PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

Improving the Efficiency and Robustness of In-Memory Computing in Emerging Technologies

Xiaoxuan Wang
Duke University

Monday, February 20, 2023
10:45 AM • MSEE 112

Zoom: https://purdue-edu.zoom.us/j/96778669371 ~ Meeting ID: 9677866 9371

Abstract
Advanced computing systems have been a key enabler for the resounding success of computationally intensive artificial intelligence (AI) models, and computing efficiency has become a critical measurement for computing tasks. To achieve better efficiency, one promising approach is to utilize emerging nonvolatile memory technologies to build the AI accelerators. Resistive random-access memory (ReRAM) is one of the most promising emerging technologies featuring high density, low access energy, and the feasibility of realizing multi-level cells. Prior ReRAM-based processing-in-memory (PIM) designs have demonstrated their potential in performing vector-matrix multiplications (VMM) compared with pure CMOS architectures. However, the prior designs cannot achieve high efficiency in the new and appealing attention-based models, such as Transformer. Besides, the hardware non-idealities will degrade the inferencing accuracy of the in-memory computing system, especially when the hardware is approaching the endurance limit. To address the above issues, this talk will focus on improving the efficiency and robustness of in-memory computing. I will start with a case study on efficient ReRAM-based PIM design for Transformer to elaborate on key computing step optimization and function construction with in-memory logic. Next, I will highlight a systematic framework to mitigate the impact of device stochastic noise and uncover Pareto-optimal solutions for high-performance and energy-efficient PIM. In the end, I will discuss a structured stochastic gradient pruning method, which enables the endurance-aware ReRAM-based training process.

Bio
Xiaoxuan Yang is a Ph.D. candidate in Electrical and Computer Engineering at Duke University, under the supervision of Dr. Hai Helen Li and Dr. Yiran Chen. She received the B.S. degree in Electrical Engineering from Tsinghua University and the M.S. degree in Electrical Engineering from the University of California, Los Angeles (UCLA). Her research interests include emerging nonvolatile memory technologies, robustness and reliability enhancement in processing-in-memory designs, and hardware accelerators for deep learning applications. Her research work won Best Research Award at ACM SIGDA Ph.D. Forum at DAC and Third Place of ACM Student Research Competition at ICCAD. She is also selected as a Rising Star in EECS by UT Austin.

Host: Professor Jing Gao, jinggao@purdue.edu