

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

Hardware & Architecture: Computer Architecture, In-Memory Computing, Emerging Memory Technology, AI Accelerator Design, Hardware-Software Co-design

Software & Algorithm: Deep Learning, ML Model Optimization and Deployment, Secure Machine Learning, Quantization

Education

University of Pittsburgh

PH.D. OF COMPUTER SCIENCE

- Co-advised by Youtao Zhang and Jun Yang

Pittsburgh, PA, U.S.

August 2014 - April 2022

Northwestern Polytechnical University

MASTER OF COMPUTER SCIENCE

Xi'an, Shaanxi, China

August 2011 - April 2014

Northwestern Polytechnical University

BACHELOR OF SOFTWARE ENGINEERING

Xi'an, Shaanxi, China

August 2007 - July 2011

Work Experience

Hewlett Packard Enterprise, USA

ARTIFICIAL INTELLIGENCE RESEARCH LAB

Postdoc Research Scientist

August 2023 - Present

- Architecture, Algorithms, and system design for ReRAM/CMOS hardware accelerators
- Analog CAM based in-memory computing accelerator design for Transformers

Meta, Inc., USA

REALITY LABS

Postdoc Research Scientist

January 2022 - August 2023

- ReRAM based in-memory computing accelerator design for graphics pipeline.
- AR/VR deep learning model optimization for in-house systolic array accelerator / SOC.
- AR/VR deep learning model deployment on SOC in existing head mounted devices.

National Key Lab of Process Optimization and Intelligent Decision, Ministry of Education, China

SMART MEDICAL SYSTEM TEAM

Research Intern

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

SMART MEDICAL SYSTEM TEAM

Research Intern

August 2017 - September 2017

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

Previous Projects

Machine Learning Accelerators

DESIGNER AND CONDUCTOR

University of Pittsburgh

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.
- Three conference papers in [ICCAD'17] [ICS'19] [DAC'20] and one U.S. patent.

Privacy-Preserving Medical Data Analysis

RESEARCH ASSISTANT

Hefei University of Technology

August 2017 - August 2018

- Evaluate Neural Networks on medical data with homomorphic encryption to protect patient's private information.
- One journal paper in ACM Transactions on Internet Technology (TOIT).

Privacy-Preserving Neural Network Framework

DESIGNER AND DEVELOPER

University of Pittsburgh

August 2017 - September 2017

- A neural network framework that targets at privacy-preserving and architecture research.
- Written in C++ and CUDA. Support fixed-point computing, stochastic-computing and homomorphic encryption.
- Open sourced on github: <https://github.com/leizhaocs/ArchNet>.

Emerging Memory Design

DESIGNER AND CONDUCTOR

University of Pittsburgh

September 2014 - April 2016

- Circuit level optimization of non-volatile memory 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>.
- 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**, 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**, Yuecheng Li, 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-Based DNN 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

- 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

- Youtao Zhang, **Lei Zhao**, and Jun Yang. (2019). **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.

Skills

Programming	C/C++, Python, Java, CUDA
Algorithms	Deep Learning, Reinforcement Learning, Homomorphic Encryption
Tools	Gem5, DRAMSim2, numpy, PyTorch, Homomorphic Encryption Libraries(SEAL, HEAAN), Linux/Unix

Honors & Awards

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