Speaker 1: Dr. Priyadarshini Panda, Assistant Professor, Yale University, USA

Priya Panda is an assistant professor in the electrical engineering department at Yale University, USA. She received her B.E. and Master's degree from BITS, Pilani, India in 2013 and her Ph.D. from Purdue University, USA in 2019. During her Ph.D., she interned in Intel Labs where she developed large scale spiking neural network algorithms for benchmarking the Loihi chip. She is the recipient of the 2019 Amazon Research Award, 2022 Google Research Scholar Award, 2022 DARPA Riser Award, 2023 NSF CAREER Award, 2023 DARPA Young Faculty Award. She has also received the 2022 ISLPED Best Paper Award, 2022 IEEE Brain Community Best Paper Award and 2024 ASP-DAC Best Paper Award Nomination. Her research interests lie in Neuromorphic Computing, Spiking Neural Networks, and In-Memory Computing.

Speaker 2: Abhishek Moitra, Ph.D. Student, Yale University, USA

Abhishek Moitra received his B.E. degree in Electrical Engineering from Birla Institute of Technology and Science Goa, India in 2019. He is pursuing his Ph.D. in the Intelligent Computing Lab under the supervision of Prof. Priya Panda. His research works have been published in reputed journals such as IEEE TCAS-1, IEEE TCAD, IEEE TETCI and conferences such as DAC and DATE. His research interests involve hardware-algorithm co-design and co-exploration for designing robust and energy-efficient hardware architectures for deep learning tasks.

Title:

In-Memory Computing for Robust & Efficient Spiking Neural Networks: Opportunities and Challenges

Abstract:

Spiking Neural Networks (SNNs) are being actively researched as an energy efficient alternative to traditional artificial neural networks (ANNs). Compared to conventional ANNs, SNNs use temporal spike data and bio-plausible neuronal activation functions such as Leaky-Integrate Fire/Integrate Fire (LIF/IF) for data processing. Today, In-Memory Computing (IMC) architectures have been proposed to alleviate the "memory-wall bottleneck" prevalent in von-Neumann architectures. In this tutorial, we will talk about the key synergies between SNNs and IMC architectures. In that regard, we will describe our recent SNN-specific IMC accelerator: SpikeSim [1], which is an integrated circuit-architecture-system design framework that can perform realistic energy-latency-area-accuracy benchmarking. With SpikeSim [1], we discover previously overlooked insights around membrane potential overhead (in terms of area, memory size) due to repeated timestep processing. With the insights provided by the SpikeSim tool on the area-energy overheads, we will discuss efficient co-design strategies such as input-aware dynamic timestep inference [2] to lower the costs. Additionally, we will talk about the significance of non-idealities in IMC pertaining to SNNs. To overcome the impact of non-idealities, we will describe our recent works that showcase training-less mitigation strategies to overcome the non-idealities [3]. Finally, we will discuss the need to co-explore the network and IMC-peripheral circuit design space to achieve optimal performance. To this end, we will describe our tool XPert [4] that performs optimization-driven design space exploration for energy-efficient IMC implementations.

[1] Moitra, Abhishek, et al. "Spikesim: An end-to-end compute-in-memory hardware evaluation tool for benchmarking spiking neural networks." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2023).
Code: https://github.com/Intelligent-Computing-Lab-Yale/SpikeSim
[2] Y. Li, A. Moitra, T. Geller and P. Panda, "Input-Aware Dynamic Timestep Spiking Neural Networks for Efficient In-Memory Computing," *2023 60th ACM/IEEE Design Automation Conference (DAC)*, San Francisco, CA, USA, 2023.
Code: https://github.com/Intelligent-Computing-Lab-Yale/SEENN
[3] Bhattacharjee, Abhiroop, et al. "Examining the robustness of spiking neural networks on non-ideal memristive crossbars." *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*. 2022.
[4] A. Moitra, A. Bhattacharjee, Y. Kim and P. Panda, "XPert: Peripheral Circuit & Neural Architecture Co-search for Area and Energy-efficient Xbar-based Computing," *2023 60th ACM/IEEE Design Automation Conference (DAC)*, San Francisco, CA, USA, 2023.
Code: https://github.com/Intelligent-Computing-Lab-Yale/XPert