Yufeng Gu

📍 *4844 Bob and Betty Beyster Building, 2260 Hayward*
Ann Arbor, MI, USA, 48109-2121
✉ *yufenggu@umich.edu* 📱 *+1-734-272-5271*

Education

University of Michigan — **Ann Arbor, MI, USA**
Ph.D., Computer Science and Engineering Department — *Sep. 2020 - Apr. 2026*

Zhejiang University (ZJU) — **Hangzhou, China**
B.Eng., College of Information Science and Electronic Engineering — *Sep. 2016 - June 2020*

Professional Experience

University of Michigan — **Ann Arbor, MI, USA**
Graduate Research Assistant, Advisor: Reetuparna Das — *Sep. 2020 – Dec. 2025*
Architecture and system co-design for large-scale emerging applications, including Generative AI and genome sequencing. Publications at top conferences including ISCA, ASPLOS and HPCA.

Tenstorrent Inc. — **Santa Clara, CA, USA**
Performance Architect Intern, Manager: Wei-han Lien — *May 2023 – Aug. 2023*
Built performance model for RISC-V CPU and AI accelerator.

Intel Labs — **Hillsboro, OR, USA**
Graduate Research Intern, Manager: Nilesh Jain — *June 2022 – Aug. 2022*
Optimized AI QoS on next generation heterogenous datacenter platform.

Yale University — **New Heaven, CT, USA**
Undergraduate Research Intern, Advisor: James Duncan, Xiaoxiao Li — *Nov. 2019 – Mar. 2020*
Interpretation on ASD with fMRI and Deep Learning Models. Publications on MICCAI and MedIA.

École Polytechnique Fédérale De Lausanne (EPFL) — **Lausanne, Switzerland**
Undergraduate Research Intern, Advisor: Babak Falsafi — *July 2019 – Sep. 2019*
Patched QFlex Simulator (from Makefile to CMake compilation) and accelerated activation functions for DNN.

Award & Honors

Outstanding Young Speaker Award (APPT 2025) — *July 2025*
Hardware-Software Co-Design for LLM Inference

Distinguished Artifact Honorable Mention (HPCA 2025) — *March 2025*
Multi-Dimensional Vector ISA Extension for Mobile In-Cache Computing. (3/29 Artifacts)

Communucation of ACM (CACM) Research Highlights — *July 2024*
GenDP: A Framework of Dynamic Programming Acceleration for Genome Sequencing Analysis.
CACM Research Highlights section selects 24/10,000+ papers from ACM conferences per year, reprinting the most significant and influential results across Computer Science.

University of Michigan Rackham Graduate Student Research Grant ($3000, role: PI) — *Apr. 2024*
Pangenomics Benchmark Suite and Characterization.

University of Michigan Rackham Travel Grant, ISCA/HPCA Travel Grant — *2023, 2025*

Fellowship of Summer@EPFL (2% applicants awarded) — *July 2019*

Tang Lixin Fellowship (2‰ students awarded) — *Nov. 2017, 2018, 2019*

Outstanding Student Leaders in Zhejiang University (3% students awarded) — *Oct. 2017, 2019*

First-Class Scholarship for Outstanding Students (2% students awarded) — *Oct. 2017*

Under-Review Publications

Yufeng Gu, Samel Gobriel, Nilesh Jain, Reetuparna Das. "DREAM: Data Reuse-Aware Processing-In-Memory Architecture for Efficient Large Language Model Inference. *(Under Submission to ASPLOS'26)*

Yufeng Gu, Vui Seng Chua, Nilesh Jain, Ravishankar Iyer, Reetuparna Das. "Farm: Fast Resource Management for Quality of Service-Aware Co-Location of Machine Learning Inference." *(Under Submission to SIGMETRICS'26)*

Sumanth Umesh, Ning Liang, **Yufeng Gu**, Reetuparna Das. "Addressing In-Memory Database Bottlenecks with CXL Memory Expansion." *(Under Submission to ASPLOS'26)*

Peer-Reviewed Publications

* equal contribution

Noah Kaplan, Jan-Niklas Schmelzle, **Yufeng Gu**, Christopher Batten, Reetuparna Das. "PangenomicsBench: A Benchmark Suite and Characterization of Pangenomics." To appear on IEEE International Symposium on Workload Characterization (IISWC) 2025.

Yufeng Gu, Arun Subramaniyan, Tim Dunn, Alireza Khadem, Kuan-Yu Chen, Somnath Paul, Md Vasimuddin, Sanchit Misra, David Blaauw, Satish Narayanasamy, Reetuparna Das. "GenDP: A Framework of Dynamic Programming Acceleration for Genome Sequencing Analysis." (Invited Paper) Communications of the ACM, 2025.

Alireza Khadem*, Kamalavasan Kamalakkannan*, Zhenyan Zhu, Akash Poptani, **Yufeng Gu**, Jered Benjamin Dominguez-Trujillo, Nishil Talati, Daichi Fujiki, Scott Mahlke, Galen Shipman, Reetuparna Das. "DX100: Programmable Data Access Accelerator for Indirection." In Proceedings of the 52th Annual International Symposium on Computer Architecture (ISCA'25).

Yufeng Gu*, Alireza Khadem*, Sumanth Umesh, Ning Liang, Xavier Servot, Onur Mutlu, Ravishankar Iyer, Reetuparna Das. "PIM Is All You Need: A CXL-Enabled GPU-Free System for Large Language Model Inference." In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'25). **In Progress of Transferring to SK Hynix's Next Generation Datacenter**

Alireza Khadem, Daichi Fujiki, Hilbert Chen, **Yufeng Gu**, Nishil Talati, Scott Mahlke, Reetuparna Das. "Multi-Dimensional Vector ISA Extension for Mobile In-Cache Computing." In 2025 IEEE International Symposium on High-Performance Computer Architecture (HPCA'25). **Distinguished Artifact Honorable Mention (3/29)**

Yufeng Gu, Arun Subramaniyan, Tim Dunn, Alireza Khadem, Kuan-Yu Chen, Somnath Paul, Md Vasimuddin, Sanchit Misra, David Blaauw, Satish Narayanasamy, Reetuparna Das. "GenDP: A Framework of Dynamic Programming Acceleration for Genome Sequencing Analysis." In Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA'23). **Communication of ACM Research Highlights (24/10,000+)**

Arun Subramaniyan, **Yufeng Gu**, Timothy Dunn, Somnath Paul, Md Vasimuddin, Sanchit Misra, David Blaauw, Satish Narayanasamy, and Reetuparna Das. "GenomicsBench: A Benchmark Suite for Genomics." In IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'21).

Xiaoxiao Li, **Yufeng Gu**, Nicha Dvornek, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results." Medical Image Analysis 65: 101765. *IF = 11.1*

Xiaoxiao Li, Yuan Zhou, Nicha C. Dvornek, **Yufeng Gu**, Pamela Ventola, and James S. Duncan. "Efficient Shapley Explanation for Features Importance Estimation Under Uncertainty." In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'20).

Talks

Hardware-Software Co-Design for LLM Inference

PhD Forum at APPT 2025	*July 2025*
Los Alamos National Laboratory	*May 2025*

CXL-enabled PIM system for LLM inference

SAFARI Live Seminar (Host: Prof. Onur Mutlu)	*June 2025*
MCCSys workshop co-located with ASPLOS 2025	*Mar. 2025*

Genomics Benchmark Suite and Accelerator Design

Computer Architecture Seminar at UCF (Host: Prof. Di Wu)	*Mar. 2024*
Cornel University (Host: Prof. Christopher Batten)	*Feb. 2024*
Peisu Xia Forum at ICT, CAS	*Dec. 2023*

Teaching Experience

Programming at Discovery Engineering program at University of Michigan	*July. 2022*
Mathematics at Tuanlin Primary School, Guizhou, China	*Aug. 2017*

Services

Sub-reviewer for ISCA 2023, ISCA 2024, MICRO 2025.

Artifact evaluation reviewer for IISWC 2021, ISCA 2023, ISCA 2024, ISCA 2025.

Reviewer for University of Michigan graduate student admission *Jan. 2021, 2022*

Mentorship

Dev Singhania M.S. UM	*Jan. 2025 - Apr. 2025*
Zesen Zhao M.S. UM (Now Ph.D. at UMich)	*Jan. 2025 - Aug. 2025*
Wenjie Geng M.S. UM	*Sep. 2024 - Aug. 2025*
Mayne Mei B.Eng. UM (Now Engineer at Etched)	*Jan. 2024 - Apr. 2024*
Ao Luo M.S. UM	*Jan. 2024 - June 2024*
Yuzhe Ruan B.S. UM (Now M.S. at Yale)	*Sep. 2023 - Apr. 2024*
Donglin Yu B.Eng. (Now M.Eng. at UIUC)	*Jan. 2023 - Apr. 2024*
Dawit Melka B.Eng. Addis Ababa University (Now Engineer at iCog Labs)	*May 2022 - Aug. 2022*
James Gu B.Eng. UM (Now Engineer at Amazon)	*May 2022 - Aug. 2022*