Tutorial 2: In-memory Computing for AI

Abu Sebastian, IBM Research - Zurich

The explosive growth in data-centric artificial intelligence related applications necessitates a radical departure from traditional von Neumann computing systems, which involve separate processing and memory units. In-memory computing is one such approach where certain computational tasks are performed in place in the memory itself. This is achieved by exploiting in tandem the physical attributes of the memory devices, its array-level organization and peripheral circuitry as well as the control logic. Any computational task that is realized within the confines of these three units could be called in-memory computing. However, the key distinction is that at no point during computation, the memory content is read back and processed at the granularity of a single memory element. Both charge-based as well as resistance-based memory devices are being explored for in-memory computing. In-memory computing can be applied both to reduce the computational complexity of a problem via analog computing as well as to reduce the amount of data being accessed by performing computations inside the memory arrays. In this tutorial, I will provide a broad overview of the key computational primitives enabled by these memory devices as well as their applications spanning scientific computing, signal processing, machine learning, deep learning and stochastic computing. I will conclude with a discussion on the challenges and new directions of research.