Analog memory needs for AI – compute-in-memory, Multi-level bit-cell

Shimeng Yu, Georgia Tech

Analog multilevel memories are the enabling device technologies for hardware acceleration of artificial intelligence workloads. Compute-in-memory (CIM) is an emerging paradigm that addresses the memory-wall problem in the deep learning accelerator. In this short course lecture, we will survey the landscape of the emerging non-volatile memories that could serve the multi-bit synaptic weights including RRAM, PCM, NOR Flash, 3D NAND and FeFET. We will highlight the key device properties that are required for on-chip inference and/or training of deep neural network (DNN) models. We will use a multi-bit RRAM test vehicle to characterize the variability/reliability at array-level for inference. Then we will introduce an end-to-end benchmark framework DNN+NeuroSim to that is interfaced with Tensorflow/PyTorch to evaluate versatile device technologies for DNN inference. Compute-in-SRAM and digital TPU will be used as the baseline for the benchmark. Hybrid precision synapse that combines non-volatile memories with volatile capacitor is also presented to achieve in-situ training accuracy that is comparable with software. State-of-the-art CIM prototype chips will be surveyed. We will summarize the general design challenges in CIM chip design with regards to the device non-idealities, analog to digital conversion, and process variations. Finally, we will present a promising research direction of 3D monolithic integration that aims to place the weight memories (RRAM or FeFET) on back-end-of-line (BEOL) at the legacy node, while scaling the front-end-of-line (FEOL) logic transistors to the latest node.

Shimeng Yu is an associate professor of electrical and computer engineering at the Georgia Institute of Technology. He received the B.S. degree in microelectronics from Peking University in 2009, and the M.S. degree and Ph.D. degree in electrical engineering from Stanford University in 2011 and 2013, respectively. From 2013 to 2018, he was an assistant professor at Arizona State University.

Prof. Yu's research expertise is on the emerging non-volatile memories (e.g., RRAM, ferroelectrics) for different applications such as deep learning accelerator, neuromorphic computing, monolithic 3D integration, and hardware security.

Among Prof. Yu's honors, he was a recipient of the NSF Faculty Early CAREER Award in 2016, the IEEE Electron Devices Society (EDS) Early Career Award in 2017, the ACM Special Interests Group on Design Automation (SIGDA) Outstanding New Faculty Award in 2018, the Semiconductor Research Corporation (SRC) Young Faculty Award in 2019, and the ACM/IEEE Design Automation Conference (DAC) Under-40 Innovators Award in 2020, etc.

Prof. Yu is active in professional services. He served or is serving many premier conferences as technical program committee, including IEEE International Electron Devices Meeting (IEDM), IEEE Symposium on VLSI Technology, etc. He is a senior member of the IEEE.