COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

DAICHI FUJIKI

In-Memory Acceleration for General Data Parallel Applications

Friday, January 28, 2022
12:00pm – 1:00pm
3725 BBB
Virtual
Passcode: 984409
Meeting ID: 947 2097 6063

ABSTRACT: Processing-in-Memory (PIM) has long been an attractive idea that has the potential to break the well-known memory wall problem. PIM moves compute logic near the memory, and thereby reduces data movement. In contrast, certain memories have been shown that they can morph themselves into compute units by exploiting the physical properties of the memory cells, making them intrinsically more efficient than PIM. Modern computing systems devote a large portion of aggregate die area for passive memories; thus, re-purposing them for active computing units brings substantial benefits. However, prior work has only provided low-level interfaces for computation or relied on a manual mapping of machine learning kernels to the compute-capable memories.

In this talk, I will demonstrate how in-memory computing can extend the compute capability and make itself applicable to a wide range of data-parallel applications. First, I will present a processor architecture that re-purposes resistive memory to support data-parallel in-memory computation. The proposed execution model seeks to expose the available parallelism in a memory array by supporting a programming model that merges the concepts of data-flow and vector processing. This is empowered by a compiler that compiles Data Flow Graphs of tensor programs. Second, I will present Duality Cache architecture that flexibly transforms caches into an in-memory accelerator that can execute arbitrary data-parallel programs. The proposed architecture adopts the SIMT execution model and uses CUDA/OpenACC framework as the programming frontend. Finally, I will present our efforts to multi-layer in-memory computing. In-memory computing can be implemented across multiple layers of the memory hierarchy, and we propose a framework that determines the appropriate level of memory hierarchy for in-memory computing and maximizes resource utilization.

CHAIR: Prof. Reetuparna Das