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

Imperial College London

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

Fast Deep Learning for Scientific Applications with FPGAs

Dr. Zhiqiang QUE, Research Associate
Department of Computing, Imperial College London

Abstract:

In the domain of scientific research, particularly in fields like particle physics, the demand for rapid data acquisition and in-situ processing systems is critical. These systems rely on custom processing elements with very low latency and high data bandwidth, along with real-time control modules. Integrating real-time machine learning algorithms with these processes can enable advances in scientific discovery. A critical component of such integrations is the acceleration of deep learning inference using reconfigurable accelerators such as FPGAs, which enables sophisticated processing in real-time with superior accuracy. In this talk, I first describe the FPGA-based fast Graph Neural Networks (GNNs) tailored for particle physics applications, demonstrating our achievements in minimizing latency and maximizing throughput. I then present our new studies on automation of optimizations for Fast Deep Learning (FastDL) in scientific applications.

Biography:

Dr. Zhiqiang QUE is a research associate in the Custom Computing Research Group in the Department of Computing at Imperial College London. His experience includes ARM CPU design at Marvell Semiconductor (2011-2015) and Low Latency FPGA systems at CFFEX (2015-2018). He earned his PhD under Prof. Wayne LUK at Imperial College in 2023, while completing his B.S. and M.S. at Shanghai Jiao Tong University in 2008 and 2011. His research focuses on computer architecture, embedded systems, high-performance computing, and design automation for hardware optimization.

13 Mar 2024 (Wed); 4:00pm - 6:00pm; P1402; https://cityu.zoom.us/j/96742093029