Ye, Haojie

734.239.3020 | yehaojie@umich.edu | https://web.eecs.umich.edu/~yehaojie/

EDUCATION

Ph.D. in Computer Science and Engineering June 2021 – Sept 2024
Thesis Advisor: Prof. Trevor Mudge University of Michigan, Ann Arbor

- **Overall GPA**: 4.0/4.0
- **Area Focus**: Computer Architecture, Algorithm, and System Design for emerging applications such as Recommendation Systems, Graph-based Machine Learning, and Graph Mining

M.S.E. in Computer Science and Engineering Sept 2019 – May 2021
Thesis Advisor: Prof. Trevor Mudge University of Michigan, Ann Arbor

- **Overall GPA**: 4.0/4.0

B.S.E. in Computer Engineering Sept 2017 – May 2019
University of Michigan, Ann Arbor

FULL-TIME POSITION

Nvidia — Deep Learning Performance Architect Oct 2024 – Present
Performance Modeling and Architecture Pathfinding Santa Clara, California

- Model performance of state-of-the-art AI models (e.g., Llama 4, DeepSeek, Qwen) on NVIDIA next-generation GPUs (pre-silicon) for performance optimization and product planning.
- Develop and use NVIDIA internal GPU simulators to launch large-scale modeling experiments, analyze results, and guide architecture pathfinding for upcoming products.
- Contributed to performance modeling and optimization efforts for NVIDIA's **RTX 6000 Pro** and future **Rubin** products.
- Collaborate with architecture, product, and pricing teams to support GPU performance projections used in high-profile events such as **NVIDIA GTC**.
- Engage with new Transformer models and optimizations to identify architectural bottlenecks and future design tradeoffs across compute, memory, and interconnect.

SELECTED PUBLICATIONS

1. **H. Ye**, Y. Xia, Y. Chen, KY. Chen, Y. Yuan, S. Deng, B. Kasikci, T. Mudge, N. Talati, "Palermo: Improving the Performance of Oblivious Memory using Protocol-Hardware Co-Design" (**HPCA'25**)
2. Y. Xia, J. Kim, Y. Chen, **H. Ye**, S. Kundu, C. Hao, N. Talati, "Understanding the performance and estimating the cost of llm fine-tuning" (**IISWC'24**)
3. Y. Chen, **H. Ye**, S. Vedula, A. Bronstein, R. Dreslinski, T. Mudge, N. Talati, "Demystifying graph sparsification algorithms in graph properties preservation" (**VLDB'24**)
4. Y. Yuan, **H. Ye**, S. Vedula, W. Kaza, N. Talati, "Everest: GPU-Accelerated System For Mining Temporal Motifs" (**VLDB'24**)
5. **H. Ye**, S. Vedula, Y. Chen, Y. Yang, A. Bronstein, R. Dreslinski, T. Mudge, N. Talati, "GRACE: A Scalable Graph-Based Approach To Accelerating Recommendation Model Inference" (**ASPLOS'23**)
6. H. Kim, **H. Ye**, T. Mudge, R. Dreslinski, N. Talati, "RecPIM: A PIM–Enabled DRAM–RRAM Hybrid Memory System To Accelerate Recommendation Models" (**ISLPED'23**)
7. N. Talati, **H. Ye**, S. Vedula, K. Chen, Y. Chen, D. Liu, D. Blaauw, A. Bronstein, T. Mudge, R. Dreslinski, "Mint: An Accelerator For Mining Temporal Motifs" (**MICRO'22**)
8. L. Belayneh, **H. Ye**, K. Chen, D. Blaauw, T. Mudge, R. Dreslinski, N. Talati, "Locality-aware Optimizations for Improving Remote Memory Latency in Multi-GPU Systems" (**PACT'22**)

9. N. Talati, **H. Ye**, Y. Yang, L. Belayneh, K. Chen, D. Blaauw, T. Mudge, R. Dreslinski, "NDMiner: Accelerating Graph Pattern Mining Using Near Data Processing" (**ISCA'22**)
10. N. Talati, D. Jin, Di, **H. Ye**, A. Brahmakshatriya, G. Dasika, S. Amarasinghe, T. Mudge, D. Koutra, R. Dreslinski, "A Deep Dive Into Understanding The Random Walk-Based Temporal Graph Learning" (**IISWC'21**)
11. X. He, S. Pal, A. Amarnath, S. Feng, D. Park, A. Rovinski, **H. Ye**, Y. Chen, R. Dreslinski, T. Mudge, "Sparse-tpu: Adapting systolic arrays for sparse matrices" (**ICS'20**)

INTERNSHIP

Nvidia — Deep Learning Performance Architect Intern May 2023 – Aug 2023
Performance modeling and projection at Nvidia Santa Clara, California

- Independent R&D in modeling LLMs on future GPUs.

Micron Technology — Advanced Hardware Development Intern May 2022 – Aug 2022
Enhanced in-memory function research at Micron Allen, Texas

- Independent research on enhanced in-memory functions of Micron's next-generation memory module.
- Published 1 Micron internal journal and 2 U.S. patents with Micron.

TECHNICAL SKILLS

- **Programming Languages**: C/C++, Python, CUDA, Verilog HDL, SystemVerilog, HTML, Matlab
- **Software and Design Tools**: Git, Murphi, Cadence, Design Compiler, LaTeX
- **Architecture Simulators**: Gem5, Gem5-gpu, DRAMSim, Ramulator, CACTI, Zsim

TEACHING EXPERIENCE

Graduate Student Instructor Sept 2020 – Jan 2021
EECS 598 Applied Parallel Programming with GPUs University of Michigan, Ann Arbor

- Taught with Prof. Reetuparna Das; prepared lecture slides, homework, and exams on GPU microarchitecture and CUDA programming.

AWARDS & HONORS

- Graduate Student Research Assistant (GSRA), University of Michigan (2020)
- Outstanding Undergraduate Research Award (2019)
- James B. Angell Scholar (2019)

REFERENCES

- Prof. Trevor Mudge, Bredt Family Professor, University of Michigan — tnm@umich.edu
- Prof. Ronald Dreslinski, Associate Professor, University of Michigan — rdreslin@umich.edu
- Dr. Nishil Talati, Assistant Research Scientist, University of Michigan — talatin@umich.edu