# 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** for emerging applications and **systems architecture**. I have strong expertise in workload profiling, performance analysis and prototype development.

## Education

***Ph.D. in Electrical and Computer Engineering* | Princeton University | 2027 (Expected)**

- Advisor: Prof. David Wentzlaff

***M.A. in Electrical and Computer Engineering* | Princeton University | 2023**

***B.S. in Electrical Engineering* | Washington University in St. Louis | 2018**

- Minor: Computer Science    Honors: *magna cum laude*

## Research Experience

*Work In Progress*: System-level AI safety precautions for transformer-based models

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

*Work In Submission (ISCA 2024)*: Data value aware instruction re-scheduling

- Created a trace-based profiler to detect instruction-level data value similarities, and generate hints in an architecturally efficient form for profile-guided optimization on hardware
- Designed a hardware scheduler that reorders instructions with similar values for subsequent execution to reduce toggling, and optimally balances between performance and energy consumption

Workshop paper presented at ASPLOS 2023: *Exploiting Data Commonality in Value Prediction*

## Work Experience

**Deep Learning Performance Architect | NVIDIA | 2020 - 2022**

- Conducted architectural exploration of GPU hardware synchronization primitives for DL workloads, built the prototype from CUDA programming model to hardware design, and evaluated cycle-accurate performance

  Paper presented at company's internal conference: *Hardware Accelerated Global Synchronization (2021)*
- Explored software pipelining and op-fusion for Convolution and Batch Norm layers with analytical models
- Explored kernel-level L2 prefetch strategies with lower DRAM bandwidth for GEMM and CNN workloads

**ASIC Power Engineer | NVIDIA | 2018 - 2020**

- Delivered full-chip GPU power analysis from replaying RTL stimulus for representative workloads on Synthesis and P&R netlists, and improved power modeling infrastructure