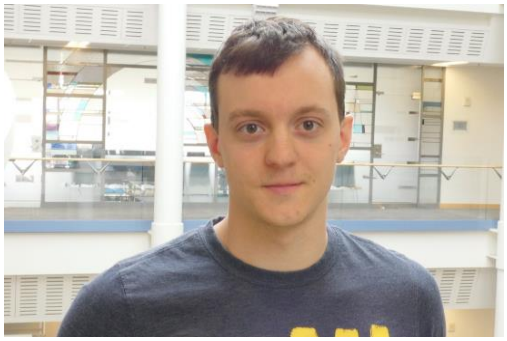
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE



Charles Eckert

In-SRAM Computing for Neural Network Acceleration

Monday, August 29, 2022
1:30 – 3:30pm
3725 Beyster

ABSTRACT: We propose using modified SRAM arrays to perform in-memory computing to accelerate neural networks. We focus on how existing processors can be repurposed to support in-memory computing while also designing a custom SRAM based ASIC to maximize the benefits from in-SRAM computing. Both the repurposed cache and custom ASIC focus on accelerating neural networks. For decades, the computing paradigm has been composed of separate memory and compute units. Processing-in-Memory(PIM) has often been proposed as a solution to break past the memory wall. With PIM, compute logic is moved near the memory, which can reduce the data movement. In-memory computing expands on PIM by morphing the memory into hybrid memory compute units where data can be stored where it is computed on. Recent works have modified SRAM arrays to allow logical operations to be performed directly inside the arrays. Our work extends basic logical operations and additionally adds support for arithmetic operations. Coinciding with the rise of increasing memory on chip and more focus on near and in-memory computing, is the ascendance of neural networks. Neural Networks are highly data parallel applications that are challenging to accelerate due to being memory bound, compute bound or both. In-memory computing can be used to help alleviate both compute and memory bottlenecks. Computing on data in place can alleviate the memory and compute bottlenecks in multiple ways. In-situ computation reduces on-chip data movement, the total amount of compute available is increased with the passive cache now transformed, and a hybrid memory compute allows for more SRAM units on chip.

CHAIR: Prof. Reetuparna Das