Abstract of PhD Thesis

Non-Volatile Memory-Centric Computing Advances

Vivek Parmar (2018EEZ8161)

Recent advances in the domain of Artificial Intelligence have led to development of specialized hardware accelerators to facilitate high-speed computations to support data-intensive workloads. While dedicated AI hardware has been dominantly explored for cloud/enterprise applications, true benefits of AI can be realized by enabling low-power edge computing. For IoT (Internet of Things) devices with constrained area and power, performing high-precision computations becomes infeasible. Additionally, physical separation between the storage/memory unit and the processor, memory↔compute bottleneck causes a further limitation. To address these concerns, we present different possible use-cases/benefits of exploiting emerging NVM technology for advanced computing applications. An exhaustive study of NMC as well as IMC techniques is presented focusing on utilizing NVM devices. As a physical realization of the NMC concept for edge-computing the NVIA (non-volatile inference accelerator) architecture based on 22nm-MRAM (magneto-resistive random-access memory) has been explored both experimentally and through large scale simulations. Benefits of non-volatile inference with resilience to harsh operating conditions (800 Oe and 125 °C) has also been demonstrated using the MNIST dataset. On the application front, the concept has also been validated for the domain of Mixed-Reality workloads (such as eye segmentation and hand detection). Differential NVM in-memory-compute bitcells have been validated both experimentally and through large-scale simulations on multiple application workloads (Thermal Images, Fashion-MNIST, CIFAR-10, Visual Wake Words). In particular, we show DM-FeFET (differential mode ferro-electric field effect transistor) realized using 28nm HKMG technology, which exhibits excellent BER (bit-error rate) tolerance of up to 10^{-2} for both storage and IMC applications involving multi-bit precisions. A novel methodology is proposed for realizing few-shot learning application with binary precision and on 130nm-RRAM based IMC hardware and validated over the miniImageNet and ORBIT datasets. Further, IMC based realization of stochastic BNN (binarized neural networks) is proposed exploiting device variability. Macro-level realization including Analog and Digital periphery circuits for RRAM IMC was demonstrated using the open-source Skywater 130nm PDK.