Coarse-Grained FPGA Overlays for On-demand Acceleration of Data Center Workloads

[Extended Abstract]

Abhishek Kumar Jain, Douglas L. Maskell
School of Computer Science and Engineering
Nanyang Technological University, Singapore
{abhishek013, asdouglas}@ntu.edu.sg

Suhaib A. Fahmy
School of Engineering
University of Warwick, United Kingdom
s.fahmy@warwick.ac.uk

ABSTRACT

We present an approach for on-demand acceleration of data center workloads using high performance architecture-centric coarse-grained FPGA overlays. Proposed approach allows on-the-fly generation of accelerators on server node and dynamic reuse of FPGA resources for multiple workloads.

1. INTRODUCTION

Heterogeneous computing platforms coupling general purpose processors with high performance FPGA fabrics, represent a possible solution for accelerating data center workloads as server processor performance has begun to plateau. We can now consider applications with hardware accelerated portions that are reconfigured at runtime. FPGA partial reconfiguration offers one way of enabling the dynamic loading of accelerators, but reconfiguration latency and design constraints limit efforts to create a truly virtualised view of resources [1]. Due to both design productivity issues and a lack of suitable hardware virtualisation approach, as has been achieved with server and desktop virtualization, FPGAs still find use in only a limited subset of application domains. Cloud computing promises virtual resources available on demand with runtime mapping to physical systems abstracted from the user. However there is, as of yet, no widely accepted way of abstracting FPGA fabrics. For true on-demand acceleration of data center workloads on FPGAs, there is a need to support dynamic generation of accelerators at runtime within seconds, enabling dynamic reuse of FPGA resources for multiple workloads within microseconds.

2. COARSE-GRAINED FPGA OVERLAYS

Coarse-grained FPGA overlays have emerged as one possible solution to this challenge, offering a number of advantages for on-demand workload acceleration because of software-like programmability, fast compilation within seconds, fast reconfiguration within microseconds, device portability, and improved design productivity [2, 3].

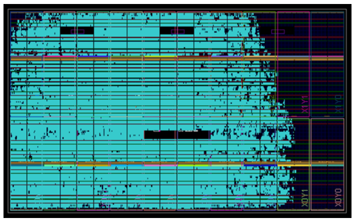

Figure 1: A coarse grained overlay consisting of 800 general purpose functional units on a Virtex 7.

The coarse-granularity of the high performance overlay architectures allows rapid hardware design and reconfiguration at a higher levels of abstraction, while also being built around the basic architecture of modern FPGA hard blocks. We propose on-demand acceleration of data center workloads by performing runtime compilation of OpenCL workloads onto a spatially configured island-style coarse-grained overlay (shown in Figure 1) implemented on a PCIe-attached Alpha Data Virtex-7 FPGA card (ADM-PCI-7V3). The overlay consists of 800 general purpose 16-bit functional units based on hard DSP primitives and can provide a peak throughput of 912 GOPS [3]. To allow on-demand acceleration of workloads, we have developed a just in time (JIT) compiler which can compile OpenCL kernels onto the overlay within seconds. To switch from one workload to another, only 8KB of configuration needs to be loaded over the PCIe interface, using OpenCL API, taking 50–100μs. The proposed approach can be effectively used for compiling and offloading deep learning kernels onto overlays at runtime and also for FPGA based database query processing where hardware needs to be changed rapidly for each new query.

3. REFERENCES

[1] S. A. Fahmy, K. Vipin, and S. Shreejith. Virtualized FPGA accelerators for efficient cloud computing. In *CloudCom*, 2015.

[2] A. K. Jain, S. A. Fahmy, and D. L. Maskell. Efficient Overlay architecture based on DSP blocks. In *FCCM*, 2015.

[3] A. K. Jain, D. L. Maskell, and S. A. Fahmy. Throughput oriented FPGA overlays using DSP blocks. In *DATE*, 2016.

ACM ISBN .
DOI: