Xiaoyang Lu

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

Education

Illinois Institute of Technology — Chicago, IL
Ph.D. in Computer Science
Advisor: *Professor Xian-He Sun* — Aug 2017 – May 2024
Thesis: *Utilizing Concurrent Data Accesses for Data-Driven and AI Applications*

New York University — New York, NY
M.S. in Computer Engineering — Aug 2015 – May 2017

Zhejiang University — Hangzhou, China
B.E. in Electronic Science and Technology — Aug 2011 – July 2015

Research Experience

Research Assistant Professor — June 2024 – Present
Illinois Institute of Technology, Gnosis Research Center (GRC) — Chicago, IL

- Lead research efforts in data-centric system design, spanning memory performance and I/O modeling, memory-system optimization, processing-in-memory (PIM), and memory-centric architectures for emerging AI and HPC workloads.
- Lead architecture and memory-system research for UniMCC, an NSF-funded GRC project on cross-layer memory-centric computing, including performance modeling, memory-system optimization, and memory-centric architectures.
- Lead performance-modeling research for PORTUS, a DOE Genesis Mission project in collaboration with Pacific Northwest National Laboratory (PNNL), targeting performance portability of dynamic scientific workflows across heterogeneous HPC and cloud resources.
- Direct and co-advise Ph.D. research in memory-system performance modeling, data-movement optimization, PIM, scalable AI architectures, and LLM systems, including research collaborations across multiple institutions.

Research Assistant — Jan 2020 – May 2024
Illinois Institute of Technology — Chicago, IL

- Established a computer-architecture research direction centered on concurrency-aware memory systems, including cache management, prefetching, and memory-performance optimization.
- Developed analytical and learning-assisted techniques for memory-system optimization and extended concurrency-aware design principles to architectures for data-intensive workloads.

Research Aide — May 2020 – Aug 2020
Argonne National Laboratory — Lemont, IL

- Characterized and modeled the performance of disaggregated memory systems, focusing on memory-access bottlenecks, interference, and system predictability.

Conference Publications

Research Theme: Data-Centric System Design, spanning Performance Modeling · Memory Systems · Memory-Centric Computing · AI Acceleration

- Improving Data Reuse across Blocks for Efficient Block-Sparse Transformers on GPUs

  Lihan Hu, **Xiaoyang Lu**, Xian-He Sun, Peng Jiang

  In the Proceedings of the 35th International Conference on Parallel Architectures and Compilation Techniques (PACT), 2026. [**AI Acceleration**]

- HyPIM: Accelerating Hyperbolic Machine Learning via Processing-In-Memory

  Jinlin Wu, Ferit Ozdaban, Selvanathan Sadhasivam, **Xiaoyang Lu**, Yunhui Guo, Minxuan Zhou

  In the Proceedings of the 2nd IEEE Cross-Disciplinary Conference on Memory-Centric Computing (CCMCC), 2026. [**Memory-Centric Computing**] [**AI Acceleration**]

- 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. [**Performance Modeling**] [**Memory-Centric Computing**] [**AI Acceleration**]

- 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. [**Performance Modeling**] [**AI Acceleration**]

- 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 Conference (DATE), 2026. [**Memory Systems**]

- 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. [**Memory Systems**]

- 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. [**Memory Systems**]

- 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. [**Memory Systems**] [**Memory-Centric Computing**]

- 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. [**Memory Systems**] [**AI Acceleration**]

- 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. [**Memory Systems**]

- 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. [**Performance Modeling**] [**Memory Systems**]

- 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. [**Performance Modeling**] [**Memory Systems**]

- 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. [**Memory Systems**]

- 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. [**Memory-Centric Computing**]

- 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. [**Performance Modeling**] [**Memory Systems**]

Journal Publications

- 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. [**Memory Systems**] [**Memory-Centric Computing**]

- 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. [**Memory-Centric Computing**] [**AI Acceleration**]

- 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. [**Performance Modeling**]

Teaching Experience

Guest Lecturer Jan 2022 – Present

Illinois Institute of Technology Chicago, IL

- Spring 2025, CS 550 Advanced Operating Systems: "Data-Centric Optimizations for LLMs".
- Fall 2024, CS 546 Parallel and Distributed Processing: "Introduction to Parallel Processing".
- Spring 2022, CS 570 Advanced Computer Architecture: "GPU Architectures".

Teaching Assistant Aug 2017 – May 2022

Illinois Institute of Technology Chicago, IL

- Supported five graduate courses, including Introduction to Advanced Studies I (CS 401), Software Engineering (CS 487), Advanced Operating Systems (CS 550), Parallel and Distributed Processing (CS 546), and Advanced Computer Architecture (CS 570).
- Designed and prepared laboratory exercises, assignments, examinations, and supporting course materials, and led laboratory sessions and technical discussions for classes of up to 60 students.
- Mentored students on course projects and provided technical guidance and assessment across data structures, algorithms, object-oriented programming, systems, parallel computing, and computer architecture.

Invited Talks

- Toward Data-Centric Computing: Data-Movement Analysis, Concurrency-Aware Optimization, and Memory-Centric Architecture, Workshop on Extreme-Scale Storage and Analysis (ESSA), 2026.

Mentoring Experience

- 2026–Present, Anwar Benhnini, Ph.D. student at Illinois Institute of Technology; co-advised with Prof. Xian-He Sun; *memory-centric architectures for AI.*
- 2026–Present, Alexander Ichetovkin, Ph.D. student at Illinois Institute of Technology; co-advised with Prof. Xian-He Sun; *memory-system performance modeling.*
- 2026–Present, Xulu Chu, Ph.D. student at Oregon State University; co-advised with Prof. Wenqian Dong; *wafer-scale architectures for AI.*
- 2026–Present, Jie Ye, Ph.D. candidate at Illinois Institute of Technology; co-advised with Prof. Xian-He Sun; *data-movement optimization for disaggregated LLM serving.*
- 2023–Present, Vadim Biryukov, Ph.D. candidate at Illinois Institute of Technology; co-advised with Prof. Xian-He Sun; *CPU architecture and memory systems.*
- 2024–2026, Lihan Hu, Ph.D. candidate at the University of Iowa; mentored in collaboration with Prof. Peng Jiang; *sparsity-aware architectures for AI.*

- 2023–2026, Haoran Geng, Ph.D. candidate at the University of Notre Dame; mentored in collaboration with Prof. X. Sharon Hu; *memory-system optimization and processing-in-memory.*
- 2025–2026, Hongrui Huang, M.S. student at Columbia University; *hardware accelerator design.*
- 2025–2026, Belthangady Akash Vi Narayana Pai, M.S. student at Illinois Institute of Technology; *near-memory processing for AI.*
- 2025–2026, Max Han, undergraduate student at the University of Illinois Urbana-Champaign; *cache and memory-system optimization.*

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

- IEEE International Symposium on High Performance Computer Architecture (HPCA), Program Committee Member, 2027
- Chips To Systems Conference (DAC), Program Committee Member, 2026
- 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

- ACM Transactions on Architecture and Code Optimization (TACO)
- Device
- Expert Systems With Applications (ESWA)
- 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)
- Neurocomputing
- Simulation: Transactions of the Society for Modeling and Simulation International