YUQI XUE

Email: yuqixue2@illinois.edu ◇ Phone: +1-217-721-4958

yuqixue.com

RESEARCH INTERESTS

I work on computer architecture for AI chips. I build the hardware and software ecosystem for neural processing units, making them more performant, cost-efficient, and energy-efficient for cutting-edge machine learning workloads like large language models (LLMs). My research explores techniques across the architectural, system, and compiler stack.

EDUCATION

University of Illinois Urbana-Champaign *Aug 2021 – Present*
Ph.D. in Electrical and Computer Engineering
Advised by Prof. Jian Huang

University of Illinois Urbana-Champaign *Aug 2017 – May 2021*
Bachelor of Science in Computer Engineering with High Honors
Bachelor of Science in Mathematics with Highest Distinction

PUBLICATION

NEURAL PROCESSING UNIT ECOSYSTEM

- **ReGate: Enabling Power Gating in Neural Processing Units.** Yuqi Xue and Jian Huang. In *58th IEEE/ACM International Symposium on Microarchitecture (MICRO'25)*. 2025.
- **Hardware-Assisted Virtualization of Neural Processing Units for Cloud Platforms.** Yuqi Xue, Yiqi Liu, Lifeng Nai, and Jian Huang. In *Proceedings of the 57th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'24)*. 2024. Received Best Paper Runner-Up Award (top 3 out of 497 submissions). Selected as IEEE Micro Top Picks.
- **V10: Hardware-Assisted NPU Multi-tenancy for Improved Resource Utilization and Fairness.** Yuqi Xue, Yiqi Liu, Lifeng Nai, and Jian Huang. In *Proceedings of the 50th International Symposium on Computer Architecture (ISCA'23)*. 2023.
- **System Virtualization for Neural Processing Units.** Yuqi Xue, Yiqi Liu, and Jian Huang. In *Proceedings of the 19th Workshop on Hot Topics in Operating Systems (HotOS'23)*. 2023.

ML COMPILERS FOR INTER-CORE CONNECTED AI CHIPS

- **Elk: Exploring the Efficiency of Inter-core Connected AI Chips with Deep Learning Compiler Techniques.** Yiqi Liu, Yuqi Xue, Noelle Crawford, Jilong Xue, and Jian Huang. In *58th IEEE/ACM International Symposium on Microarchitecture (MICRO'25)*. 2025.
- **Scaling Deep Learning Computation over the Inter-Core Connected Intelligence Processor.** Yiqi Liu, Yuqi Xue, Yu Cheng, Lingxiao Ma, Ziming Miao, Jilong Xue, and Jian Huang. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles (SOSP'24)*. 2024.

OTHER PUBLICATIONS

- **Toward Engineering AGI: Benchmarking the Engineering Design Capabilities of LLMs.** Xingang Guo, Yaxin Li, Xiangyi Kong, Yilan Jiang, Xiayu Zhao, Zhihua Gong, Yufan Zhang, ..., Yuqi Xue, et al. In *39th Annual Conference on Neural Information Processing Systems (NeurIPS'25)*. 2025.

- **Managing Scalable Direct Storage Accesses for GPUs with GoFS.** Shaobo Li*, Yirui Eric Zhou*, Yuqi Xue, Yuan Xu, and Jian Huang. In *31st Symposium on Operating Systems Principles (SOSP'25)*. 2025.
  *Co-primary authors.
- **Architecting An Efficient Memory-Semantic CXL-based SSD with OS and Hardware Co-design.** Haoyang Zhang*, Yuqi Xue*, Yirui Eric Zhou, Shaobo Li, and Jian Huang. In *Proceedings of the 31th IEEE International Symposium on High-Performance Computer Architecture (HPCA'25)*. 2025.
  *Co-primary authors.
- **G10: Enabling An Efficient Unified GPU Memory and Storage Architecture with Smart Tensor Migrations.** Haoyang Zhang*, Yirui Eric Zhou*, Yuqi Xue, Yiqi Liu, and Jian Huang. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'23)*. 2023.
  *Co-primary authors.
- **RackBlox: A Software-Defined Rack-Scale Storage System with Network-Storage Co-Design.** Benjamin Reidys, Yuqi Xue, Daixuan Li, Bharat Sukhwanim, Wen-mei Hwu, Deming Chen, Sameh Asaad, and Jian Huang. In *29th ACM Symposium on Operating Systems Principles (SOSP'23)*. 2023.
- **IceClave: A Trusted Execution Environment for In-Storage Computing.** Luyi Kang*, Yuqi Xue*, Weiwei Jia*, Xiaohao Wang, Jongryool Kim, Changhwan Youn, Myeong Joon Kang, Hyung Jin Lim, Bruce Jacob, and Jian Huang. In *Proceedings of the 54th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'21)*. 2021.
  *Co-primary authors.

WORK EXPERIENCE

Systems Platform Research Group (Illinois PlatformX), UIUC — *May 2021 - Present*
Graduate Research Assistant — *Urbana, IL*

Google LLC — *Fall 2023*
Student Researcher — *Remote*

Google LLC — *Summer 2023*
Research Intern — *Sunnyvale, CA*

TEACHING

ECE 511: Computer Architecture, UIUC, *Graduate TA* — *Spring 2024, Fall 2022*
- List of Teachers Ranked as Excellent By Their Students (Fall 2022)

ECE 522: Emerging Memory and Storage Systems, UIUC, *Graduate TA* — *Spring 2023*

ECE 411: Computer Organization and Design, UIUC, *Graduate TA* — *Fall 2023, Spring 2022*

ECE 310: Digital Signal Processing, UIUC, *Undergraduate Grader* — *Jan 2020 - Jul 2020*

SERVICE

Artifact Evaluation Committee

- IEEE/ACM International Symposium on Microarchitecture (MICRO), 2022
- USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2022, 2023
- USENIX Annual Technical Conference (ATC), 2022, 2023
- IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2023

Peer Review

- IEEE Computer Architecture Letters, 2022, 2023, 2024

· ACM Transactions on Architecture and Code Optimization, 2024, 2025

· Student PC, MICRO 2026

Organizing Committee

· Workshop on Hot Topics in System Infrastructure (HotInfra), Web Chair, 2023, 2024, 2026

Tutorial

· NexNPU: Building an Open-Source Ecosystem for Neural Processing Units (colocated with ISCA'26), 2026

AWARD

- IEEE Micro Top Picks from 2024 Computer Architecture Conferences
- Best Paper Runner-Up, MICRO 2024
- Dan Vivoli Endowed Fellowship, 2024
- Rambus Computer Engineering Fellowship, 2025
- Qualcomm Graduate Award, Spring 2025
- Yi-Min Wang and Pi-Yu Chung Research Award, 2025
- Student Travel Grants, ISCA 2023, MICRO 2024
- Oral Presentation Award, CSL Student Conference, 2024