Yayue Hou

Rensselaer Polytechnic Institute houy4@rpi.edu Github Linkedin

Research Interests

- Algorithms and Hardware co-design for **Machine Learning Systems**.
- **Compute-in-Memory**, including algorithm improvement and hardware efficiency enhancement.
- Noise resilient **Large Language Models** with new model architecture and training techniques.

Publications

DATE'25	**Yayue Hou**, Hsinyu Tsai, Kaoutar El Maghraoui, Tayfun Gokmen, Geoffrey W. Burr, Liu Liu. "NORA: Noise-Optimized Rescaling of LLMs on Analog Compute-in-Memory Accelerators." In Proceedings of Design, Automation and Test in Europe Conference, 2025
ICCAD'25	**Yayue Hou**, Zhenyu Liu, Garrett Gagnon, Hsinyu Tsai, Kaoutar El Maghraoui, Geoffrey W. Burr, Liu Liu. "SAGE: Saliency-Aware Grouping for Efficient Mapping of LLMs on Analog Compute-in-Memory." In Proceedings of International Conference on Computer-Aided Design, 2025.
ASPLOS'25	Zehao Fan, Yunzhen Liu, Garrett Gagnon, Zhenyu Liu, **Yayue Hou**, Hadjer Benmeziane, Kaoutar El Maghraoui, Liu Liu. "STARC: Selective Token Access with Remapping and Clustering for Efficient LLM Decoding on PIM Systems." In upcoming 2025 International Conference on Architectural Support for Programming Languages and Operating Systems.
Arxiv' 2025	Zehao Fan, Zhenyu Liu, Yunzhen Liu, **Yayue Hou**, Hadjer Benmeziane, Kaoutar El Maghraoui, Liu Liu. "Context-Aware Mixture-of-Experts Inference on CXL-Enabled GPU-NDP Systems."

Skills

Programming:	Python, Verilog, C/C++, CUDA
Tools & Frameworks:	PyTorch, Anaconda, Huggingface, AIHWKIT (Analog CIM), CACTI, Synopsys, Cadence, LaTeX.
Familiar areas:	In-Memory-Computing, Sparsity, Quantization, Parameter-Efficient Fine-Tuning, Hardware-Software Co-design.

Education

2023-present	**Rensselaer Polytechnic Institute**, NY.	Ph.D. student in Electrical Engineering
2019-2023	**Tongji University**, Shanghai.	B.Eng. in Microelectronics Science and Engineering

Research Experiences

2025-present **Efficient Architecture Design for LLM Inference.**

Keywords: KV cache, Heterogeneous Computing System, PIM

Summary: We do KV cache management with multiple techniques like page-wise sparsity and mixed-precision to reduce the computation and communication overhead when running LLM inference on PIM or heterogeneous computing systems.

2025-present **Resilient Model Fine-tuning on Analog Compute-in-Memory Devices.**

Keywords: Analog CIM, Fine-tuning, Parameter efficient fine-tuning, Analog foundation models.

Summary: We design a framework for LLM fine-tuning on Analog Compute-in-Memory devices, which considers limited endurance and high programming overhead of ACIM memory devices and flexibility of LLM deployment on different ACIM systems by doing data distribution aware training and parameter efficient fine-tuning.

2025-present **Efficient Compute-in-memory Heterogeneous System for LLMs and MoE.**

Keywords: Analog CIM, MoE, Heterogeneous CIM system

Summary: We apply a highly heterogeneous system integrating analog CIM, digital CIM, and conventional digital units. We design a management mechanism to balance workloads across different computing units and fully utilize bandwidth and compute resources. Based on dynamic activation model architecture like MoE, we are looking into efficient mapping and workload scheduling techniques.

2024-2025 **Noise-aware Analog Compute-in-Memory Algorithm-Hardware Co-design for LLMs.**

Keywords: Analog CIM, LLM, Post-training, ADC energy efficiency

Summary: We analyze the performance of LLMs under different Analog CIM noises and propose calibration techniques which increase model resilience and enable lower ADC working precision by managing the weight and activation data distribution of LLMs.