INITIAL EXPERIENCES WITH DEPLOYING FPGA ACCELERATORS IN DATACENTERS

Dr. Zhenman Fang

Department of Computer Science, UCLA

2016年10月25日 星期二 2:00pm

理科五号楼410会议室

ABSTRACT: With the end of CPU core scaling due to dark silicon limitations, customized accelerators on FPGAs have gained increased attention in modern datacenters due to their lower power, high performance and energy efficiency. Evidenced by Microsoft's FPGA deployment in its Bing search engine and Intel's 16.7 billion acquisition of Altera, integrating FPGAs into datacenters is considered one of the most promising approaches to sustain future datacenter growth. However, it is quite challenging for existing big data computing systems—like Apache Spark and Hadoop—to access the performance and energy benefits of FPGA accelerators.

In this talk, I will present those challenges and share our initial experiences at UCLA about efficient FPGA accelerator deployment in datacenters. First, I will discuss how to choose the right CPU-FPGA server platform based on the quantitative microarchitecture study [DAC 16]. Then I will demonstrate how to efficiently integrate FPGA accelerators into Apache Spark using next-generation DNA sequencing acceleration as a case study [HotCloud 16]. Finally, I will present our generic system Blaze, which provides programming and runtime support for enabling easy and efficient deployment of FPGA accelerators in datacenters [ACM SoCC 16]. Blaze abstracts FPGA accelerators as a service (FaaS) and provides a set of clean programming APIs for multiple big data applications (e.g., Spark applications) to easily utilize and share those accelerators. By integrating a PCIe-based FPGA board into each commodity server, we have improved the system throughput by 1.7x to 3x (and energy efficiency by 1.5x to 2.7x) for a range of big data applications, compared to a conventional CPU-only cluster.

BIOGRAPHY: Dr. Zhenman Fang is a postdoc in Department of Computer Science, UCLA, under the supervision of Prof. Jason Cong and Prof. Glenn Reinman. He is also a member of the NSF/Intel funded multi-university Center for Domain-Specific Computing (CDSC) and the SRC/DARPA funded multi-university Center for Future Architectures Research (C-FAR). Zhenman earned his PhD degree in July 2014 from School of Computer Science, Fudan University, under the supervision of Prof. Binyu Zang. He also spent the last 15 months of his PhD life visiting Department of Computer Science and Engineering, University of Minnesota at Twin Cities, under the supervision of Prof. Pen-Chung Yew. Zhenman's research interests include big data and cloud computing, heterogeneous and energy-efficient accelerator-rich architectures and systems, near data computing, performance evaluation methodology, emerging workload characterization and optimization (especially for computational genomics and machine learning applications), and compiler optimizations. He has published 10+ papers in top venues such as DAC, ICCAD, ACM SoCC, ACM TACO, ICS, FCCM, and LCTES. More details can be found in Zhenman's personal website: https://sites.google.com/site/fangzhenman/.