A Case for FPGA Accelerators in the Cloud

Suhaib A. Fahmy, Kizheppatt Vipin

School of Computer Engineering
Nanyang Technological University, Singapore
sfahmy@ntu.edu.sg

Abstract

Hardware accelerators use custom architectures to significantly speed up computations in a wide range of domains. As performance scaling in server-class CPUs slows, we propose integration of hardware accelerators in the cloud as a way to maintain a positive performance trend. Field programmable gate arrays represent the ideal resource for such integration since they can be reprogrammed as needs change and allow dynamic sharing of communication infrastructure among multiple accelerators. We briefly discuss service levels for cloud accelerators and make a case for adoption.

1. Introduction

Virtualisation of resources is central to the efficient scaling and sharing provided by the cloud. However, it has a cost, and server CPUs are not scaling in performance at the rates previously observed. One strategy for overcoming this stalled scaling is to add heterogeneity to the cloud; providing access to resources better suited to complex computation. Some efforts have already emerged for integrating GPUs for computation [1], but while they can offer raw performance, GPUs are not power efficient and sharing between users/tasks is also complex.

We propose integration of FPGAs, hosting custom hardware accelerators that can be changed on demand. While offering significant speed-ups in execution of a wide variety of tasks over CPUs, FPGAs are also significantly more power-efficient, resulting in a computational efficiency improvement in the orders of magnitude over both CPUs and GPUs [2]. At the server hardware level, manufacturers have recently announced changes that will ease integration of FPGAs in servers, including the IBM POWER8 CAPI port and Intel XEON+FPGA solution. This integrates an FPGA with XEON processor in a single chip package, resulting in $10\times$ performance gains, while maintaining reprogrammability for a diverse client base. Microsoft recently showed such benefits integrating FPGAs in their Bing servers, resulting in a doubling in performance for the affected tasks, at a small increase in power consumption [3]. Evolving to a dynamically programmable, user-accessible paradigm is central to wider adoption of FPGAs; their dynamic programmability makes them the ideal resource for offering hardware performance with the generality demanded by the cloud.

Poster at ACM SoCC '14, November, 2014, Seattle, WA, USA.

2. Hardware in the Cloud Service Models

We envisage FPGAs being integrated in the cloud at different service levels depending on application. In the **Vendor Accelerator** model, a typical Software as a Service (SaaS) deployment can be accelerated by the cloud provider in a manner that is invisible to the user. This results in accelerated execution and increased (software) computational capacity in the CPU. Most work to date has focussed exclusively at this level, including that in [3].

In the **Accelerators as a Service** model, FPGAs are indirectly accessible to clients, through a library of accelerators that can be accessed via an API. An application can offload suitable computations to these accelerators through specific API calls. Providers can charge a premium for access to these accelerators. No FPGA experience whatsoever is required by the client, and the vendor can offer specialised libraries for different application domains.

In **Fabric as a Service**, FPGAs are made available as a distinct resource. Virtual FPGAs are made accessible for clients to load their own accelerator designs while the resource allocator manages allocation, reconfiguration, and data movement. This closely relates to IaaS, and the provider can charge for access to FPGAs as for other resources. Expert clients would use this to access otherwise expensive FPGA devices, and for large scale custom applications.

Facilitating these service models requires a virtualised framework for managing FPGAs in cloud servers. This involves sharing communication bandwidth between accelerators, dynamic loading into accelerator slots, and managing user requests. We have developed an initial demonstrator platform that allows dynamic loading of accelerators through API requests, based on work in [4], achieving near saturation of PCIe bandwidth shared among accelerators.

References

[1] W. Zhu *et al.*, "Multimedia cloud computing," *IEEE Sig. Proc. Mag.*, vol. 28, no. 3, pp. 59–69, 2011.

[2] S. Kestur *et al.*, "BLAS comparison on FPGA, CPU and GPU," in *Proc. Int. Symp. VLSI (ISVLSI)*, 2010, pp. 288–293.

[3] A. Putnam *et al.*, "A reconfigurable fabric for accelerating large-scale datacenter services," in *Proc. Int. Symp. Comp. Arch. (ISCA)*, 2014, pp. 13–24.

[4] K. Vipin and S. A. Fahmy, "DyRACT: A partial reconfiguration enabled accelerator and test platform," in *Proc. Int. Conf. Field Prog. Logic and Appl. (FPL)*, 2014.