IN-MEMORY COMPUTING WITH ASSOCIATIVE MEMORIES — A CROSS-LAYER PERSPECTIVE

Prof. Xiaobo Sharon Hu, University of Notre Dame

Data transfer between processors and memory is a major bottleneck in improving application-level performance. This is particularly true for data intensive tasks such as many machine learning and security applications. In-memory computing, where certain data processing is performed directly in the memory array, can be an effective solution to address this bottleneck. Associative memory (AM), a type of memory that can efficiently “associate” an input query with appropriate data words/locations in the memory, is a powerful in-memory computing core. Nonetheless harnessing the benefits of AM requires cross-layer efforts spanning from devices and circuits to architectures and systems. In this talk, I will showcase representative cross-layer AM based design efforts. In particular, I will highlight how different non-volatile memory technologies (such as RRAM, FeFET memory and Flash) can be exploited to implement various AM functions (e.g., exact and approximate match, ternary and multi-bit data representation, and different distance functions). I will use several popular machine learning and other applications to demonstrate how they can profit from these different AM designs. End-to-end (from device to application) evaluations will be analyzed to reveal the benefits contributed by each design layer, which can serve as guides for future research efforts. More information about the event and the speaker:

https://www.ict.tuwien.ac.at/staff/taherinejad/MiM/next.html