ABU SEBASTIAN
IBM ZURICH

IMPRS STNS
curiosiTea

IN-MEMORY COMPUTING FOR DEEP LEARNING AND BEYOND

ABSTRACT

The rise of AI and in particular, deep learning (DL), is a key driver for innovations in computing systems. There is a significant effort towards the design of custom accelerator chips based on reduced precision arithmetic and highly optimized dataflow. However, the need to shuttle millions of synaptic weight values between the memory and processing units, remains unaddressed. In-memory computing (IMC) is an emerging computing paradigm that addresses this challenge of processor-memory dichotomy. Attributes such as synaptic efficacy and plasticity can be implemented in place by exploiting the physical attributes of memory devices such as phase-change memory (PCM).

In this talk, I will give a status update on where in-memory computing stands with respect to DL acceleration. I will present some recent algorithmic advances for performing accurate DL inference and training with imprecise IMC. I will also touch upon some system-level aspects and will present a world's first IMC compute core based on PCM fabricated in 14nm CMOS technology. I will also provide a brief overview of photonic in-memory computing that could facilitate unprecedented latency and compute density. Finally, I will present some applications of IMC that transcend conventional DL such as memory-augmented neural networks and spiking neural networks.

JULY 07, 2021
4:00 PM
ONLINE

MAX PLANCK INSTITUTE
OF MICROSTRUCTURE PHYSICS
WEINBERG 2 | 06120
HALLE (SAALE) | GERMANY