

Xiaoyang Lu

xlu40@illinoistech.edu | xiaoyang-lu.github.io | Google Scholar | Chicago, IL

RESEARCH AREA & FOCUS

Computer Architecture & Systems, with a focus on **data-centric system design** for memory-intensive AI/HPC workloads.

- **Concurrency-Aware Memory-System Optimization:** Quantifying and exploiting memory-access concurrency to guide cache management, prefetching, and memory-hierarchy optimization [C7, C9, C10, C11, C13, C15].
- **Performance & I/O Modeling:** Modeling memory and data-movement costs to guide architecture and system decisions across AI/HPC workloads [C3, C4, C12, J3, P2].
- **Learning-Assisted System Optimization:** Using lightweight machine-learning methods, including reinforcement learning, to predict memory behavior and guide system optimization [C1, C6, C7, C10].
- **Data-Centric AI/HPC Architectures:** Reducing data movement through PIM/near-data execution and memory-centric system design for emerging AI workloads [C2, C3, C8, C14, J1, J2, P1].

ACADEMIC APPOINTMENTS

Research Assistant Professor

June 2024 – Present

Illinois Institute of Technology, Gnosis Research Center (GRC)

Chicago, IL

- Initiate and lead research projects in data-centric computer systems, spanning concurrency-aware memory-system optimization, performance modeling, and memory-centric architectures for AI and HPC workloads.
- Develop model-guided performance-matching methods that translate workload data-access and data-movement requirements into architecture, data-placement, resource-selection, and heterogeneous-execution decisions.
- Co-advise and mentor Ph.D. students across memory-centric and AI systems research, guiding projects from problem formulation and system design through evaluation and publication, including multi-institutional collaborations.

EDUCATION

Illinois Institute of Technology

Chicago, IL

Ph.D. in Computer Science

Aug 2017 – May 2024

Advisor: Professor Xian-He Sun

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 – Jul 2015

PRIOR RESEARCH EXPERIENCE

Research Assistant

Jan 2020 – May 2024

Illinois Institute of Technology

Chicago, IL

- Developed a computer-architecture research direction in concurrency-aware memory systems, spanning cache management, prefetching, and memory-performance modeling for data-intensive workloads.
- Developed analytical and learning-assisted methods that quantify memory-access concurrency and translate it into optimization decisions across the memory hierarchy.

Research Aide

May 2020 – Aug 2020

Argonne National Laboratory

Lemont, IL

- Characterized disaggregated-memory performance, focusing on memory-access bottlenecks, interference, and performance variability.

- **NSF SHF: Small - Optimization of Memory Architectures: A Foundation Approach (CCF-2008907).**
Role and contributions: Technical lead in developing concurrency-aware memory-performance modeling and optimization, including prefetching, cache partitioning, replacement, and holistic cache control.
Representative publications: [C10, C11, C12, C13, C15, J3].
- **NSF CSR: Medium - Towards a Unified Memory-Centric Computing System with Cross-Layer Support (CNS-2310422).**
Role and contributions: Technical lead in developing memory-centric systems and execution methods for data-intensive AI workloads, including sparse-matrix acceleration, I/O analysis for long-sequence attention, and PIM-based execution.
Representative publications: [C1, C3, C4, C9].
- **DOE Genesis Mission - PORTUS (Phase I; PNNL-led): Automatic Performance Portability for Distributed Scientific Workflows on the Genesis Mission Platform.**
Role and contributions: Co-lead GRC's workflow performance-modeling effort, developing dataflow-aware models and resource-matching methods to guide storage selection, data placement, and task parallelism across heterogeneous HPC and cloud resources, in collaboration with PNNL.

Submitted Research Proposal (Under Review)

- **NSF CSR - Semantics-Driven Proactive Memory-Centric Computing for Stateful AI.**
Co-PI; submitted September 2026.

CONFERENCE PUBLICATIONS

- | | |
|------------------------|--|
| PACT'26 [C1] | Lihan Hu, Xiaoyang Lu , Xian-He Sun, Peng Jiang. <i>Improving Data Reuse across Blocks for Efficient Block-Sparse Transformers on GPUs</i> . In the 35th International Conference on Parallel Architectures and Compilation Techniques (PACT), 2026. |
| CCMCC'26 [C2] | Jinlin Wu, Ferit Ozdaban, Selvanathan Sadhasivam, Xiaoyang Lu , Yunhui Guo, Minxuan Zhou. <i>HyPIM: Accelerating Hyperbolic Machine Learning via Processing-In-Memory</i> . In the 2nd IEEE Cross-Disciplinary Conference on Memory-Centric Computing (CCMCC), 2026. |
| IPDPS'26 [C3] | Xiaoyang Lu , Lihan Hu, Hongrui Huang, Peng Jiang, Xian-He Sun. <i>I/O-Aware PIM Acceleration for Long-Sequence LLM Inference with Hybrid Sparse Attention</i> . In the 40th IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2026. |
| ASPLOS'26 [C4] | Xiaoyang Lu , Boyu Long, Xiaoming Chen, Yinhe Han, Xian-He Sun. <i>I/O Analysis is All You Need: An I/O Analysis for Long-Sequence Attention</i> . In the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2026. |
| DATE'26 [C5] | Vadim Biryukov, Xiaoyang Lu , Zirui Liu, Kaixiong Zhou, Xian-He Sun. <i>Zion: A Comprehensive, Adaptive, and Lightweight Hardware Prefetcher</i> . In the Design, Automation and Test in Europe Conference (DATE), 2026. |
| MICRO'25 [C6] | Haoran Geng, Xiaoyang Lu , Yuezhi Che, Ziang Tian, Dazhao Cheng, Xian-He Sun, Michael Niemier, X. Sharon Hu. <i>COSMOS: RL-Enhanced Locality-Aware Counter Cache Optimization for Secure Memory</i> . In the 58th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2025. |
| GLSVLSI'25 [C7] | Yuping Wu, Xiaoyang Lu , Xiaoming Chen, Yinhe Han, Xian-He Sun. <i>Concurrency-Aware Cache Miss Cost Prediction with Perceptron Learning</i> . In the 35th ACM Great Lakes Symposium on VLSI (GLSVLSI), 2025. |
| ICCD'24 [C8] | Liang Yan, Xiaoyang Lu , Xiaoming Chen, Sheng Xu, Xingqi Zou, Yinhe Han, Xian-He Sun. <i>AceMiner: Accelerating Graph Pattern Matching using PIM with Optimized Cache System</i> . In the 42nd IEEE International Conference on Computer Design (ICCD), 2024. |
| ASPLOS'24 [C9] | Xiaoyang Lu , Boyu Long, Xiaoming Chen, Yinhe Han, Xian-He Sun. <i>ACES: Accelerating Sparse Matrix Multiplication with Adaptive Execution Flow and Concurrency-Aware Cache Optimizations</i> . In the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2024. |
| HPCA'24 [C10] | Xiaoyang Lu , Hamed Najafi, Jason Liu, Xian-He Sun. <i>CHROME: Concurrency-Aware Holistic Cache Management Framework with Online Reinforcement Learning</i> . In the 30th IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2024. |

HPCA'23 [C11]	Xiaoyang Lu , Rujia Wang, Xian-He Sun. <i>CARE: A Concurrency-Aware Enhanced Lightweight Cache Management Framework</i> . In the 29th IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2023.
WSC'22 [C12]	Hamed Najafi, Xiaoyang Lu , Jason Liu, Xian-He Sun. <i>A Generalized Model For Modern Hierarchical Memory System</i> . In the Winter Simulation Conference (WSC), 2022.
ICCD'21 [C13]	Xiaoyang Lu , Rujia Wang, Xian-He Sun. <i>Premier: A Concurrency-Aware Pseudo-Partitioning Framework for Shared Last-Level Cache</i> . In the 39th IEEE International Conference on Computer Design (ICCD), 2021.
ISLPED'21 [C14]	Liang Yan, Mingzhe Zhang, Rujia Wang, Xiaoming Chen, Xingqi Zou, Xiaoyang Lu , Yinhe Han, Xian-He Sun. <i>CoPIM: A Concurrency-aware PIM Workload Offloading Architecture for Graph Applications</i> . In the IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2021.
ICCD'20 [C15]	Xiaoyang Lu , Rujia Wang, Xian-He Sun. <i>APAC: An Accurate and Adaptive Prefetch Framework with Concurrent Memory Access Analysis</i> . In the 38th IEEE International Conference on Computer Design (ICCD), 2020.

JOURNAL PUBLICATIONS

TCAD'25 [J1]	Liang Yan, Xiaoyang Lu , Sheng Xu, Xiaoming Chen, Xingqi Zou, Yinhe Han, Xian-He Sun. <i>ProMiner: Enhancing Locality, Parallelism, and Offloading for Graph Mining on Processing-in-Memory Systems</i> . IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 44(12): 4577–4590, 2025.
CAL'25 [J2]	Liang Yan, Xiaoyang Lu , Xiaoming Chen, Yinhe Han, Xian-He Sun. <i>Pyramid: Accelerating LLM Inference with Cross-Level Processing-in-Memory</i> . IEEE Computer Architecture Letters, 24(1): 121–124, 2025.
JCST'23 [J3]	Xian-He Sun, Xiaoyang Lu . <i>The Memory-Bounded Speedup Model and Its Impacts in Computing</i> . Journal of Computer Science and Technology, 38(1): 64–79, 2023.

PREPRINTS

arXiv'26 [P1]	Xiaoyang Lu , Belthangady Akash Vi Narayana Pai, Xian-He Sun. <i>DynaNDE: Dynamic Near-Data Expert Scheduling for Batched MoE Inference</i> . arXiv preprint, 2026.
arXiv'26 [P2]	Haoran Geng, Tomas Sousa Pereira, Xiaoyang Lu , Xian-He Sun, Michael Niemier, X. Sharon Hu. <i>VIPER: Architecture-Aware Performance Modeling for Processing-in-Memory Design-Space Exploration</i> . arXiv preprint, 2026.

RESEARCH MENTORING

Ph.D. Co-Advising

- **Vadim Biryukov**, Ph.D. candidate, Illinois Institute of Technology, 2023–Present; co-advised with Prof. Xian-He Sun; *CPU architecture and memory systems*; 1× **DATE** [C5].
- **Anwar Benhnini**, Ph.D. student, Illinois Institute of Technology, 2026–Present; co-advised with Prof. Xian-He Sun; *memory-centric architectures for AI*.
- **Alexander Ichetovkin**, Ph.D. student, Illinois Institute of Technology, 2026–Present; co-advised with Prof. Xian-He Sun; *memory-system performance modeling*.
- **Xulu Chu**, Ph.D. student, Oregon State University, 2026–Present; co-advised with Prof. Wenqian Dong; *wafer-scale architectures for AI*.

Collaborative Ph.D. Mentoring

- **Haoran Geng**, Ph.D. candidate, University of Notre Dame, 2023–Present; mentored in collaboration with Prof. X. Sharon Hu; *memory-system optimization and processing-in-memory*; 1× **MICRO** [C6], 1 **preprint** [P2].
- **Lihan Hu**, Ph.D. candidate, University of Iowa, 2024–2026; mentored in collaboration with Prof. Peng Jiang; *sparsity-aware architectures for AI*; 1× **PACT** [C1], 1× **IPDPS** [C3].

M.S. and Undergraduate Mentoring

- **Hongrui Huang**, M.S. student, Columbia University, 2025–2026; *processing-in-memory for AI acceleration*; 1× IPDPS [C3].
- **Belthangady Akash Vi Narayana Pai**, M.S. student, Illinois Institute of Technology, 2025–2026; *near-memory processing for AI*; 1 preprint [P1].
- **Max Han**, undergraduate student, University of Illinois Urbana-Champaign, 2024–2026; *cache and memory-system optimization*.

TEACHING EXPERIENCE

Guest Lectures

Illinois Institute of Technology

- 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 computer science 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 laboratory exercises, assignments, examinations, and course materials; led laboratory sessions and technical discussions for classes of up to 60 students; and mentored student course projects.

Teaching Interests: Computer Architecture, AI/ML Systems, Operating Systems, and Parallel and Distributed Systems.

PROFESSIONAL SERVICE

Editorial Service

- 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 Service

- **Program Committee Member:** DAC’27, EuroSys’27, HPCA’27, DAC’26, IEEE Cloud Summit’26, ICCD’25.
- **External Review Committee Member:** MLSys’26.
- **Shadow Program Committee Member:** EuroSys’26.
- **Session Chair:** PACT’26.

Selected Journal Reviewing

- ACM TACO; IEEE TC, TPDS, TII, TNSE, TCE; JSA; MICPRO; FGCS; Neurocomputing; ESWA; Device; MEMORI; Simulation.

INVITED TALKS

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

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