COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Siying Feng

Acceleration Techniques of Sparse Linear Algebra on Emerging Architectures

Thursday, October 27, 2022
9:00 – 11:00am
Hybrid Event
3725 BBB / Zoom

ABSTRACT: Recent years have witnessed a tremendous surge of interests in accelerating sparse linear algebra applications. Sparse linear algebra is a fundamental building block and usually the performance bottleneck of a wide range of applications, such as machine learning, graph processing and scientific computing. The key challenge of sparse linear algebra lies in the irregular access pattern induced by the sparseness nature, which renders the deep cache hierarchy in general-purpose processors useless and makes sparse linear algebra applications notoriously memory intensive. This dissertation proposes to optimize the performance and efficiency of sparse linear algebra kernels by taking advantage of emerging architecture techniques, including hardware specialization, architecture reconfiguration, and near-memory processing.

This dissertation first proposes Transmuter, a reconfigurable architecture that features massively-parallel PEs and a reconfigurable on-chip memory hierarchy that can adapt to different applications. Transmuter demonstrates significant efficiency gains over the CPU and GPU across a diverse set of commonly-used kernels while offering GPU-like programmability. More importantly, Transmuter retains high performance for sparse linear algebra kernels, achieving an energy efficiency within 4.1× compared to state-of-the-art functionally-equivalent accelerators.

Next, this dissertation further improves the performance of Transmuter on Sparse Matrix-Vector Multiplication (SpMV) and graph analytics using CoSPARSE. CoSPARSE is an intelligent framework that judiciously reconfigures to the best-performing software algorithm and hardware configuration on-the-fly based on the input characteristics. The synergistic software and hardware reconfiguration amass a net speedup up to 2.0x, over a naive baseline implementation with no software or hardware reconfiguration.

The algorithm reconfiguration of CoSPARSE requires either a fast runtime sparse matrix transposition implementation or twice the storage overhead. Therefore, the final part of this dissertation presents MeNDA, a scalable near-memory multi-way merge accelerator for sparse matrix transposition and sparse merging dataflows. The wide application of multi-way merge sorting allows MeNDA to be easily adapted to other sparse primitives such as SpMV. Compared to two state-of-the-art implementations of sparse matrix transposition on a CPU and a sparse library on a GPU, MeNDA achieves a speedup of 19.1x, 12.0x, and 7.7x, respectively, while incurring a power overhead of 78.6 mW per DRAM rank.

CHAIR: Prof. Ronald Dreslinski