Michael T. Shen

Electrical and Computer Engineering PhD Student

908-723-6035 | mts247@cornell.edu | michaeltshen.github.io | michaeltshen | Michael T. Shen | michael-t-shen

Research Interests

My research interests lie in computer architecture and systems, with a focus on building efficient and adaptable infrastructure for machine learning, security, and privacy. Recently, I have been optimizing inference pipelines for Retrieval-Augmented Generation (RAG) large language models and compound multi-agent systems. Broadly, I am interested in designing machine learning systems that are both efficient and extensible, capable of dynamically adapting to diverse workloads and meeting varied Service Level Objectives (SLOs).

Education

Cornell University

Ithaca, New York → Roosevelt Island, New York

Doctor of Philosophy in Electrical & Computer Engineering — *Aug. 2023 - Present*

- **Advisors**: Professor Udit Gupta & Professor G. Edward Suh
- **Selected Coursework**: Computer Architecture, High Level Design Automation, Sustainable Computing
- **Doctoral Committee**: Udit Gupta, G. Edward Suh, Zhiru Zhang

Northeastern University

Boston, Massachusetts

Bachelor of Science in Computer Engineering — *Sep. 2019 - May 2023*

- **GPA**: 3.93/4.00, *Summa Cum Laude*
- **Clubs & Organizations**: Northeastern University Computer Architecture Research Laboratory, IEEE-HKN, Tau Beta Pi, AerospaceNU - NUAV Software & Hardware
- **Selected Coursework**: Digital Design and Computer Organization, Computer Architecture and Organization, Introduction to Software Security, Computer Hardware and System Security, Introduction to Machine Learning and Pattern Recognition, Simulation and Performance Evaluation

Research Experience

Cornell Computer Systems Laboratory

Ithaca, New York → Roosevelt Island, New York

Graduate Researcher — *Aug. 2023 - Present*

- **Advisors**: Professor Udit Gupta & Professor G. Edward Suh
- **Overview**: *Developed frameworks and systems for extensible and dynamic machine learning workload acceleration.*
  - **Hermes: Retrieval–Augmented Generation Acceleration**
    - Designed and implemented an algorithm–systems co-design framework for trillion-token Retrieval-Augmented Generation, reducing retrieval-stage overhead in both latency and energy.
    - Redesigned retrieval datastores into partitioned structures with hierarchical search, reducing retrieval latency and enhancing throughput and energy efficiency while maintaining retrieval quality at scale.
    - Performed design space exploration across key hyperparameters (e.g., number of clusters, sampling depth, nProbe, retrieval stride) to identify optimal trade-offs between latency, throughput, energy, and retrieval quality.
    - Conducted extensive case studies across diverse RAG workloads, models, and hardware platforms to demonstrate the extensibility and scalability of Hermes, validating consistent improvements in latency and energy efficiency at trillion-token scale.
  - **ML Workload Characterization**
    - Evaluated end-to-end Retrieval-Augmented Generation pipelines from a systems perspective, identifying retrieval as a key bottleneck and quantifying its impact on latency, throughput, and energy efficiency.
    - Characterized how LLM agents perform across diverse hardware platforms, analyzing GPU vs. CPU execution trade-offs and uncovering how platform-specific constraints shape agent behavior.
    - Investigated emerging platforms such as the NVIDIA GH200, revealing novel trends like executing traditionally CPU-bound workloads (e.g., retrieval) directly on GPUs and leveraging the high-bandwidth CPU–GPU interconnect to efficiently manage data movement.
  - **Compound AI & Agentic Systems**
    - Explored methodologies for hardware definition in multi-agent systems, examining strategies to assign ML agents to heterogeneous platforms while optimizing for different SLOs.
    - Developed and evaluated diverse end-to-end multi-agent ML pipelines, analyzing accuracy, latency, and efficiency trade-offs across agent stages such as inference and retrieval.

Scalable Architecture Laboratory

(Remote) Williamsburg, Virginia

Undergraduate Researcher

Jan. 2022 - Jul. 2023

- **Principle Investigator**: Professor Yifan Sun
- **Overview**: *Investigated strategies for enhancing the development of lightweight, event-driven computer architecture simulators.*
- **Odyssey: Automated Simulator Calibration and Fine-Tuning**
  - Built Odyssey, a framework that automatically takes pre-constructed GPU simulators and calibrates them for closer alignment with real hardware performance metrics.
  - Designed a ranking system to identify the most influential simulator parameters for effectively tuning benchmark workloads.
  - Integrated an autotuning loop with MIT's OpenTuner to efficiently explore the GPU simulator design space and validate tuned parameters against comprehensive microbenchmark suites.
  - Tuned and modified GPU simulators (e.g., NaviSim with AMD timing models) to achieve performance deviations of less than 30% on new architectures.

Northeastern University Computer Architecture Research laboratory

Boston, Massachusetts

Research Assistant

Jun. 2020 - May 2023

- **Principle Investigator**: Professor David Kaeli
- **Overview**: *Contributed to the development of open-source computer architecture simulators for HPC and security research.*
- **NaviSim/MGPUSim: Multi-GPU Simulation**
  - Contributed to the design and development of NaviSim, an architectural simulator that supports AMD GPU architectures.
  - Built an instruction disassembler and emulator that supports AMD GCN3, RDNA, and CDNA ISAs.
  - Implemented data tracers and filtering tools to analyze microbenchmark behavior.
  - Designed and integrated an event-driven simulation for cache write-evictions.
  - Developed a validation tool to detect register conflicts across workloads, enhancing debugging and accuracy.
- **Yori: Side-Channel Attack Simulation**
  - Extended the functionality of Yori, a RISC-V simulator (built on Chipyard/BOOM) to study Spectre-style side-channel attacks.
  - Implemented and architecturally validated RISC-V privileged timing instructions.
  - Integrated modern branch predictors (GShare and TAGE) into the simulator.
  - Developed a custom benchmark suite for evaluating branch predictor performance.
- **GME: Microarchitectural GPU-Based FHE Acceleration**
  - Co-developed GME, a framework using microarchitectural extensions to accelerate fully homomorphic encryption (FHE) on GPUs.
  - Created a new modular instruction within the NaviSim ISA to optimize computationally expensive NTT operations.
  - Designed and implemented a microbenchmark suite to quantify GME performance gains in FHE operations.

Industry Experience

Advanced Micro Devices

Boxborough, Massachusetts

GPU Architecture Modeling Co-op

Jan. 2022 - Jul. 2022, Jan. 2023 - May 2023

- **Organization**: Radeon Technology Group
- Worked as a member of the Graphics Architecture Modeling Team to develop and improve upon the software tools within AMD's GPU simulators to service next-generation gaming GPUs
- Created a visualization tool integrated with AMD's GPU simulator to create graphical representations of graphics pipelines, accurately illustrating the timing and sequence of draw calls for gaming render simulation traces.
- Performed a case study on AMD's GPU simulator to identify and address discrepancies in hardware performance counters between actual hardware traces and simulation traces
- Developed a set of software tools to identify and highlight notable non-uniform memory access patterns and imbalances in performance counter analytics (such as memory reads and writes) across simulation traces

Portal Instruments

(Remote) Cambridge, Massachusetts

Software Engineering Co-op

Jan. 2021 - Jul. 2021

- Worked on the software for Portal Instrument's needle-free jet injection drug delivery device collaboratively with other software, mechanical, & electrical engineering teams
- Improved and extended software tools to increase the device's usability and functionality, including expanding injection support to enable multiple sequential injections
- Altered the design of data capture and storage functionalities to prevent information leakages and better ensure the security/validity of user information

Honors & Awards

04/2026	**ML and Systems Rising Star** MLCommons	*Santa Clara, CA*
01/2025	**Top Pick in Hardware and Embedded Security**	*Newark, NJ*
04/2023	**Huntington 100 Class Inductee** Northeastern University	*Boston, MA*
03/2023	**NSF Graduate Research Fellowship** National Science Foundation	*Boston, MA*
12/2022	**Tau Beta Pi Inductee (Engineering Honor Society)** Northeastern University	*Boston, MA*
11/2022	**1st Place ECE Capstone Competition Project** Northeastern University	*Boston, MA*
06/2022	**PEAK Experience Summit Grant** Northeastern University	*Boston, MA*
04/2022	**RISE Expo Best Undergraduate Engineering and Technology Poster** Northeastern University	*Boston, MA*
03/2021	**Eta Kappa Nu Inductee (IEEE Honor Society)** Northeastern University	*Boston, MA*
02/2021	**HackBeanpot Best UI/UX** HackBeanpot	*Boston, MA*

Service

03/2026 – Present	**Computer Architecture Long-term Mentor** Computer Architecture Student Organization (CASA)
08/2025	**Artifact Evaluation Committee Member** IISWC
07/2025	**Mentor** ISCA Undergrad Architecture Mentoring (uArch) Workshop Mentor
09/2024 – Present	**Mentor** ECE PhD Mentorship Program, Cornell University

Publications

** Denotes Equal Contribution*

- **Towards Understanding Challenges Surrounding Agentic System Design**
  M. Shen*, X. Zhou*, I. Patwardhan*, Y. Li, U. Gupta.
  North East AI Agents Day 2026

- **Hermes: Algorithm-System Co-design for Efficient Retrieval Augmented Generation At-Scale**
  M. Shen, M. Umar, K. Maeng, G.E. Suh, U. Gupta.
  International Symposium on Computer Architecture (ISCA 2025)
  Artifact Badges: Available, Functional, and Reproducible

- **Towards Understanding Systems Trade-offs in Retrieval-Augmented Generation Model Inference**
  M. Shen, M. Umar, K. Maeng, G. E. Suh, U. Gupta
  arXiv preprint, arXiv:2412.11854

- **GME: GPU-based Microarchitectural Extensions to Accelerate Homomorphic Encryption**
  K. Shivdikar, Y. Bao, R. Agrawal, M. Shen, G. Jonatan, E. Mora, A. Ingare, N. Livesay, J. Abellan, J. Kim, A. Joshi, D. Kaeli.
  International Symposium on Microarchitecture (MICRO 2023)
  Top Pick in Hardware and Embedded Security

- **NaviSim: A Highly Accurate GPU Simulator for AMD RDNA GPUs**
  Y. Bao, Y. Sun, Z. Feric, M. Shen, M. Weston, J. Abellán, T. Baruah, J. Kim, A. Joshi, D. Kaeli.
  International Conference on Parallel Architectures and Compilation Techniques (PACT 2022)
  Artifact Badges: Available, Functional, and Reproducible