Yian Wang

(+1) 217-721-2793 | yianw11@uci.edu | canying.me

EDUCATION

University of California, Irvine, Master of Science in Computer Engineering — GPA: 3.95/4.0 — June 2025

University of Illinois, Urbana-Champaign, Bachelor of Science in Computer Engineering — GPA: 3.50/4.0 — May 2023

Courses: Computer Architecture, Distributed System, Heterogeneous System, Memory & Storage System, Parallel Computing, HPC, Analog & Digital Circuits, Artificial Intelligence, Computer Vision, Machine Learning

Skills: Verilog/SystemVerilog, UVM, Vitis HLS, Intel Altera Quartus, Xilinx, Vivado, Synopsys VCS, Design Compiler; Python, C/C++, System C, Java, Vue.js, MATLAB, Cmake, Git, Linux Bash Shell, Docker, OpenCV, Pytorch, CUDA, ROCm/HIP, OpenCL & OpenGL, numba

PUBLICATIONS

APQ: Toward Arbitrary-dimensional Quantization of Large Language Models with Extended Product Quantization (To be submitted) — June 2024 - May 2025

- Current model quantization techniques primarily focus on data format precision quantization. This paper explores high-level quantization instead, offering fine-grained model size over traditional quantization.
- Wrote the entire codebase in PyTorch with full reproducibility. To demonstrate that my method works with other works, it is designed in a way such that it can work seamlessly with the implementation of other works. Added support for parallel quantization. Designed a custom ASIC to accelerate inference. Preliminary result shows an average of 4×computation reduction compared to APQ with GPUs.

FlexiBit: Fully Flexible Precision Bit-parallel Accelerator Architecture for Arbitrary Mixed Precision AI (To be submitted) — Sep 2024 - April 2025

- Current hardware fails to support non-power-of-two precisions such as FP6 efficiently, due to padding solutions in modern GPUs and accelerators. The paper proposes a novel ASIC architecture that computes flexible precision floating-point matrix multiplications with near full utilization.
- Write the RTL of the hardware design in SystemVerilog and perform thorough testing via UVM. Utilized several DRAM simulators such as Ramulator2 to get an estimate for our DRAM model. Implemented several works in the literature as our baselines. Evaluated them with the same memory model to ensure a fair comparison. Result shows FlexiBit achieves an average of 1.64× higher performance per area compared to tensor core and bit-parallel accelerators.

COD-BT: Co-Optimized Binary Transformer Accelerator for Edge FPGA (To be submitted) — Aug 2024 - April 2025

- Binary transformers are desired for edge devices. However, current binary transformers are still involved in complex floating point operations.
- Designed a binary kernel with thresholds to replace softmax operations. Explored different granularities of the threshold. Used a greedy search algorithm to search for the optimal thresholds without training. With further hardware-friendly optimizations in the attention block, COD-BT achieves up to 3894.7 GOPS throughput and 448.7 GOPS/W energy efficiency on edge FPGAs, delivering a 311× energy efficiency improvement over GPUs and a 3.5× throughput improvement over the state-of-the-art binary accelerator, with only negligible accuracy degradation.

WORK EXPERIENCE

Max-Optics Information Technology Ltd., Accelerated Computing Intern | Shanghai, China — May 2024 - Aug 2024

- Developed parallel computing programs using various frameworks (e.g. CUDA, ROCm, OpenCL, HIP, etc.) due to the outdated CUDA versions of previous programs.
- Profile existing solutions using Nsight Compute and analyze their bottlenecks. Study textures and use NCCL in multi-GPU program for scalability.
- Accelerated current single-GPU solution by ~31%, the multi-GPU version by ~19%.

University of Illinois, Urbana-Champaign Course Assistant | Champaign, IL, USA — Aug 2022 - May 2023

- Improved the legacy class final project of designing a RISC-V 5-stage pipelined & Out-of-Order processor. Provide options to students where they can skip final exam if their processors can boot up a minimized Linux on provided FPGA board.
- We migrated from FPGA flow to ASIC flow, which improves average clock per instructions(CPI). On software side, we switched from Intel Altera Quartus to Synopsys VCS.
- Designed autograder using FastAPI and shell scripts to host on department server. It relies on git to fetch students' assignments on schedule, run designated tests, and generate report.

PROJECTS

Toolbox for calcium imaging analysis, Research Intern | *Columbia University* — Mat 2022 - Sep 2023

- I developed a graphical application in Python, aiming to support the calcium signal analysis for researchers with no programming experience in Columbia University Department of Psychiatry.
- I added machine learning techniques to help filter noisy signals, which brings 32× speedup and 1.2× accuracy. In addition, I packaged the project to be a installable application on Windows, Mac, and Linux.

FPGA game in SystemVerilog: Plant v.s. Zombie, ECE385 Final Project, *UIUC* — Aug 2020 - Dec 2020

- I developed an interactive, console-style plant v.s. zombie game on Intel DE-10 Lite FPGA board with MAX3421E to work with keyboard inputs. I designed core game logic in combination of SystemVerilog and C, including interactions between plants and zombies, and the sunlight-planting mechanism.
- Support VGA graphic display and significantly reduced FPGA resource bottleneck by using compressed graphic representations.