Thesis Title: Investigation of In-Memory Computing Architectures for Edge AI Applications

Abstract: The memory wall problem is a fundamental bottleneck in modern computer architecture, where the increasing speed of processors has far outpaced the rate at which data can be transferred to and from memory. To investigate these challenges, we present a heterogeneous simulation framework, which models the performance of heterogeneous memory subsystems with systolic arrays for the simulation of various memory configurations, including hybrid SRAM-NVM setups, for different neural network workloads. Upon examination of the memory subsystem's impact on the accelerator, we present novel and energy-efficient In-Memory Computing (IMC) architectures designed to tackle the "memory wall problem" by performing multi-bit multiply and accumulate (MAC) operations directly within an SRAM-based IMC accelerator. The proposed architecture features a unique input and weight mapping strategy that removes the need for power-hungry Digital-to-Analog Converters (DACs), which also incorporates an innovative analog carry computation method that computes the final multi-bit product directly within the IMC macro, eliminating the latency and power consumption of external digital circuitry.

The study applies these notions to RRAM (Resistive Random-Access Memory), a form of non-volatile memory. This encompasses the creation of a variable bit-precision vector co-processor that incorporates an RRAM-based IMC unit alongside a RISC-V core. The system employs an innovative data mapping technique for vector-matrix multiplication, enhancing energy efficiency and executing multi-bit MVM in a single cycle, therefore reducing latency. These RRAM-based designs seek to deliver efficient and high-performance solutions for deep learning inference on edge devices.

In addition to deep learning, the thesis briefly explores applications in genomics by exploiting an RRAM-based IMC accelerator for DNA sequence analysis. This architecture addresses the high computational complexity of DNA sequencing by converting DNA into a binary format to enable in-memory bitwise XOR and AND operations, showcasing the versatility of IMC architectures beyond conventional machine learning tasks. This research presents a collection of specialized hardware accelerators and a simulation framework that jointly enhance the area of IMC by showcasing high-performance, energy-efficient solutions for intricate computational challenges.