PhD Viva Announcement

Research Scholar: Sandeep Kaur Kingra

Entry ID: 2016EEZ8070

Title: Exploring Computing Applications with Non-Volatile Memory

Advisor: Prof. Manan Suri

Abstract:

In conventional computing, intensive workloads dissipate most resources (time and energy) on data shuttling between isolated processing- and storage- blocks, leading to fundamental limitations such as the ``Memory Wall''. Promising solutions to eliminate these bottlenecks utilize concepts such as “In-Memory Computing/Near Memory Computing" (IMC/NMC), where computations are performed either in-situ or near the actual storage. We present an exhaustive study of new IMC techniques and IMC based application mapping based on emerging resistive non-volatile memory (NVM) devices. We propose a novel ``Simultaneous Logic in-Memory'' (SLIM) methodology wherein the bitcells are capable of implementing both Memory and Logic operations simultaneously in space (silicon) and time. We demonstrate novel non-stateful SLIM bitcells (1T-1R/2T-1R) and propose a detailed programming scheme, array level implementation, and controller architecture. To study the impact of the proposed SLIM approach for real-world implementations, we performed analysis for three applications: (i) Sobel Edge Detection, (ii) Binary Neural Networks-Multi layer Perceptron (BNN-MLP), and (iii) Keccak-f hash function. For edge-AI applications, we propose an IMC based low-precision deep neural network (DNN) implementation that utilizes oxide-based random access memory (OxRAM) devices. We experimentally demonstrate a dual-configuration stateful 2T-2R XNOR IMC bitcell using fabricated 1T-1R OxRAM arrays and analyze the trade-off in terms of circuit overhead, energy, and latency. Additionally, using the proposed 2T-2R bitcell, we present a fully-binarized XOR based In-Memory Similarity Search (IMSS) operation. It enables simultaneous match operation across multiple stored data vectors by performing analog column-wise XOR operation and summation to compute Hamming Distance (HD). We also propose an efficient hardware mapping methodology for vector matrix multiplication (VMM) using analog OxRAM based IMC technique. For implementing Quantized Neural Networks (QNNs), two key building blocks (CMOS based neurons and OxRAM-synaptic blocks) are experimentally demonstrated. Our evaluations indicate that emerging NVM based IMC architectures can attain significant improvement in system-level performance compared to conventional von Neumann computing systems.