COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Hiwot Kassa

Efficient utilization of heterogeneous compute and memory systems

Tuesday, August 2, 2022
9:00 – 11:00am
Virtual Event
Zoom – Passcode: 817580

ABSTRACT: Conventional compute and memory systems scaling to achieve higher performance and lower cost and power have diminished. Concurrently, we have diverse compute and memory demanding workloads that continue to grow and stress traditional systems with only CPUs and DRAM. Heterogeneous compute and memory systems establish the opportunity to boost performance for these demanding workloads by providing hardware units with specialized characteristics. However, these systems have unique characteristics compared to traditional systems, and we must carefully design how workloads utilize these units to harness their full benefits. This dissertation presents software, system, and hardware techniques that maximize heterogeneous systems' performance, energy, and cost-efficiency based on the compute and memory access patterns of various application domains.

First, this thesis proposes ChipAdvisor, a machine learning-based framework, to identify the best platform for an application. ChipAdvisor considers the intrinsic characteristics of applications such as parallelism, locality, and synchronization patterns and archives 98% accuracy in predicting the best performant and energy-efficient platform for diverse workloads when considering a system with CPU, GPU, and FPGA. Second, we propose a heterogeneous memory-enabled system design with DRAM and storage class memory (SCM) for key-value stores, one of the largest workloads in data centers. We characterize an extensive deployment of key-value stores in a commercial data center and design optimal server configurations with heterogeneous memories to increase performance by 80% while reducing the total cost of ownership (TCO) by 43-48%.

Third, this dissertation designs MTrainS, an end-to-end recommendation system trainer that utilizes heterogeneous compute and memory systems. MTrainS efficiently divides recommendation model training tasks between CPUs and GPUs, and it hierarchically utilizes various memory types, such as HBM, DRAM, and SCMs, by studying the temporal locality and bandwidth requirements of recommendation system models in data centers and reduces the number of hosts used for training by up to 8×. Last, this dissertation proposes CoAct, fine-grain cache, and memory sharing techniques for collaborative workloads running in integrated CPU-GPU systems. CoAct uses the collaborative pattern of applications to fine-tune cache partitioning and interconnect and memory controller utilization for CPU and GPU, improving performance by 25%.

CHAIR: Prof. Ronald Dreslinski