CECS
CENTER FOR EMBEDDED & CYBER-PHYSICAL SYSTEMS
UNIVERSITY OF CALIFORNIA · IRVINE

CECS Seminar



"Software-Defined Accelerator-Centric Systems - A Journey to AI and Big Data Acceleration"

Dr. Zhenman Fang

Associate Professor, University of Minnesota, Twin Cities

Wednesday, July 29th
11:00-12:00 p.m.
Location: ISEB 4020

Abstract: With the end of CPU scaling due to dark silicon limitations, customizable hardware accelerators have gained increasing attention in modern datacenters and edge devices, due to their low power, high performance, and high energy-efficiency. Unfortunately, not everyone can design hardware accelerators and there remains a substantial gap for the large community of software developers to effectively leverage the benefits of domain-specific hardware acceleration.

In this talk, I will share my research on software-defined accelerator-centric systems. First, I will give an overview of the unique advantages of hardware accelerators using the roofline model and illustrate common hardware customization techniques that can be defined by software. Next, I will share my research journey to AI and big data acceleration, and demonstrate two successful case studies of accelerator-centric systems: 1) HeatViT, a hardware-efficient image-adaptive token pruning framework for efficient yet accurate vision transformer (ViT) acceleration on edge FPGAs, with a key focus on algorithm/hardware codesign; and 2) SQL2FPGA, a compiler that automatically compiles big data queries onto datacenter FPGA accelerators, with a key focus on improving software programmability. Finally, I will discuss future research directions of software-defined accelerator-centric systems.

Biography: Dr. Zhenman Fang is an Associate Professor in Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. He is on leave from School of Engineering Science, Simon Fraser University, Canada, where he joined since Apr 2019. Zhenman founded and directs the HiAccel lab. Zhenman's recent research focuses on customizable computing with software-defined hardware acceleration, which aims to sustain the ever-increasing performance and energy-efficiency demands of important application domains in post-Moore's law era. It spans the entire computing stack, including emerging application characterization and acceleration (including machine learning, big data analytics, high-performance computing, and precision medicine), novel accelerator-rich and near-data computing architecture designs, and corresponding programming, runtime, and tool support.

Zhenman has published over 80 papers in top-tier conferences and journals and holds two US patents, including the ICCAD 10 Year Retrospective Most Influential Paper Award in 2025, three best paper awards (FPL 2024 Stamatis Vassiliadis Best Paper, TCAD 2019 Donald O. Pederson Best Paper, and MEMSYS 2017 Best Paper), four best paper nominees (ICCAD 2025, FCCM 2025, HPCA 2017 and ISPASS 2018), and an invited paper from Proceedings of the IEEE 2019. His research has also been recognized with the inaugural SFU Research Excellence Award - Horizon Award (2025), a NSERC (Natural Sciences and Engineering Research Council of Canada) Alliance Award (2020), a CFI JELF (Canada Foundation for Innovation John R. Evans Leaders Fund) Award (2019), a Xilinx University Program Award (2019), a Team Award from Xilinx Software and IP Group (2018), and a Postdoc Fellowship from UCLA Institute for Digital Research and Education (2016-2017). Zhenman has also actively served in the organizing and program committee of top-tier conferences in the areas of FPGA and reconfigurable computing, design automation, and computer architecture. Recently, he has served as Program Vice Chair of FCCM 2026, Journal Track Chair of FPL 2026, General Chair of ASAP 2025, Program Chair of RAW 2024 and RAW 2025. More details can be found in his website: https://zhenman.github.io/.

Hosted By: Prof. Sitao Huang