Department of Electrical Engineering
香港城市大學
City University of Hong Kong

SIMON FRASER UNIVERSITY

CityU Architecture Lab for Arithmetic and Security (CALAS) Seminar Series

Software-Programmable Accelerator-Centric Systems

Dr. Zhenman Fang, Assistant Professor, Computer Engineering
Simon Fraser University

Abstract: With the end of CPU scaling due to dark silicon limitations, customizable hardware accelerators on FPGAs have gained increasing attention in modern datacenters and edge devices due to their low power, high performance and energy-efficiency. Evidenced by Microsoft's FPGA deployment in its Bing search engine and Azure cloud, the public cloud access by Amazon and Alibaba, Intel's US$16.7B acquisition of Altera, and AMD's US$50B acquisition of Xilinx, FPGA-based customizable acceleration is considered one of the most promising approaches to sustain the ever-increasing performance and energy-efficiency demand of emerging application domains such as machine learning and big data analytics.

In this talk, Dr. Fang will first explain how FPGA hardware accelerators achieve amazing improvements and give an overview of our research on software-programmable accelerator-centric systems [PIEEE 2019, TRETS 2021]. Following that, he will present a few successful case studies on software-defined hardware acceleration for machine learning and big data analytics, including 1) Caffeine for early-day CNN acceleration [TCAD 2019 best paper], 2) HeatViT for vision transformer pruning and acceleration [HPCA 2023], 3) CHIP-KNN for k-nearest neighbors acceleration [FPT 2020, TRETS 2023], and 4) SQL2FPGA for compiling Spark SQL onto FPGAs [FCCM 2023].

Biography: Dr. Zhenman Fang is a Tenure-Track Assistant Professor in School of Engineering Science, Computer Engineering Option, Simon Fraser University, Canada, where he founded and directs the HiAccel lab. His recent research focuses on customizable computing with specialized hardware acceleration, which aims to sustain the ever-increasing performance, energy-efficiency, and reliability demand 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, computational genomics, big data analytics, and high-performance computing), novel accelerator-rich and near-data computing architecture designs, and corresponding programming, runtime, and tool support. Dr. Fang has published over 50 papers in top conferences and journals and two US patents, including two best paper awards (TCAD 2019 Donald O. Pederson best paper award and MEMSYS 2017), two best paper nominees (HPCA 2017 and ISPASS 2018), and an invited paper from Proceedings of the IEEE 2019. His research has also been recognized with 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). More details can be found in his personal website: https://www.sfu.ca/~zhenman/.

15 Nov 2023 (Wed), 10:00am - 12:00nn, P1402, https://cityu.zoom.us/j/96742093029