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 the areas of computer architecture and systems, with an emphasis on building efficient and adaptable infrastructure for machine learning, security, and privacy. Recently, my work has focused on optimizing inference pipelines for Retrieval-Augmented Generation (RAG) large language models and compound AI systems. I am especially interested in developing extensible machine learning systems capable of dynamically adapting to diverse workloads and deployment scenarios.

Education

Cornell University

Ithaca, New York → Roosevelt Island, New York

DOCTOR OF PHILOSOPHY IN ELECTRICAL & COMPUTER ENGINEERING — *Aug. 2023 - Present*

- **Advisors**: Dr. Udit Gupta & Dr. G. Edward Suh
- **Selected Coursework**: Computer Architecture, High Level Design Automation, Sustainable Computing

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

Industry Experience

Advanced Micro Devices

Boxborough, Massachusetts

GRAPHICS 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

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

Research Experience

Cornell Computer Systems Laboratory

Ithaca, New York → Roosevelt Island, New York

GRADUATE RESEARCHER — *Aug. 2023 - Present*

- **Group Affiliations**: Scalable, Specialized, Sustainable, Systems for AI, Suh Research Group
- **Advisors**: Dr. Udit Gupta & Dr. G. Edward Suh
- Conducted research on the characterization and acceleration of Retrieval Augmented Generation (RAG) language model systems
- Characterization of RAG Systems: In an effort to begin tackling and mitigating the performance penalties associated with RAG from a systems perspective, we taxonomized and characterized the different elements within the RAG ecosystem for LLMs that explore trade-offs within latency, throughput, and memory. We conducted a study that revealed underlying inefficiencies in RAG for systems deployment, that can result in TTFT latencies that are twice as long and unoptimized datastores that consume terabytes of storage.
- Hermes: In an effort to address the unique bottlenecks presented within the retrieval stage of large-scale RAG systems, we explore and develop algorithm-systems co-design frameworks to mitigate these overheads. Our system can significantly reduce retrieval latency by partitioning and distributing datastores across multiple nodes, while also enhancing throughput and energy efficiency through an intelligent hierarchical search that dynamically directs queries to optimized subsets of the datastore.

Scalable Architecture Laboratory

Williamsburg, Virginia

Undergraduate Researcher

Jan. 2022 - Jul. 2023

- **Principle Investigator**: Dr. Yifan Sun
- Contributed to the development of lightweight, event-driven computer architecture simulators for large-scale workloads: Odyssey (A framework for simulator timing model auto-generation) and Pipesim
- Odyssey Contributions: Designed a methodology for automatically calibrating NaviSim (An AMD Architecture GPU Simulator) against actual hardware with weighting and configuring scripts, integrated an autotuning framework with custom accuracy metrics into NaviSim, created an automated testing validation framework for evaluating the efficacy of simulated microbenchmarks, conducted a case study on Odyssey that found it could create AMD architecture simulator timing models that deviated by less than 30% from actual hardware.

Northeastern University Computer Architecture Research laboratory

Boston, Massachusetts

Research Assistant

Jun. 2020 - May 2023

- **Principle Investigator**: Dr. David Kaeli
- Contributed to the development of several open-source computer architecture simulators for high-performance and security research, including NaviSim/MGPUSim (a multi-architecture GPU simulator for high-performance computer architecture research), Yori (a lightweight RISC-V architecture simulator for side channel research), and GME (an extension of NaviSim for exploring acceleration techniques for homomorphic encryption).
- NaviSim/MGPUSim Contributions: Lead developer for the frontend instruction disassembler and emulator for the AMD RDNA and CDNA ISAs, created data tracers and filters for microbenchmarks, implemented the cache write eviction scheme, developed a validation tool for identifying register conflicts in microbenchmarks
- Yori Contributions: Implemented and validated RISCV privileged timing instructions architecturally, integrated modern branch predictor designs into the simulator design (GShare and TAGE), developed a benchmark suite for branch predictor evaluation
- GME Contributions: Created a new modular instruction within the NaviSim ISA for theoretical timing analysis, developed a microbenchmark suite for evaluating our microarchitectural performance enhancements on homomorphic encryption algorithms

Honors & Awards

01/2025	**2024 Top Pick in Hardware and Embedded Security**	*Newark, NJ*
04/2023	**Inductee** Northeastern 2023 Huntington 100 Class	*Boston, MA*
03/2023	**Awardee** National Science Foundation Graduate Research Fellowship	*Boston, MA*
12/2022	**Inductee** Tau Beta Pi	*Boston, MA*
11/2022	**1st Place** Northeastern Electrical and Computer Engineering Capstone Competition	*Boston, MA*
06/2022	**Awardee** PEAK Experience Summit Grant	*Boston, MA*
04/2022	**Best Undergraduate Engineering and Technology Poster** Northeastern 2022 RISE Expo	*Boston, MA*
03/2021	**Inductee** IEEE Eta Kappa Nu	*Boston, MA*
02/2021	**Best UI/UX** Hack Beanpot	*Boston, MA*

Publications

- **Hermes: Algorithm-System Co-design for Efficient Retrieval Augmented Generation At-Scale**
  M. Shen, M. Umar, K. Maeng, G.E. Suh, U. Gupta.
  [To Appear] 52nd ACM/IEEE 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.
  56th IEEE/ACM 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.
  31st International Conference on Parallel Architectures and Compilation Techniques (PACT 2022)
  Artifact Badges: Available, Functional, and Reproducible