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

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

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

The Next Wave of HLS: Fully Automated PyTorch-to-Accelerator Design Flow

Prof. Deming Chen, Abel Bliss Professor

University of Illinois Urbana-Champaign (UIUC), USA

ACM TRETS, Editor-in-Chief, IEEE Fellow

Abstract:

In this talk, we introduce a new design flow, ScaleHLS, that established a new High-Level Synthesis (HLS) solution translating AI models described in PyTorch to customized AI accelerators automatically. By adopting PyTorch as input for AI designs (instead of traditional C/C++ for HLS), the lines of code and design simulation time can be reduced by about 10× and 100×, respectively. Meanwhile, despite being fully automated and able to handle various applications, this new flow achieves a 1.29x higher throughput over DNNBuilder, a state-of-the-art RTL-based neural network accelerator on FPGAs. Such AI model-to-RTL flows pave the way for a new wave of HLS that could drive the high-productivity designs of AI circuits with high density, high-energy efficiency, low cost, and short design cycle. And such high-level model-to-RTL flows can be expanded to other non-AI domains. However, we are also facing existing and new challenges for such HLS solutions, such as ensuring the correctness of the high-level design, accommodating accurate low-level timing/energy information, handling the complexity of 3D circuits and/or chiplet-based design flows, and achieving all these in a highly scalable manner.

Biography:

Deming Chen is the Abel Bliss Professor of the Grainger College of Engineering at University of Illinois at Urbana-Champaign (UIUC). His current research interests include reconfigurable and heterogenous computing, hybrid cloud, system-level design methodologies, machine learning and acceleration, and hardware security. He has published more than 280 research papers, received 10 Best Paper Awards and one ACM/SIGDA TCFPGA Hall-of-Fame Paper Award, and given more than 150 invited talks. His research has generated high impact, with open-sourced solutions adopted by both academia and industry (e.g., FCUDA, DNNBuilder, CSRNet, SkyNet, ScaleHLS). He is an IEEE Fellow, an ACM Distinguished Speaker, and the Editor-in-Chief of ACM Transactions on Reconfigurable Technology and Systems (TRETS). He is the Director of the AMD-Xilinx Center of Excellence and the Hybrid-Cloud Thrust Co-Lead of the IBM-Illinois Discovery Accelerator Institute at UIUC. He has been involved in several startup companies, such as AutoESL and Inspirit IoT. He received his Ph.D. from the Computer Science Department of UCLA in 2005.

7 June 2024 (Fri); 11:00am – 12:00nn HKT; P1402; https://cityu.zoom.us/j/96742093029