HAIYUE MA

Princeton, NJ • +1 (609) 453-8654 • hm1@princeton.edu • https://haiyuem.github.io/

SUMMARY

Ph.D. student in Computer Architecture seeking internship in summer 2024. My research interests are **hardware/software co-design and systems architecture**. I am particularly interested in exploring new forms of architecture for emerging applications.

SKILLS & EXPERTISE

Areas: System design for ML models, performance analysis, RTL-level power analysis

Tools: Architecture simulators (ChampSim, Gem5, Nvidia internal simulators) and profilers (Intel PIN, Intel VTune, Nsight compute/system)

Languages and frameworks: C/C++, CUDA, Python, Verilog

EDUCATION

Princeton University

Ph.D., Electrical and Computer Engineering — 2027 (Expected)

M.A., Electrical and Computer Engineering — 2023

Advisor: Prof. David Wentzlaff

Washington University in St. Louis

B.S., Electrical Engineering — 2018

Minor: Computer Science Honors: *magna cum laude*

RESEARCH EXPERIENCE

Princeton University — **2022 - Present**

Architecture-level AI safety precautions for ML model inference — 2023 - Present

- Use Performance Counters to monitor the model's hardware resource usage and detect misaligned behavior
- Design system response to misalignment: an integrated "kill-switch" feature in FP units at the microarchitecture level, and a cut-off strategy for global communication between compute units at the system level

Data value commonality profiling and value-aware hardware scheduler — 2022 - 2023

- Proposed an instruction-level profiler for detecting data value similarities, and an efficient form to condense value similarity hints in hardware for architectural exploitation
- Proposed a hardware scheduler for reordering instructions with similar values to execute consecutively to save dynamic energy usage

- Applied value prediction to domain-specific, data-intensive applications with high degree of data commonality

PUBLICATIONS & WORKSHOPS

Haiyue Ma and David Wentzlaff. *"Exploiting Data Commonality in Value Prediction".* The Fifth Young Architect Workshop, ASPLOS 2023.

Papers In Submission:

Haiyue Ma, and David Wentzlaff. *"A Hardware Kill-Switch: AI Abnormality Detection and Response".* In Submission to Young Architect Workshop in ASPLOS 2024.

Haiyue Ma, Kaifeng Xu, and David Wentzlaff. *"VASER: Value-Aware Scheduler for Energy Reduction".* In Submission to ISCA 2024.

Kaifeng Xu, **Haiyue Ma**, and David Wentzlaff. *"Preps: Temporal Branch Predictor Prefetching for Serverless Invocations".* In Submission to ASPLOS 2024.

WORK EXPERIENCE

NVIDIA **2018 - 2022**

Deep Learning Performance Architect 2020 - 2022

- Conducted GPU architectural exploration for DL workloads: built the prototype for a hardware synchronization primitive from CUDA programming to hardware design, and evaluated end-to-end performance

  Haiyue Ma, et. al. *"Hardware Accelerated Global Synchronization".* (Internal Conference)
- Optimized end-to-end DL inference performance with software pipelining and kernel fusion
- Explored sub-kernel level L2 prefetch strategies for neural network applications

ASIC Power Engineer 2019 – 2020

Intern, ASIC Power Engineer 2018

- Delivered full-chip GPU power analysis for the Ampere and Orin architectures by replaying RTL stimulus for representative workloads on Synthesis and P&R netlists
- Designed and implemented power flow infrastructure improvements for speed and accuracy
- Resolved tool issues including unmapped pins, mismatched switching activities, stimulus replay errors and clock gating inefficiencies

Samsung Semiconductor

Intern, GPU Power-Performance-Area Methodology Lab 2017

- Co-developed an early-stage power modeling tool by fitting a regression curve on architectural event counters and RTL power numbers collected from the existing design, and predicting power consumption for future architectures based on available event counters

TEACHING EXPERIENCE

Princeton University
Teaching Assistant
Computer Architecture — Spring 2023

Washington University in St. Louis
Teaching Assistant
Computer Architecture — Fall 2018
Introduction to Digital Logic and Computer Design — Fall 2017
Data Structures and Algorithms — Spring 2017
Introduction to Electrical and Electronic Circuits — Fall 2016
Computer Science I — Spring 2016

HONORS & AWARDS

Non-Degree Graduate Student, Stanford Center for Professional Development — 2020 - 2021
Took graduate-level Stanford classes for credits while working full-time
Courses: Parallel Computing / Computer Systems Architecture / Advanced Topics in OS
GPA: 3.91/4.00

HONORS & AWARDS

Princeton First-Year Fellowship — 2021
Dean's List (All Semesters) — 2015 – 2018

SERVICE & VOLUNTEERING

Co-Chair, Computer Architecture Long-term Mentorship Program — 2023 – Present
Steering Committee Member, Computer Architecture Student Association — 2022 – Present
Social Chair, Computer Architecture Student Association — 2023