SEMINAR NOTICE

Preston M. Green Department of Electrical and Systems Engineering

Early-Stage Design Space Exploration Tool for Neural Network Inference Accelerators

MS Dissertation Defense

By

Liu Ke

MS Candidate

Preston M. Green Department of Electrical and Systems Engineering
Washington University in St. Louis

Abstract: Deep neural networks (DNNs) have achieved spectacular success in recent years. In response to DNN's enormous computation demand and extensive memory footprint, numerous inference accelerators have been proposed. However, the diverse nature of DNNs, both at the algorithm level and the parallelization level, makes it difficult to arrive at an "one-size-fits-all" hardware implementation. In this paper, we develop NNest, an early-stage design space exploration tool that can speedily and accurately estimate the area/performance/energy of DNN inference accelerators based on high-level network topology and architecture traits, without the need for low-level RTL codes. Equipped with a generalized spatial architecture framework, NNest is able to perform fast high-dimensional design space exploration across a wide spectrum of architectural/microarchitectural parameters. Our proposed novel date movement strategies and multi-layer fitting schemes allow NNest to more effectively exploit parallelism inherent in DNN. Results generated by NNest demonstrate: 1) previously-undiscovered accelerator design points that can outperform state-of-the-art implementation by 39.3% in energy efficiency; 2) Pareto frontier curves that comprehensively and quantitatively reveal the multi-objective tradeoffs in custom DNN accelerators; 3) holistic design exploration of different level of quantization techniques including recently-proposed binary neural network (BNN).

DATE:	Friday, April 27, 2018
TIME:	9:00 am
PLACE:	Green Hall. Room 0120

Research Advisor:
Dr. Xuan 'Silvia' Zhang

This seminar is in partial fulfillment
of the Masters Degree

Washington University in St. Louis
SCHOOL OF ENGINEERING & APPLIED SCIENCE