Analog In-Memory Computing for Deep Learning Inference

Abu Sebastian, IBM Research - Zurich, CH-8803 Rüschlikon, Switzerland

Dr. Abu Sebastian is a Distinguished Scientist and technical manager at IBM Research – Zurich. He is one of the technical leaders of IBM's research efforts towards next generation AI Hardware and manages the in-memory computing group at IBM Research - Zurich. He is the author of over 200 publications in peer-reviewed journals/conference proceedings and holds over 90 US patents. In 2015 he was awarded the European Research Council (ERC) consolidator grant and in 2020, he was awarded an ERC Proof-of-concept grant. He was an IBM Master Inventor and was named Principal and Distinguished Research Staff Member in 2018 and 2020, respectively. In 2019, he received the Ovshinsky Lectureship Award for his contributions to "Phase-change materials for cognitive computing". In 2023, he was conferred the title of Visiting Professor in Materials by University of Oxford. He is a distinguished lecturer and fellow of the IEEE.

Abstract: Deep neural networks (DNNs) are revolutionizing the field of artificial intelligence and are key drivers of innovation in device technology and computer architecture. While there has been significant progress in the development of specialized hardware for DNN inference, many of the existing architectures physically split the memory and processing units. This means that DNN models are typically stored in a separate memory location, and that computational tasks require constant shuffling of data between the memory and processing units – a process that slows down computation and limits the maximum achievable energy efficiency. Analog in-memory computing (AIMC) is a promising approach that addressing this challenge by borrowing two key features of how biological neural networks are realized. Synaptic weights are physically localized in nanoscale memory elements and the associated computational operations are performed in the analog/mixed-signal domain.

In the first part of the course, I will introduce AIMC based on non-volatile memory technology. The focus will be on the key concepts and the associated terminology. Subsequently, a multi-tile mixed-signal AIMC chip for deep learning inference will be presented. This chip fabricated in 14nm CMOS technology comprises 64 AIMC cores/tiles based on phase-change memory technology. It will serve as the basis to delve deeper into the device, circuits, architectural and algorithmic aspects of AIMC. Of particular focus will be achieving floating point-equivalent classification accuracy while performing the bulk of computations in the analog domain with relatively less precision. I will also present an architectural vision for a next generation AIMC chip for DNN inference. I will conclude with an outlook for the future.