasking), integrated with **HuggingFace**.
- Designed and benchmarked **test-time optimizations** across reasoning and QA benchmarks (e.g., MMLU, GSM8K, PIQA), enabling scalable long-context (>1024 tokens) generation.
- First author on *Accelerating Diffusion Language Model Inference via Efficient KV Caching and Guided Diffusion* (arXiv:2505.21467), demonstrating leadership and collaboration in **efficient LLM inference** research.

Distributed Framework for Recommendation Retrieval and Ranking

- Designed and deployed a production-style framework supporting training and serving of **recommendation models** (two-tower, transformer, and DLRM) using **TensorFlow**, with serving enabled through a microservices architecture built on **gRPC**, **Ray Serve**, and **NVIDIA Triton Server**.
- Implemented a scalable **data pipeline** with feature engineering in **NVTabular**, fast data access through **Redis**, and integrated vector search indexes with **FAISS**.
- Instrumented the design space exploration framework with profiling using **NVIDIA Nsight Systems** and **Intel VTune**, driving **system-level optimizations** that reduced latency by 40% and doubled throughput.

Amazon Web Services (AWS) Software Development Engineer Intern | Boston, MA May 2022 – Aug 2022

- Revamped AWS FSx frontend codebase with **TypeScript/React**; Upgraded **UI** libraries to align with AWS design standards; Implemented scalable state management and async workflows with **Redux Toolkit** and **Redux Observable**.
- Improved **user experience** and code quality by leveraging strong typing, reducing bugs, and eliminating duplication.
- Integrated **automated testing into CI/CD pipelines** and delivered migration tools adopted by 5+ teams.

Skills

ML Frameworks: PyTorch, TensorFlow, HuggingFace, vLLM, llama.cpp, verl (RLHF)

Systems & Infra: AWS, NVIDIA Triton, Ray Serve, FAISS, gRPC, TVM, Redis, MySQL, Docker, Kubernetes, Linux, Slurm

Programming: Python, C/C++, Java, JavaScript, OCaml, SystemVerilog, Bash, Shell, LaTeX

Expertise: AI/ML, ML Systems and Infra, Distributed Systems, Computer Architecture, Compilers, Operating Systems

Projects

2D Convolution Kernel Optimization Individual Contributor Fall 2023

- Implemented optimized 2D convolution kernels on FPGA (**Verilog**), CPU (**C++/OpenMP**), and GPU (**CUDA**).
- Achieved 12× latency reduction and 10× throughput gain on FPGA compared to baseline.

Custom Compiler Targeting x86-64 Assembly Team Lead Spring 2023

- Led a team of 3 to build a compiler in **Java** targeting **x86-64 assembly**, implementing 12.5K lines of code from scratch.
- Implemented end-to-end **compiler workflow** from parsing to assembly code generation, validating across hundreds of unit and integration tests.
- Implemented **compiler optimizations** (e.g., register allocation, copy propagation) to improve code size and performance.

MLSys Teaching Frameworks (Cornell ECE 5545) Technical Lead June 2022 – Jan 2023

- Led development of user-friendly **PyTorch** frameworks and tools for speech recognition **model training**, **fine-tuning**, **quantization**, and **deployment**, including export to **ONNX** and **TensorFlow Lite (TFLite)**.
- Wrote tutorials and assignments on the efficient kernel programming with **Apache TVM**; mentored students on optimizing tensor operations across CPU, GPU, and FPGA platforms.
- Directed deployment of a **TinyML Keyword Spotting** application on embedded devices (**Arduino** Nano 33 BLE).

Quantization Frameworks for Efficient Deep Learning Research Lead July 2020 – Sep 2022

- Built *QFX*, a PyTorch toolkit that emulates variable-length fixed-point arithmetic using **floating-point operations**, enabling differentiable **quantization-aware training**.
- Developed *OverQ*, an algorithm that improved post-training quantization accuracy by 5% Top-1 gain on ResNet50.

RISC-V Multicore Processor in Verilog Team Lead Fall 2021

- Implemented a **quad-core RISC-V processor** in **Verilog** with pipelined execution, bypassing, and a variable-latency **ALU** with iterative multiplier.
- Designed a **memory subsystem and cache hierarchy** (two-way set-associative) and benchmarked the multicore system using PyMTL on parallel C workloads to validate performance scaling.

Publications

Accelerating Diffusion LLM Inference via KV Caching and Guided Diffusion preprint
Zhanqiu Hu, Jian Meng, Yash Akhauri, Mohamed S. Abdelfattah, Jae-sun Seo, Zhiru Zhang, Udit Gupta

A Full-Stack HW/SW Co-Design Analysis for Recommendation System Inference Submitted to IEEE Micro
Zhanqiu Hu, Mark Zhao, Zhiru Zhang, Udit Gupta

OverQ: Opportunistic Outlier Quantization for Neural Network Accelerators preprint
Ritchie Zhao, Jordan Dotzel, **Zhanqiu Hu**, Preslav Ivanov, Christopher De Sa, Zhiru Zhang