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, AI Accelerator Design, Hardware-Compiler-Algorithm Co-design, In-Memory Computing, Emerging Memory Technology,

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

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
- LLVM and MLIR based compiler design and optimization for LLM on in-memory computing accelerators
- Quantization exploration for LLM and Transformers
- Transformer accelerator design based on analog CAM in-memory computing.
- Tools and skills: Pytorch, LLVM, MLIR, Hugging Face, scikit-learn, Chisel, RISC-V, Python, C/C++.

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.

Previous Projects

Machine Learning Accelerators
University of Pittsburgh

Designer and Conductor — *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
Hefei University of Technology

Research Assistant — *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
University of Pittsburgh

Designer and Developer — *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
University of Pittsburgh

Designer and Conductor — *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

- 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

- 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) — *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*