

Xiaoyang Lu

917-755-1369 | xlu40@illinoistech.edu | Chicago, IL

EDUCATION

Illinois Institute of Technology <i>Ph.D. in Computer Science</i> Advisor: Professor Xian-He Sun Thesis: Utilizing Concurrent Data Accesses for Data-Driven and AI Applications	Chicago, IL Aug 2017 – May 2024
New York University <i>M.S. in Computer Engineering</i>	New York, NY Aug 2015 – May 2017
Zhejiang University <i>B.E. in Electronic Science and Technology</i>	Hangzhou, China Aug 2011 – July 2015

RESEARCH EXPERIENCE

Research Assistant Professor <i>Illinois Institute of Technology</i>	June 2024 – Present Chicago, IL
<ul style="list-style-type: none">Conduct comprehensive research in memory-centric computer architectures and scalable memory systems, focusing on optimizing high-performance computing systems.Explore and develop hardware/software co-designed accelerators for machine learning workloads, achieving significant improvements in data access speeds and computational efficiency.Investigate and implement processing-in-memory (PIM) architectures to minimize data movement and maximize computational speed, enhancing system performance.Direct and supervise PhD research, mentoring students in advancing the field of computer architecture and high-performance computing.	
Research Assistant <i>Illinois Institute of Technology</i>	Jan 2020 – May 2024 Chicago, IL
<ul style="list-style-type: none">Focused on memory performance optimizations, developing sophisticated models and pioneering machine learning-assisted architectural innovations.Designed and implemented intelligent frameworks aimed at enhancing cache performance, focusing on efficiency and innovative design principles.Mentored multiple graduate students, guiding their research projects and fostering both their academic development and practical engineering skills.	
Research Aide <i>Argonne National Laboratory</i>	May 2020 – Aug 2020 Lemont, IL
<ul style="list-style-type: none">Conducted comprehensive performance testing on disaggregated memory systems, identifying key areas for improvement.Developed and refined performance models for disaggregated memory systems, enhancing predictive accuracy and system efficiency.Quantified and mitigated interference in disaggregated memory systems, ensuring optimal operation and reliability.	

CONFERENCE PUBLICATIONS

- [IPDPS 2026] I/O-Aware PIM Acceleration for Long-Sequence LLM Inference with Hybrid Sparse Attention
Xiaoyang Lu, Lihan Hu, Hongrui Huang, Peng Jiang, Xian-He Sun
In the Proceedings of the 40th IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2026
- [ASPLOS 2026] I/O Analysis is All You Need: An I/O Analysis for Long-Sequence Attention
Xiaoyang Lu, Boyu Long, Xiaoming Chen, Yinhe Han, Xian-He Sun
In the Proceedings of the 31st International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2026

- [DATE 2026] Zion: A Comprehensive, Adaptive, and Lightweight Hardware Prefetcher
Vadim Biryukov, **Xiaoyang Lu**, Zirui Liu, Kaixiong Zhou, Xian-He Sun
In the Proceedings of the Design, Automation, and Test in Europe (DATE), 2026
- [MICRO 2025] COSMOS: RL-Enhanced Locality-Aware Counter Cache Optimization for Secure Memory
Haoran Geng, **Xiaoyang Lu**, Yuezhi Che, Ziang Tian, Dazhao Cheng, Xian-He Sun, Michael Niemier, X. Sharon Hu
In the Proceedings of the 58th International Symposium on Microarchitecture (MICRO), 2025
- [GLSVLSI 2025] Concurrency-Aware Cache Miss Cost Prediction with Perceptron Learning
Yuping Wu, **Xiaoyang Lu**, Xiaoming Chen, Yinhe Han, Xian-He Sun
In the Proceedings of the 35th Great Lakes Symposium on VLSI (GLSVLSI), 2025
- [ICCD 2024] AceMiner: Accelerating Graph Pattern Matching using PIM with Optimized Cache System
Liang Yan, **Xiaoyang Lu**, Xiaoming Chen, Sheng Xu, Xingqi Zou, Yinhe Han, Xian-He Sun
In the Proceedings of the 42nd International Conference on Computer Design (ICCD), 2024
- [ASPLOS 2024] ACES: Accelerating Sparse Matrix Multiplication with Adaptive Execution Flow and Concurrency-Aware Cache Optimizations
Xiaoyang Lu, Boyu Long, Xiaoming Chen, Yinhe Han, Xian-He Sun
In the Proceedings of the 29th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2024
- [HPCA 2024] CHROME: Concurrency-Aware Holistic Cache Management Framework with Online Reinforcement Learning
Xiaoyang Lu, Hamed Najafi, Jason Liu, Xian-He Sun
In the Proceedings of the 30th International Symposium on High-Performance Computer Architecture (HPCA), 2024
- [HPCA 2023] CARE: A Concurrency-Aware Enhanced Lightweight Cache Management Framework
Xiaoyang Lu, Rujia Wang, Xian-He Sun
In the Proceedings of the 29th International Symposium on High-Performance Computer Architecture (HPCA), 2023
- [WSC 2022] A Generalized Model For Modern Hierarchical Memory System
Hamed Najafi, **Xiaoyang Lu**, Jason Liu, Xian-He Sun
In the Proceedings of the Winter Simulation Conference (WSC), 2022
- [ICCD 2021] Premier: A Concurrency-Aware Pseudo-Partitioning Framework for Shared Last-Level Cache
Xiaoyang Lu, Rujia Wang, Xian-He Sun
In the Proceedings of the 39th International Conference on Computer Design (ICCD), 2021
- [ISLPED 2021] CoPIM: A Concurrency-Aware PIM Workload Offloading Architecture for Graph Applications
Liang Yan, Mingzhe Zhang, Rujia Wang, Xiaoming Chen, Xingqi Zou, **Xiaoyang Lu**, Yinhe Han, Xian-He Sun
In the Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED), 2021
- [ICCD 2020] APAC: An Accurate and Adaptive Prefetch Framework with Concurrent Memory Access Analysis
Xiaoyang Lu, Rujia Wang, Xian-He Sun
In the Proceedings of the 38th International Conference on Computer Design (ICCD), 2020

JOURNAL PUBLICATIONS

- [TCAD 2025] ProMiner: Enhancing Locality, Parallelism, and Offloading for Graph Mining on Processing-in-Memory Systems
Liang Yan, **Xiaoyang Lu**, Sheng Xu, Xiaoming Chen, Xingqi Zou, Yinhe Han, Xian-He Sun
IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2025

- [CAL 2025] Pyramid: Accelerating LLM Inference with Cross-Level Processing-in-Memory

Liang Yan, **Xiaoyang Lu**, Xiaoming Chen, Yinhe Han, Xian-He Sun

IEEE Computer Architecture Letters (CAL), 2025, 24(1): 121-124

- [JCST 2023] The Memory-Bounded Speedup Model and its Impacts in Computing

Xian-He Sun, **Xiaoyang Lu**

Journal of Computer Science and Technology (JCST), 2023, 38(1): 64-79

TEACHING EXPERIENCE

Guest Lecture

Jan 2022 – Present

Chicago, IL

Illinois Institute of Technology

- Spring 2025 CS 550 Advanced Operating Systems, “Data-Centric Optimizations for LLM”.
- Fall 2024 CS 546 Parallel and Distributed Processing, “Introduction of Parallel Processing”.
- Spring 2022 CS 570 Advanced Computer Architecture, “GPU Architectures”.

Teaching Assistant

Aug 2017 – May 2022

Chicago, IL

Illinois Institute of Technology

- Assisted in teaching five graduate courses at Illinois Institute of Technology, each with 9-60 students, covering topics such as Java Programming (CS 401), Software Engineering (CS 487), Advanced Operating Systems (CS 550), Parallel and Distributed Processing (CS 546), and Advanced Computer Architecture (CS 570).
- Developed and prepared comprehensive course materials, including laboratory experiments, lectures, exams, homework, and practice problems.
- Led weekly lab sessions and problem-solving discussions for groups of up to 30 students, enhancing their understanding and application of course materials.
- Supervised and guided students in final projects, provided detailed feedback, and graded exams and weekly homework assignments.

MENTORING EXPERIENCE

- 2025-Present Max Han, undergraduate student at UIUC, Hardware-Assisted OS Primitive.
- 2025-2026 Belthangady Akash Vi Narayana Pai, master student at Illinois Tech, Near Memory Processing.
- 2025-2026 Hongrui Huang, master student at Columbia University, Accelerator for LLM Serving.
- 2024-Present Lihan Hu, PhD student at University of Iowa, Infrastructure for Efficient LLM Serving.
- 2023-Present Haoran Geng, PhD student at University of Notre Dame, Architecture for Secure Memory.
- 2023-Present Vadim Biryukov, PhD student at Illinois Tech, Hardware Prefetcher for Data-Intensive Workloads.

ACADEMIC HONORS AND AWARDS

- 2024 DAC PhD Forum Travel Award
- 2024 Illinois Institute of Technology Computer Science Department Best Student Paper Award (2023-2024)
- 2024 Illinois Institute of Technology College of Computing Best Poster Award
- 2024 ASPLOS Student Travel Award
- 2023 Top 100 Chips Achievements (2022-2023)
- 2023 HPCA Student Travel Award
- 2015 New York University Scholarship
- 2015 Zhejiang University Excellent Bachelor Thesis Award

PROFESSIONAL SERVICE

Editor

- Guest Editor, IEEE Internet of Things Journal, Special Issue on Integrated Sensing, Memory, Communication, and Computation for Large-Scale AI Based IoT Systems, 2026.

Conference Committee Service

- The Chips To Systems Conference (DAC), Technical Program Committee Member, 2026
- The Conference on Machine Learning and Systems (MLSys), External Review Committee Member, 2026
- European Conference on Computer Systems (EuroSys), Shadow Program Committee Member, 2026
- IEEE Cloud Summit, Technical Program Committee Member, 2026
- IEEE International Conference on Computer Design (ICCD), Program Committee Member, 2025

Invited Reviewer for Journals & Transactions

- Device
- Future Generation Computer Systems (FGCS)
- IEEE Transactions on Computers (TC)
- IEEE Transactions on Consumer Electronics (TCE)
- IEEE Transactions on Industrial Informatics (TII)
- IEEE Transactions on Network Science and Engineering (TNSE)
- IEEE Transactions on Parallel and Distributed Systems (TPDS)
- Journal of Systems Architecture (JSA)
- Memories - Materials, Devices, Circuits and Systems (MEMORI)
- Microprocessors and Microsystems (MICPRO)
- Simulation: Transactions of the Society for Modeling and Simulation International