From Model to FPGA: Software-Hardware Co-Design for Efficient Neural Network Acceleration

**ABSTRACT**

Artificial neural networks, which dominate artificial intelligence applications such as object recognition and speech recognition, are in evolution. To apply neural networks to wider applications, customized hardware are necessary since CPU and GPU are not efficient enough. FPGA can be an ideal platform for neural network acceleration since it is programmable and can achieve much higher energy efficiency compared with general-purpose processors. However, the long development period and insufficient performance of traditional FPGA acceleration solutions prevent it from wide utilization. In this work, we propose a complete design flow to achieve both fast deployment and high energy efficiency for accelerating neural networks on FPGA. Deep compression and data quantization are employed to exploit the redundancy in algorithm and reduce both computational and memory complexity. Two architecture designs for CNN and DNN/RNN will be introduced together with the compilation environment. Evaluated on Xilinx Zynq 7000 and Kintex Ultrascale series FPGA with real-world neural networks, up to 10 times higher energy efficiency can be achieved compared with mobile GPU and desktop GPU.

**BIOGRAPHY**

Yu Wang received his B.S. degree in 2002 and Ph.D. degree (with honor) in 2007 from Tsinghua University, Beijing. He is currently a Tenured Associate Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include brain inspired computing, application specific hardware computing, parallel circuit analysis, and power/reliability aware system design methodology. Dr. Wang has authored and coauthored over 130 papers in refereed journals and conferences. He has received Best Paper Award in ISVLSI 2012 and Best Poster Award in HEART 2012 with 6 Best Paper Nominations. He is a recipient of IBM X10 Faculty Award in 2010. He served as TPC chair for ICFPT 2011 and Finance Chair of ISLPED 2012-2016, and served as program committee member for leading conferences in these areas, including top EDA conferences such as DAC, DATE, ICCAD, ASP-DAC, and top FPGA conferences such as FPGA and FPT. Currently he serves as Associate Editor for IEEE Transactions on CAD and Journal of Circuits, Systems, and Computers. He also serves as guest editor for Integration, the VLSI Journal and IEEE Transactions on Multi-Scale Computing Systems.

All are welcome!