

# Xiaoyang Lu

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

## EDUCATION

### Illinois Institute of Technology

*Ph.D. in Computer Science, Department of Computer Science*

Chicago, IL

Aug 2017 – May 2024

### New York University

*M.S. in Computer Engineering, Department of Electrical and Computer Engineering*

New York, NY

Aug 2015 – May 2017

### Zhejiang University

*B.E. in Electronic Science and Technology*

Hangzhou, China

Aug 2011 – July 2015

## RESEARCH EXPERIENCE

### Research Assistant Professor

*Illinois Institute of Technology*

June 2024 – Present

Chicago, IL

- 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

*Illinois Institute of Technology*

Jan 2020 – May 2024

Chicago, IL

- 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

*Argonne National Laboratory*

May 2020 – Aug 2020

Lemont, IL

- 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

- [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 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 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 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
- **[JCST 2023]** The Memory-Bounded Speedup Model and its Impacts in Computing  
Xian-He Sun, Xiaoyang Lu  
Journal of Computer Science and Technology, 2023, 38(1): 64-79

## TEACHING EXPERIENCE

---

### Teaching Assistant

Aug 2017 – May 2022

*Illinois Institute of Technology*

Chicago, IL

- 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

### Guest Lecture

Jan 2022 – Present

*Illinois Institute of Technology*

Chicago, IL

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

## MENTORING EXPERIENCE

---

- 2023-Present Vadim Biryukov, PhD student at Illinois Tech, Hardware Prefetcher for Data-Intensive Workloads.
- 2023-Present Haoran Geng, PhD student at University of Notre Dame, Architecture for Secure Memory.
- 2024-Present Lihan Hu, PhD student at University of Iowa, Infrastructure for Efficient LLM Serving.
- 2025-Present Max Han, undergraduate student at UIUC, Hardware-Assisted OS Primitive.

## 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

## SERVICES

---

### Conference Technical Program Committee

- IEEE International Conference on Computer Design (ICCD), 2025

### Invited Reviewer for Journals & Transactions

- IEEE Transactions on Parallel and Distributed Systems (TPDS)
- IEEE Transactions on Industrial Informatics (TII)
- IEEE Transactions on Network Science and Engineering (TNSE)
- IEEE Transactions on Consumer Electronics (TCE)
- Journal of Systems Architecture (JSA)
- Future Generation Computer Systems (FGCS)
- Simulation: Transactions of the Society for Modeling and Simulation International