Mingi Kwon

linkedin.com/in/mingikwon00 | mingi001025.github.io | mingi001025@gmail.com

Education

University of California, San Diego — *La Jolla, CA*
MS in ECE, Computer Engineering (GPA: 3.814/4.0) — *Sep. 2024 – Mar. 2026*

Hanyang University — *Seoul, Republic of Korea*
BS in Electrical Engineering (GPA: 3.97/4.5) — *Mar. 2019 – Aug. 2024*

Courses

Relevant Courses: Computer Architecture, Low Power VLSI Implementation for ML, VLSI Algorithm & Architecture, SoC Architecture, Optimization and Acceleration of Deep Learning

Work Experience

Research Assistant, STABLE Lab (Advisor: Prof. Jishen Zhao) — *Mar. 2025 - Present*

- Analyzed AlphaFold3's inference pipeline (Tensorflow) using `JAX Profiler`, `Perf`, `uProf`, and `Nsight` to characterize CPU-level bottlenecks including cache contention, thread scalability limits, and memory hierarchy inefficiencies; co-developed `AFSysBench` to support reproducible system-level benchmarking across diverse platforms. Co-authored a paper on this work, accepted at `IISWC` 2025.
- Characterized multi-agent LLM serving workloads using `vLLM` and `SGLang`, benchmarking GPT-OSS-120B under **MT-Bench** multi-turn dialogues. Analyzed prefill–decode disaggregation, pipeline, and **tensor parallelism** across 2×H100 GPUs—revealing decode-phase and KV-cache transfer bottlenecks that limited throughput scaling (~1.8–2.0 Req/s, 95% decode utilization).

CPU Microarchitecture Intern, Samsung SAIT (Corporate R&D Division) — *Jun. 2025 – Aug. 2025*

- Optimized the frontend pipeline of a high-performance Out-of-Order **RISC-V CPU**, including **I-Cache** redesign, variable shifter, and priority encoder improvements—synthesized with `Synopsys RTL Architect`, achieving a **25.03%** area reduction and **98.91%** reduction in Total Negative Slack (TNS) in the I-Cache @2.5GHz.
- Refactored all **I-Cache** logic along the `paddr` path to relocate PLRU eviction from IF0 to IF1, removing redundant computations and shortening IF0's critical path; enabled use of single-ported PLRU memory by eliminating concurrent access requirements.
- Redesigned the **variable shifter** as a configurable-width, bidirectional barrel shifter and implemented a tree-based **priority encoder**, reducing total area by **25.1%** and TNS by **29.9%**; verified functionality using `VCS` and `Verdi`.
- Conducted performance modeling and architectural simulation in `gem5` to evaluate custom **TAGE** predictors for RISC-V CPU integration; proposed a 9-table, ~12KB configuration that matched 8KB TAGE-SC-L performance with significantly lower implementation complexity and better area-efficiency.

Researcher, Qualcomm Institute — *Jan. 2025 - Feb. 2025*

- Led instruction for undergraduate AI/ML research, guiding students through implementation of MLPs, quantization, and pruning using KNIME; mentored three teams that successfully published papers based on their research projects.

Project Experience

RTL2GDS For LLM Attention HW Accelerator — *Jan. 2025 - Mar. 2025*

- RTL design, verification, logic synthesis, and place-and-route (PnR) to generate gate-level netlist and layout via Synopsis Design Compiler (DC) and Cadence Innovus in TSMC 65nm technology.
- Implemented clock gating and pipelining on MAC unit and SFP to optimize power efficiency and meet timing constraints.
- Digital design fundamentals of clock domain crossing such as multi-cycle path design and managing asynchronous interface with synchronizer, hand-shaking protocol, and FIFO.

Reconfigurable 2D Systolic Array for Al Acceleration — *Sep. 2024 - Dec. 2024*

- Developed a reconfigurable 8×8 weight- and output-stationary 2D systolic MAC array in Verilog to enable parallel computations for convolution layers for deep learning networks, verified using real parameters from VGG16 and ResNet20.
- Designed zero-detection logic for zero-skipping to reduce latch toggling, allowing low-power computation in sparse deep neural network following pruning techniques.
- Optimized parallel read and write operations across stages (SRAM, FIFO, MAC array) to minimize chip area, reducing FIFO depth, optimized power consumption(0.52x) and frequency(1.08x).

Publication

Kim, J.; **Kwon, M.**; Zhao, J. "**AlphaFold3 Workload Characterization: A Comprehensive Analysis of Bottlenecks and Performance Scaling**" Accepted at *IEEE International Symposium on Workload Characterization (IISWC)*, 2025.

Shin, H.; **Kwon, M.**; Lee, Y.; Kim, Y.; Cho, M.-K.; Song, I. "**Circuit-Centric Genetic Algorithm for Optimization of a Radio-Frequency Receiver**" Electronics 2025, 14, 770. **Best Researcher Award**

Technical Skills

Fluent Programming: VHDL, Verilog, SystemVerilog, Python, Pytorch, C/C++, Shell Script, TCL, Linux

Tools: Synopsys DC, VCS, Verdi, RTL Architect, Gem5, Cadence Virtuoso, Innovus, Intel Quartus, Modelsim, Xilinx Vivado, Perf, uProf, Nsight