Haiyue Ma

Princeton, NJ | +1 (609) 453-8654 | hm1@princeton.edu

Summary

Third-year Ph.D. student in Computer Architecture seeking research internships in summer 2024. My research interests are hardware/software co-design for parallel computing and data movement patterns. I have strong expertise in workload profiling, performance analysis and prototype development.

Education

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

- PI: 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

Research Focus:

- Architecture-level data value profiling and value locality exploitation
- Hardware scheduling policy design for performance and energy optimization
- Low-power architecture design for edge computing

Contributions:

- Developed a trace-based profiler that detects value similarities among input variants across program runs, and investigated new opportunities for architectural optimizations made possible by such profiling
- Designed a value-aware hardware scheduling policy for energy reduction, and conducted design space explorations for scheduler configurations to optimally balance performance and energy efficiency

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

Work Experience

Compute Architect | NVIDIA | 2020 - 2021

- 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

  Poster 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
- Designed and implemented power modeling methodology to improve both speed and accuracy