Lei Zhao

PhD of Computer Science

AI Research Lab, Hewlett Packard Enterprise
3404 E Harmony Rd, Fort Collins, CO 80528, USA

(+1) 412-708-4514 | lei.zhao@hpe.com, leizhao0403@gmail.com | https://leizhaocs.github.io

Research and Expertise

Model & Optimization: Large Language Models (LLM), Convolutional Neural Networks (CNN), Quantization, Mixed-Precision, Efficient Inference, Model Deployment

Hardware & Architecture: Computer Architecture, AI Accelerators, Hardware-Software Co-design, Architecture Simulation, Performance and Energy Modeling, Memory Hierarchy, Emerging Memory Technology

Software & Framework: ML Compiler (MLIR), Dataflow and Mapping Optimization, Compute Graph Optimization, Runtime and Toolchain Integration

Skills

Programming C/C++, Python, CUDA, Chisel
Frameworks PyTorch, Hugging Face Transformers, MLIR, LLVM, Gem5, Timeloop/Accelergy

Education

University of Pittsburgh — *Pittsburgh, PA, U.S.*
PH.D. OF COMPUTER SCIENCE — *August 2014 - April 2022*
- Co-advised by Youtao Zhang and Jun Yang

Northwestern Polytechnical University — *Xi'an, Shaanxi, China*
MASTER OF COMPUTER SCIENCE — *August 2011 - April 2014*

Northwestern Polytechnical University — *Xi'an, Shaanxi, China*
BACHELOR OF SOFTWARE ENGINEERING — *August 2007 - July 2011*

Work Experience

Hewlett Packard Enterprise, USA — *Postdoc Research Scientist*
ARTIFICIAL INTELLIGENCE RESEARCH LAB — *August 2023 - Present*
- Architecture, algorithms, compiler and system design for ReRAM/CMOS hardware accelerators
- Project Lead of a full-stack ML accelerator for Transformers, responsibilities include:
  - Designed the architecture of an in-memory computing accelerator for Transformer models.
  - Optimized models and performed noise-aware fine-tuning for efficient deployment.
  - Developed and optimized an LLVM/MLIR-based compiler targeting the accelerator.
  - Conducted performance and energy modeling via simulation, analyzing accelerator efficiency and scalability.
- Current project status: The finalized architecture has been handed over to the circuit design team, and the tape-out is planned within the next two years.

Meta, Inc., USA — *Postdoc Research Scientist*
REALITY LABS — *Janurary 2022 - August 2023*
- Neural network model optimization and deployment on mobile hardware platforms.
- Cross-platform AR/VR demo system development (Windows, Linux, MacOS).
- ReRAM based in-memory computing accelerator design for graphics pipeline.
- Software simulation for graphics pipeline.
- Tools and skills: Pytorch, OpenGL(GLSL), OpenCV, Python, C/C++.

National Key Lab of Process Optimization and Intelligent Decision, Ministry of Education, China — *Research Intern*

SMART MEDICAL SYSTEM TEAM — *July 2018 - August 2018*

- Maintaining previously developed privacy-preserving neural network training framework.
- Convert and optimize DNN model into a privacy-preserving friendly fashion.

National Key Lab of Process Optimization and Intelligent Decision, Ministry of Education, China — *Research Intern*

SMART MEDICAL SYSTEM TEAM — *August 2017 - Septempber 2017*

- Developed a privacy-preserving neural network training framework in C++ with CUDA acceleration.

PhD Projects

Machine Learning Accelerators — *University of Pittsburgh*

PROJECT OWNER / PRIMARY RESEARCHER — *April 2017 - April 2022*

- Design ASIC accelerators based on existing or emerging memory technologies to improve performance and energy efficiency of machine learning computations with a focus on model security and user privacy.
- Key Achievements: Three conference papers in [ICCAD'17] [ICS'19] [DAC'20] and one U.S. patent.

Emerging Memory Design — *University of Pittsburgh*

PROJECT OWNER / PRIMARY RESEARCHER — *September 2014 - April 2016*

- Emerging memory system design for machine learning acceleration.
- Developed a cycle-accurate out-of-order architecture simulator based on MIPS ISA. Open sourced on github: https://github.com/leizhaocs/Monichi.
- Key Achievements: Four conference papers in [ICCD'15] [ISQED'17] [NVMSA'17] [ICCAD'17] and one journal paper in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD).

Publications

PhD Thesis

- **Lei Zhao**. (2022). Secure Accelerator Design for Deep Neural Networks. University of Pittsburgh.

Conference Proceedings

- **Lei Zhao**, Aishwarya Natarajan, Luca Buonanno, Archit Gajjar, Ron Roth, Sergey Serebryakov, John Moon, Omar Eldash, Jim Ignowski and Giacomo Pedretti. (2025). RACE-IT: A Reconfigurable Analog Computing Engine for In-Memory Transformer Acceleration. International Conference on Computer Design. (ICCD'25)
- London, Yanir, Luca Buonanno, Lei Zhao, Giacomo Pedretti, Bassem Tossoun, Stanley Cheung, Yuan Yuan, Yiwei Peng, Matej Hejda and Marco Fiorentino. (2025). Optical content-addressable memories for tree-based machine learning. AI and Optical Data Sciences VI
- Archit Gajjar, **Lei Zhao**, Omar Eldash, Aishwarya Natarajan, Xia Sheng, Giacomo Pedretti, Paolo Faraboschi, Jim Ignowski, Aman Aro and Luca Buonanno. (2025). Analog In-memory Computing Enhanced FPGA for High-Throughput and Energy-Efficient Acceleration. IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM'25)
- Archit Gajjar, **Lei Zhao**, Omar Eldash, Aishwarya Natarajan, Rand Jean, Xia Sheng, Giacomo Pedretti, Paolo Faraboschi, Jim Ignowski and Luca Buonanno. (2025). Enhancing FPGAs with Analog In-Memory Computing Macros. Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (ISFPGA'25)
- **Lei Zhao**, Luca Buonanno, Aishwarya Natarajan, Jim Ignowski and Giacomo Pedretti. (2024). Noise Aware Finetuning for Analog Non-Linear Dot Product Engine. ML with New Compute Paradigms (MLNCP) at NeurIPS 2024 (MLNCP workshop at NeurIPS 2024)
- Luca Buonanno, Giacomo Pedretti, **Lei Zhao**, Aishwarya Natarajan, Todd Richmond, John Moon, Rand Jean, Xia Sheng, Ron Roth and Jim Ignowski. (2024). Memristive Quaternary Content-Addressable Memories for

Implementing Boolean Functions. IEEE International Symposium on Circuits and Systems (ISCAS'24)

- **Lei Zhao**, Luca Buonanno, Ron M. Roth, Sergey Serebryakov, Archit Gajjar, John Moon, Jim Ignowski and Giacomo Pedretti. (2023). RACE-IT: A Reconfigurable Analog CAM-Crossbar Engine for In-Memory Transformer Acceleration. (arXiv:2312.06532)
- **Lei Zhao**, Jae-sun Seo, H. Ekin Sumbul, Edith Beigne, and Dawei Wang. (2023). ReARVR: A ReRAM-Based DNN Accelerator for Mobile Devices. Design Automation Conference (Poster Session). (DAC'23)
- **Lei Zhao**, Youtao Zhang, and Jun Yang. (2022). A DNN Protection Solution for PIM Accelerators With Model Compression. IEEE Computer Society Annual Symposium on VLSI. (ISVLSI'22)
- **Lei Zhao**, Youtao Zhang, and Jun Yang. (2022). SRA: A Secure ReRAM-Based DNN Accelerator. Design Automation Conference. (DAC'22)
- **Lei Zhao**, Youtao Zhang, and Jun Yang. (2021). Flipping Bits to Share Crossbars in ReRAM-BasedDNN Accelerator. International Conference on Computer Design. (ICCD'21)
- **Lei Zhao**, Youtao Zhang, and Jun Yang. (2020). SCA: A Secure CNN Accelerator for both Training and Inference. Design Automation Conference. (DAC'20)
- **Lei Zhao**, Quan Deng, Youtao Zhang, and Jun Yang. (2019). RFAcc: A 3D ReRAM Associative Array based Random Forest Accelerator. International Conference on Supercomputing. (ICS'19)
- **Lei Zhao**, Youtao Zhang, and Jun Yang. (2017). AEP: An Error-bearing Neural Network Accelerator for Energy Efficiency and Model Protection. International Conference On Computer Aided Design. (ICCAD'17)
- Wen Wen, **Lei Zhao**, Youtao Zhang, and Jun Yang. (2017). Speeding Up Crossbar Resistive Memory by Exploiting In-memory Data Patterns. International Conference On Computer Aided Design. (ICCAD'17)
- **Lei Zhao**, Youtao Zhang, and Jun Yang. (2017). Mitigating Shift-Based Covert-Channel Attacks in Racetrack Last Level Caches. Non-Volatile Memory Systems and Applications Symposium. (NVMSA'17)
- **Lei Zhao**, Lei Jiang, Youtao Zhang, Nong Xiao, and Jun Yang. (2017). Constructing Fast and Energy Efficient 1TnR based ReRAM Crossbar Memory. International Symposium on Quality Electronic Design. (ISQED'17)
- Xianwei Zhang, **Lei Zhao**, Youtao Zhang, and Jun Yang. (2015). Exploit Common Source-Line to Construct Energy Efficient Domain Wall Memory based Caches. International Conference on Computer Design. (ICCD'15)

Journal Articles

- Giacomo Pedretti, John Moon, Pedro Bruel, Sergey Serebryakov, Ron M. Roth, Luca Buonanno, Archit Gajjar, **Lei Zhao**, Tobias Ziegler, Cong Xu, Martin Foltin, Paolo Faraboschi, Jim Ignowski and Catherine E. Graves. (2024). X-TIME: accelerating large tree ensembles inference for tabular data with analog CAMs. IEEE Journal of Exploratory Solid-State Computational Devices and Circuits (JXCDC'24)
- Zijie Yue, Shuai Ding, **Lei Zhao**, Youtao Zhang, Zehong Cao, M. Tanveer, Alireza Jolfaei, and Xi Zheng. (2020). Privacy-preserving Time Series Medical Images Analysis Using a Hybrid Deep Learning Framework. ACM Transactions on Internet Technology.
- Wen Wen, **Lei Zhao**, Youtao Zhang, and Jun Yang. (2019). Exploiting In-memory Data Patterns for Performance Improvement on Crossbar Resistive Memory. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.

Patents

- **Lei Zhao**, Luca Buonanno, and Giacomo Pedretti. (2024). Analog content addressable memory for general computing. U.S. Patent, 18/651,218 (Allowed, to be issued).
- **Lei Zhao**, Giacomo Pedretti and Ron M Roth. (2024). Gray encoding for analog content addressable memory in general computing. U.S. Patent, 18/668,540 (Filed).
- Luca Buonanno, **Lei Zhao**, and Giacomo Pedretti. (2024). Analog computation of shift and add for dot product engines. U.S. Patent, 18/651,338 (Filed).
- **Lei Zhao**, Luca Buonanno, and Giacomo Pedretti. (2024). Programmable in-memory accelerator architecture for transformer models. U.S. Patent, 18/772,940 (Filed).

- **Lei Zhao**, Luca Buonanno, Archit Gajjar, Giacomo Pedretti, and Jim Ignowski. (2024). Recursive analog content addressable memory device. U.S. Patent, 18/933,512 (Filed).
- Giacomo Pedretti, **Lei Zhao**, and Luca Buonanno. (2025). Analog content addressable memory approximating activation functions. U.S. Patent, 19/195,605 (Filed).
- Luca Buonanno, Archit Gajjar, **Lei Zhao**, Aishwarya Natarajan, Omar Eldash, Giacomo Pedretti, Paolo, Faraboschi, and Jim Ignowski. (2025). Field programmable gate array with integrated analog in-memory computing dot product. U.S. Patent, 19/067,179 (Filed).
- Luca Buonanno, Giacomo Pedretti, **Lei Zhao**, and Jim Ignowski. (2024). Current-mode operation of analog content addressable memories. U.S. Patent, 18/914,938 (Filed).
- Luca Buonanno, **Lei Zhao**, Archit Gajjar, Giacomo Pedretti, and Jim Ignowski. (2025). An Inequality engine performing a recursive bitwise analysis. U.S. Patent, 19/352,926 (Filed).
- Youtao Zhang, **Lei Zhao**, and Jun Yang. (2022). System and method of deploying an artificial neural network on a target device. U.S. Patent, 11,531,877 B2.
- Shuai Ding, **Lei Zhao**, Shanlin Yang, Hao Wang and Zijie Yue. (2020). The architecture, image processing method and process of an artificial intelligence chip for medical endoscope. Chinese. Patent, CN108055454B.

Honors & Awards

2021	**Best Paper Candidate**, 39th International Conference on Computer Design (ICCD)	*Virtual Conference*
2017	**Best Paper Candidate**, 18th International Symposium on Quality Electronic Design (ISQED)	*Santa Clara, CA, USA*
2014	**Outstanding Master Degree Thesis**, Graduation Commencement of Northwestern Polytechnical University	*Xi'an, Shaanxi, China*