HAIYUE MA

hm1@princeton.edu · (609) 453-8654 · haiyuem.github.io

SUMMARY

Ph.D. candidate in Computer Architecture with research interests spanning GPU systems architecture, hardware/software co-design, and microarchitecture. My recent work focuses on GPU bottleneck modeling, scheduling, and performance analysis for next-generation machine learning workloads.

EDUCATION

Princeton University, Princeton, NJ
Ph.D. in Electrical & Computer Engineering — *2026 (Expected)*
M.A. in Electrical & Computer Engineering — *2023*
Advisor: Prof. David Wentzlaff

Washington University in St. Louis, St. Louis, MO — *2018*
B.S. in Electrical Engineering, magna cum laude
Minor: Computer Science

RESEARCH & INTERNSHIP EXPERIENCE

NVIDIA Corporation — US, Remote

Deep Learning Architecture Intern — *Summer 2025*
Manager: Ronny Krashinsky, Mentor: Chris Mei

- Develop performance models to quantify the impact of load imbalance in Mixture-of-Experts (MoE) inference on large-scale GPU systems, and derive insights for mapping strategies and workload configurations in compute- and memory-bound scenarios.
- Evaluate load balancing strategies such as expert placement and duplication and quantify the performance impact.

Paper in Preparation (Collaboration between Architecture Research Group and Deep Learning Architecture Team): *Performance Impact of MoE Load Imbalance in Large-Scale Inference Decode*.

Deep Learning Architecture Intern — *Summer 2024*
Manager: Ronny Krashinsky, Mentor: Timmy Liu

- Studied kernel overlapping to accelerate LLM training by overlapping random number generation (RNG) for dropout with GEMMs based on their distinct fine-grained hardware bottlenecks. Built an analytical performance model, cross-validated with silicon results, to guide efficient scheduling and GPU-level resource utilization.

Princeton University, Princeton Parallel Group

Ph.D. Candidate — Princeton, NJ
Advisor: Prof. David Wentzlaff — *Jan. 2022–Present*

LLM Performance Analysis and Architecture Design — *2024–Present*

- Developed an analytical framework to find optimal strategies to reduce load imbalance and maximize throughput in Mixture-of-Experts (MoE) inference.
- Proposed systematic techniques for controllable, hardware-based performance throttling of AI workloads, which involves extending roofline-model for workload characterization and cost-efficient manufacturing. The proposed hardware supports use cases in AI safety, export control, and product segmentation.

Data-Aware Instruction-Level Scheduling — *2022–2023*

- Designed microarchitecture-level techniques to detect, profile, and exploit instruction operand value similarities in modern processors, enabling value-aware scheduling, dynamic profiling of data locality, and optimized value prediction for energy-efficient and high-performance computation.

PUBLICATIONS, PREPRINTS AND POSTERS

Haiyue Ma*, August Ning, and David Wentzlaff. *Differential Architecture: Limiting Performance of Targeted Applications.* Submitted to HPCA 2026, preprint available upon request.
Proposes Differential Architecture, a methodology to build hardware that intentionally bounds AI workload performance while preserving efficiency for other general applications, validated with an extended roofline model and GPU simulation. It provides practical design and production strategies for regulatory and safety-driven chip differentiation.

Haiyue Ma*, Zhixu Du, and Yiran Chen. *MoE-GPS: Guidlines for Prediction Strategy for Dynamic Expert Duplication in MoE Load Balancing.* arXiv:2506.07366, 2025. Under submission.
Develops a framework that models and selects optimal prediction strategies for dynamic expert duplication in Mixture-of-Experts (MoE) inference, improving end-to-end system throughput by balancing prediction accuracy and overhead across diverse hardware and workloads.

Haiyue Ma*, Timmy Liu, and Ronny Krashinsky. *Reducing the Cost of Dropout in Flash-Attention by Hiding RNG with GEMM.* arXiv:2410.07531, 2024 (work conducted at NVIDIA). Under submission.
Analyzes key architecture bottlenecks of random number generation (RNG) in Dropout layers compared to other operators in LLM training. Proposes RNG and GEMM overlapping to maximize GPU resource utilization. Proposes an analytical model based on bottleneck analysis, backed up by silicon results.

Haiyue Ma*, Kaifeng Xu, and David Wentzlaff. *VASER: Value-Aware Scheduler for Energy Reduction.* Poster, PACT 2025. To appear.
Introduces a microarchitecture level, value-aware scheduler that reduces dynamic energy usage by reordering instructions with similar operand values for consecutive execution.

Haiyue Ma* and David Wentzlaff. *DVProf: Profiling Dynamic Value Locality Between Instructions.* Poster, IISWC 2024.
Proposes an instruction-level profiling tool designed to detect and analyze dynamic data value similarities that are invisible to static analysis.

Haiyue Ma* and David Wentzlaff. *Exploiting Data Commonality in Value Prediction.* The Fifth Young Architect Workshop, ASPLOS 2023.
Demonstrates how value prediction can be leveraged in domain-specific, data-intensive applications by exploiting inherent data commonality patterns.

(* indicates first author contribution)

WORK EXPERIENCE

NVIDIA Corporation Shanghai, China
Deep Learning Performance Architect *2020–2021*

- 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.
- Optimized end-to-end DL inference performance with software pipelining and kernel fusion.
- Explored sub-kernel level L2 prefetch strategies for neural network applications.
- **Haiyue Ma** et al., "Hardware Accelerated Global Synchronization," nTech 2021 Poster.

ASIC Power Engineer (Intern/Full Time) *2018–2020*

- 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

TECHNICAL STRENGTHS

Languages	C/C++, CUDA, Python, Verilog
Modeling & Tools	GPGPU-Sim, ChampSim, gem5, Intel PIN, Nsight, VTune
Architectures	GPU microarchitecture, ML hardware co-design, performance modeling, power/perf analysis

NON-DEGREE COURSES

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

HONORS & AWARDS

First-Year Ph.D. Fellowship, Princeton University	*2021*
Dean's List (All Semesters), Washington University in St. Louis	*2015–2018*

SERVICE & VOLUNTEERING

Co-Chair, Computer Architecture Long-term Mentorship Program	*2023–Present*
Steering Committee Member, Computer Architecture Student Association	*2022–Present*
Artifact Reviewer, ISCA 2024 & ISCA 2025	*2024, 2025*
Social Chair, ASPLOS 2023	*2023*
Member, WiCArch (Women in Computer Architecture)	*2023–Present*