What To Do With Datacenter FPGAs Besides Deep Learning

Moderator:
Andrew Putnam
Microsoft
Redmond, WA 98052
anputnam@microsoft.com

Panel Overview

FPGAs have been deployed in datacenters worldwide and are now available for use by in both public and private clouds. Enormous focus has been given to optimizing machine learning workloads for FPGAs, especially for deep neural networks (DNNs) in areas like web search, image classification, and translation.

However, major cloud applications encompasses a variety of areas that aren't primarily machine learning workloads, including databases, video encoding, text processing, gaming, bioinformatics, productivity and collaboration, file hosting and storage, e-mail, and many more. While machine learning can certainly play a role in each of these areas, is there more that can be done to accelerate these more traditional workloads?

Even more challenging than identifying promising workloads is figuring out how developers can practically create and deploy useful applications using FPGAs to the cloud. While FPGAs-as-a-Service allow access to FPGAs in the cloud, there is a huge gap between raw programmable hardware and a customer paying money to use an application powered by that hardware.

A wide variety of FPGA IP exists for developers to use, but individual IP blocks are a long way from being a fully functional cloud application. Building block IPs like Memcached, regex matching, protocol parsing, and linear algebra are only a subset of the necessary functionality for full cloud applications. Developing or acquiring IP and integrating it into a full application that customers will pay for is a significant task. And even when a customer pays, how should the money be distributed between IP vendors. Should it be a onetime fee? By usage? By number of FPGAs deployed? Who should have the burden for support if something goes wrong? In traditional cloud applications, FPGA IP block functions are implemented in software libraries. However, few examples of optimized software libraries are commercially successful, so is selling FPGA IP even a viable commercial model for cloud applications?

High-level synthesis (HLS) tools promise to provide one path to enable software developers to make effective use of FPGAs for computing tasks, but are any tools really capable of accelerating cloud-scale applications? Many HLS tools require substantial microarchitectural guidance in the form of pragmas or configuration files to come out with good results. Real cloud applications also rarely have a single dominant function and have significant data movement, so without proper partitioning and tuning, the acceleration gains from the FPGA are quickly wiped out by data movement and Amdahl's Law.

This panel will gather experts in using FPGAs for cloud application areas beyond machine learning, and how those applications can be built and successfully deployed. We will cover topics such as:

- What are the most important cloud workloads for FPGAs to target besides machine learning?
- Are there specific changes to the FPGA architecture that would benefit these cloud applications?
- What are the economic models that will work for IP developers, application developers, and cloud providers?
- How can we make development of FPGA applications easier for the Cloud?
- Will open source IP make it impossible for IP vendors to make commercially successful libraries?
- What advances are necessary for HLS tools to be practical in the Cloud?

The panel is comprised of experts in applications, IP development, and cloud deployment. Each will give a short presentation of what they find as the most important applications and how they see FPGA development for the cloud going forward, then we will open the floor to an interactive discussion with the audience.

CCS Concepts/ACM Classifiers

• Hardware → Reconfigurable logic and FPGAs;
• Computer systems organization → Reconfigurable computing;
• Computer systems organization → Cloud computing

Keywords: Reconfigurable Computing; FPGAs; Cloud Computing

FPGA '20, February 23–25, 2020, Seaside, CA, USA.

ACM ISBN 978-1-4503-7099-8/20/02.
DOI: https://doi.org/10.1145/3373087.3375885