北京大学高能效计算与应用中心学术报告

Invited Talk, Center for Energy-Efficient Computing and Applications

SMART AND FAST: EFFICIENT MACHINE LEARNING GREETS AGILE HARDWARE DESIGN

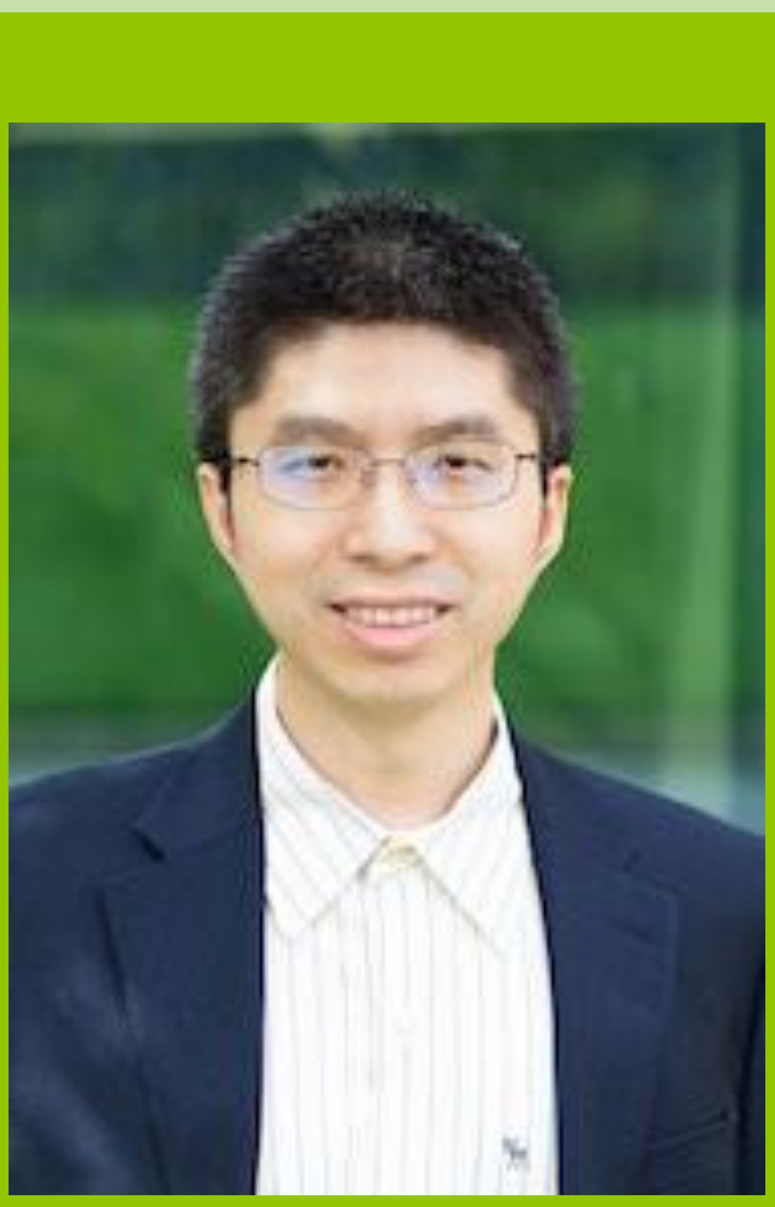
Associate Professor Zhiru Zhang

ECE，Cornell University

Host：梁云 长聘副教授

2019年8月19日 星期一 10:00am

理科五号楼 410会议室



ABSTRACT: The growing complexity of machine learning (ML) models, the proliferation of edge devices, and the diminished benefits of technology scaling together place strong demands on continued improvements in performance and energy efficiency of computer hardware. In line with this trend, deep neural network (DNN) processing is shifting from general-purpose CPUs/GPUs to specialized architectures in both academic and commercial settings, and there has been a active body of research into hardware-friendly DNN optimization. Increasing specialization also motivates the deployment of new high-level design languages and ML-aided automation tools to implement hardware accelerators in a much more productive and agile manner.

This talk introduces our recent research results from three related projects that navigate the intersection between ML and hardware design: (1) Channel gating, a dynamic, fine-grained, and trainable technique for DNN acceleration. Unlike static network pruning, channel gating exploits input-dependent dynamic sparsity at run time. This results in a significant reduction in compute cost with a minimal impact on accuracy. (2) HeteroCL, a new open-source programming infrastructure for accelerator-rich computing with decoupled algorithm and compute/data customization, and support for mixed declarative and imperative code. (3) PRIMAL, an ML-based power inference framework that enables fast and accurate power estimation for reusable ASIC IPs. PRIMAL is on average 50x faster than the best commercial gate-level power analysis tool, with an average error less than 5%.

BIOGRAPHY: Zhiru Zhang is an Associate Professor in the School of ECE at Cornell University. His current research investigates new algorithms, design methodologies, and automation tools for heterogeneous computing. His research has been recognized with a Google Faculty Research Award (2018), the DAC Under-40 Innovators Award (2018), the Rising Professional Achievement Award from the UCLA Henry Samueli School of Engineering and Applied Science (2018), a DARPA Young Faculty Award (2015), the IEEE CEDA Ernest S. Kuh Early Career Award (2015), an NSF CAREER Award (2015), the Ross Freeman Award for Technical Innovation from Xilinx (2012), a Best Paper Award from FPGA (2019), a Best Short Paper Award from FCCM (2018), a Best Paper Award from the ACM Transactions on Design Automation of Electronic Systems (2012), and three best paper nominations (ICCAD'2009, FPGA'2017, FPGA'2018). On the teaching side, he received the Michael Tien'72 Excellence in Teaching Award from College of Engineering in 2016.