Datacenter-Scale Customized Computing

Jason Cong
Chancellor's Professor, UCLA Computer Science Department
Director, Center for Domain-Specific Computing
http://vast.cs.ucla.edu/people/faculty/jason-cong

Customized computing has been of interest to the research community for over three decades. The interest has intensified in the recent years as the power and energy become a significant limiting factor to the computing industry. For example, the energy consumed by the datacenters of some large internet service providers is well over 10^9 Kilowatt-hours. FPGA-based acceleration has shown 10-100X performance/energy efficiency over the general-purpose processors in many applications. However, programming FPGAs as a computing device is still a significant challenge. Most of accelerators are designed using manual RTL coding. The recent progress in high-level synthesis (HLS) has improved the FPGA programming productivity considerably. But extensive source code rewriting is still required to achieve high-performance acceleration.

In this talk, I shall present our ongoing work to enable further automation for customized computing. One effort is on automated compilation to combining source-code level transformation for HLS with efficient parameterized architecture template generations. I shall highlight our progress on loop restructuring and code generation, memory partitioning, data prefetching and reuse, combined module selection, duplication, and scheduling with communication optimization. These techniques allow the programmer to easily compile computation kernels to FPGAs for acceleration. Another direction is to develop efficient runtime support for scheduling and transparent resource management for integration of FPGAs for cloud-scale acceleration. Our runtime system provides scheduling and resource management support at multiple levels, including server node-level, job-level, and datacenter-level so that programmer can make use the existing programming interfaces, such as MapReduce, Hadoop, and Spark, for large-scale distributed computation. Finally, I shall discuss some acceleration results we have in multiple application domains, such as machine learning, medical imaging, and bioinformatics.

Speaker Bio: Jason Cong received his B.S. degree in computer science from Peking University in 1985, his M.S. and Ph. D. degrees in computer science from the University of Illinois at Urbana-Champaign in 1987 and 1990, respectively. Currently, he is a Chancellor's Professor at the UCLA Computer Science Department, the director of Center for Domain-Specific Computing (CDSC). He served as the department chair from 2005 to 2008. Dr. Cong's research interests include electronic design automation, energy-efficient computing, customized computing for big-data applications, and highly scalable algorithms. He has over 400 publications in these areas, including 10 best paper awards, and the 2011 ACM/IEEE A. Richard Newton Technical Impact Award in Electric Design Automation. He was elected to an IEEE Fellow in 2000 and ACM Fellow in 2008. He is the recipient of the 2010 IEEE Circuits and System Society Technical Achievement Award "For seminal contributions to electronic design automation, especially in FPGA synthesis, VLSI interconnect optimization, and physical design automation."

Dr. Cong has graduated 33 PhD students. Nine of them are now faculty members in major research universities, including Cornell, Fudan Univ., Georgia Tech., Peking Univ., Purdue, SUNY Binghamton, UCLA, UIUC, and UT Austin. One of them is now an IEEE Fellow, six of them got the highly competitive NSF Career Award, and one of them received the ACM SIGDA Outstanding Dissertation Award. Dr. Cong has successfully co-founded three companies with his students, including Aplus Design Technologies for FPGA physical synthesis and architecture evaluation (acquired by Magma in 2003, now part of Synopsys), AutoESL Design Technologies for high-level synthesis (acquired by Xilinx in 2011), and Neptune Design Automation for ultra-fast FPGA physical design (acquired by Xilinx in 2013). Currently, he is a co-founder and the chief scientific advisor of Falcon Computing Solutions, a startup dedicated to enabling FPGA-based customized computing in data centers.