

Title: In-memory computing: The next frontier in deep learning acceleration?

Abstract:

The rise of AI and in particular deep learning is a key driver for innovations in computing systems. There is a significant effort towards the design of custom ASICs 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). It is shown that, using custom “additive noise training”, software equivalent accuracy deep learning inference is possible. Moreover, using a mixed-precision training approach, iso-accuracy training is also possible. The IMC approach can be easily extended to spiking neural networks and to also implement additional entities such as explicit associative memory in an efficient manner for memory augmented neural networks. I will also present deep learning demonstrations based on a first of its kind IMC compute core based on PCM integrated in 14nm CMOS technology. Finally, I will provide a brief overview of photonic in-memory computing that could facilitate unprecedented latency and compute density.

Short Bio: Abu Sebastian is a Distinguished Research Staff Member at IBM Research – Zurich. He was a contributor to several key projects in the space of storage and memory technologies and currently leads the research effort on in-memory computing at IBM Zurich. Dr. Sebastian is a co-recipient of the 2009 IEEE Control Systems Technology Award and the 2009 IEEE Transactions on Control Systems Technology Outstanding Paper Award. In 2013 he received the IFAC Mechatronic Systems Young Researcher Award for his contributions to the field of micro-/nanoscale mechatronic systems. 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 is an IBM Master Inventor since 2016. He 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".

Contact Information: ase@zurich.ibm.com