Ruokai Yin

PhD candidate, ECE, Yale University

ruokai.yin@yale.edu Google Scholar Github Personal Website

Research Interests

- Low-power computer architectures and systems design and modeling for energy-efficient AI workloads, particularly those involving asymmetric operand precision or sparsity.
- AI algorithm-hardware co-design for model compression (pruning and quantization)

Education

Ph.D., Electrical and Computer Engineering, Yale University **Sep. 2021 — Present**
Advisor: Prof. Priyadarshini Panda Expected: May. 2026

B.S., Electrical Engineering & Computer Science & Math, University of Wisconsin - Madison **Sep. 2018 — May. 2021**
Graduated with Distinction, GPA: 3.98/4.00
Advisor: Prof. Joshua San Miguel

Employment

Research Intern, Azure-AI Architecture and System team, Microsoft, **May. 2025 — Present**
Mentors: Apala Guha & Xuan Zuo

Research Intern, ASIC team, Cerebras Systems, **May. 2024 — Aug. 2024**
Mentor: Vipin Sharma
Architecture design and modeling for Cerebras's next-generation wafer-scale engine, with a focus on intra-PE, inter-PE, and IO level.

Publications [Selected]

Computer architecture & Domain-specific acceleration:

PacQ: A SIMT Microarchitecture for Efficient Dataflow in Hyper-asymmetric GEMMs.
Ruokai Yin, Yuhang Li, and Priyadarshini Panda
ACM/IEEE Design Automation Conference (DAC) 2025.

LoAS: Fully Temporal-Parallel Dataflow for Dual-Sparse Spiking Neural Networks.
Ruokai Yin, Youngeun Kim, Di Wu, and Priyadarshini Panda
International Symposium on Microarchitecture (MICRO) 2024. open-source artifact.

SATA: Sparsity-Aware Training Accelerator for Spiking Neural Networks.
Ruokai Yin, Abhishek Moitra, Abhiroop Bhattacharjee, Youngeun Kim, and Priyadarshini Panda
IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2022. open-source simulator.

uGEMM: Unary Computing Architecture for GEMM Applications.
Di Wu, Jingjie Li, **Ruokai Yin**, Hsuan Hsiao, Younghyun Kim, Joshua San Miguel
International Symposium on Computer Architecture (ISCA) 2020, **IEEE Top-pick 2020.** open-source simulator.

AI Algorithm-Architecture Co-Design:

DuoGPT: Training-free Dual Sparsity through Activation-aware Pruning in LLMs.
Ruokai Yin, Yuhang Li, Donghyun Lee, Priyadarshini Panda
Conference on Neural Information Processing Systems (NeurIPS), 2025. *under submission*

GPTAQ: Efficient Finetuning-Free Quantization for Asymmetric Calibration.
Yuhang Li, **Ruokai Yin**, Donghyun Lee, Shiting Xiao, Priyadarshini Panda
International Conference on Machine Learning (ICML), 2025. open-source code.

MINT: Multiplier-less Integer Quantization for Spiking Neural Networks.
Ruokai Yin, Yuhang Li, Abhishek Moitra, and Priyadarshini Panda
Asia and South Pacific Design Automation Conference (ASP-DAC) 2024, **Best Paper Award Nomination.** open-source code.

Workload-balanced Pruning for Sparse Spiking Neural Networks.
Ruokai Yin, Youngeun Kim, Yuhang Li, Abhishek Moitra, Nitin Satpute, Anna Hambitzer, Priyadarshini Panda
IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI), 2024. open-source code.

Complete publication list available at here.

Awards & Honors

Research:

DAC Young Fellow, DAC, 2025
Best Paper Award Nomination, ASP-DAC, 2024
Spotlight Paper, NeurIPS Workshop on Learning from Time Series for Health, 2022
IEEE Micro Top Pick, Computer Architecture, 2020

Academic:

Conference Travel Fellowship, Yale University, Fall 2024
John Bennett Fenn Fellowship Fund, Yale University, Fall 2021 – Spring 2022
Distinctive Scholastic Achievement, University of Wisconsin - Madison, Spring 2021
Dean's Honor List, University of Wisconsin - Madison, Fall 2018 – Spring 2021

Talks

LoAS: Fully Temporal-Parallel Dataflow for Dual-Sparse Spiking Neural Networks
57th MICRO (Austin, USA), Nov 2024

MINT: Multiplier-less Integer Quantization for Energy Efficient Spiking Neural Networks
29th ASP-DAC (Incheon, South Korea), Jan 2024

SATA: Sparsity-Aware Training Accelerator for Spiking Neural Networks
Center for Brain-Inspired Computing (C-BRIC, SRC), Nov 2022

UnarySim and Characterizing Early Termination in Stochastic Computing
2020 UW Computer Architecture Industrial Affiliates (Madison, WI, USA), Sep 2020

Teaching Experience

TA - EENG 439, Neural Networks & Learning Systems, Fall 2023
Instructor: Prof. Priya Panda

TA - EENG 348, Digital Systems, Spring 2023
Instructor: Prof. Rajit Manohar

Academic Activities

Reviewer
- IEEE Transactions on Neural Networks and Learning Systems
- IEEE International Symposium on Circuits and Systems, 2024
- IEEE Journal on Emerging and Selected Topics in Circuits and Systems
- IEEE Transactions on Very Large Scale Integration Systems
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems
- AI Communications