Energy-Efficient Design Methodologies for ReRAM-based Deep Neural Network Accelerators on the Edge

Abstract

As most of the emerging applications (in machine learning, distributed sensing, IoT) are increasingly more data-centric, the energy and performance cost of data movement between memory and processing elements can easily dominate the overall system energy and become the bottleneck for system performance. This proposal aims to mitigate this issue by leveraging emerging resistive random access memory (ReRAM) devices. These devices have the potential to replace conventional memory technologies in high performance computing applications due to their high density, ultra-low energy consumption, and compatibility with conventional CMOS-based fabrication flows. Thus, unlike the conventional dynamic random access memory (DRAM), ReRAM can be integrated on the chip with the processing elements, dramatically reducing the energy cost of accessing off-chip DRAM. Furthermore, ReRAM devices can be configured in a crossbar array to implement deep neural network (DNN)-based machine learning algorithms with unprecedented energy efficiency. This approach is typically referred to as in-memory computing since the ReRAM devices not only store the data (i.e. weights of the neural network), but also perform multiplication and accumulation operations that are essential for DNNs.

Despite these significant advantages, system-level integration of ReRAM is challenging due to issues related to long term reliability, limited endurance, high device variability, and high sensitivity to temperature. In this collaborative effort, we will produce preliminary results to address these issues via both circuit-level (digital and analog) and algorithmic innovations. Specifically, PI Salman will develop design optimization methods to achieve thermally feasible integration of ReRAM with processing elements. These methods include efficient thermal analysis and optimization of various parameters such as ReRAM size/configuration, systolic array size, supply voltage, frequency, and placement of these elements on the chip. PI Stanacevic will focus on the analog in-memory computing mechanism of ReRAM to tolerate high variability with minimum overhead. PIs will work together to develop novel mapping techniques between DNN models and ReRAM based hardware to optimize utilization, performance while considering the on-chip device variation and reliability concerns of ReRAM devices.

PIs will have access to fabricated ReRAM devices and ReRAM-based crossbar chips through an existing research collaboration with Air Force Research Labs (AFRL) in Rome, NY. Thus, the proposed methodologies will be evaluated and tested on practical ReRAM chips. The broad objective of this seed grant proposal is to experimentally demonstrate the proposed methodologies and establish a long term, funded research collaboration with AFRL through future DoD-AFOSR Broad Agency Announcements.