FPGAs in Data Centers

BY GUSTAVO ALONSO

**EXPERT-CURATED
GUIDES TO
THE BEST OF
CS RESEARCH**

Research for Practice combines the resources of the ACM Digital Library, the largest collection of computer science research in the world, with the expertise of the ACM membership. In every RfP column experts share a short curated selection of papers on a concentrated, practically oriented topic.

This installment of Research for Practice features a curated selection from Gustavo Alonso, who provides an overview of recent developments utilizing FPGAs (field-programmable gate arrays) in datacenters. As Moore's Law has slowed and the computational overheads of datacenter workloads such as model serving and data processing have continued to rise, FPGAs offer an increasingly attractive point in the trade-off between power and performance. Gustavo's selections highlight early successes and practical deployment considerations that inform the ongoing, high-stakes debate about the future of datacenter- and cloud-based computation substrates. Please enjoy! —*Peter Bailis*

Most of today's IT is being driven by the convergence of three trends: the rise of big data, the prevalence of large clusters as the main computing platform (whether as the cloud, data centers, or data appliances), and the lack of a dominating processor architecture. The result is a fascinating cacophony of products and ideas around hardware acceleration and novel computer architectures, along with the systems and languages needed to cope with the ensuing complexity.

One key aspect of these developments is energy consumption, which is a crucial cost factor in IT and can no longer be ignored as a social issue. Power consumption in computing has many causes, but a well-known culprit is the movement required to bring data from storage to the processors along complex memory hierarchies. Such data transfers consume a proportionally enormous amount of energy without actually doing anything useful in terms of computation. Data movement also has a side effect, often overlooked in research: the performance penalty caused by moving the data to and from an accelerator; this movement often eats up most of the advantages provided by that accelerator.

It is in this context that FPGAs have attracted the attention of system architects and have started to appear in commercial cloud platforms. An FPGA allows the development of digital circuits customized to a given application. The customization makes them efficient in terms of both resource and energy consumption. Existing FPGAs typically consume one order of magnitude less power than CPUs or GPUs, even less in closely integrated systems that do not require a separate board. Unlike ASICs (application-specific integrated circuits), FPGAs are programmable in the sense that the circuit implemented can be swapped for a different one when the need arises (updates, upgrades, different uses, etc.).

The four papers presented here provide an overview of how FPGAs are being integrated into data centers and how they are being used to make data processing more efficient. They are presented in two groups, one showing

how designs in this area are quickly evolving and one detailing some of the ongoing debates around FPGAs.

A Reconfigurable Fabric for Accelerating Large-scale Datacenter Services

A. Putnam, A. M. Caulfield, E. S. Chung, et al.
41st ACM/IEEE International Symposium on Computer Architecture (ISCA), 2014
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Catapult_ISCA_2014.pdf

A Cloud-Scale Acceleration Architecture

A. M. Caulfield, E. S. Chung, A. Putnam, et al.
49th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2016
<https://www.microsoft.com/en-us/research/wp-content/uploads/2016/10/Cloud-Scale-Acceleration-Architecture.pdf>

These two papers are part of a series of publications by Microsoft describing Project Catapult (<https://www.microsoft.com/en-us/research/project/project-catapult/>). The first paper provides insights into the development process of FPGA base systems. The target application is accelerating the Bing web search engine. The configuration involves one FPGA per server, connected to the host through PCI (peripheral component interconnect). A separate network, independent of the conventional network, connects the FPGAs to each other using a six-by-eight, two-dimensional torus topology. The paper shows how such a system can improve the throughput of document ranking or reduce the tail latency for such operations by 29 percent.

The second paper builds on the lessons learned from the first. The web-search accelerator was based on a unit of 48 machines, a result of the decision to use a torus network to connect the FPGAs to each other. Not only is the cabling of such units cumbersome, but it also limits how many FPGAs can talk to each other and requires routing to be provided in each FPGA, complex procedures to achieve fault tolerance, etc.

In the cloud, scaling and efficiently using such a design is problematic. Hence, the second paper describes the solution being deployed in Azure: the FPGA is placed between the NIC (network interface controller) of the host and the actual network, as well as having a PCI connection to the host. All network traffic goes through the FPGA. The motivation for this is that the regular 40-Gbps network available in the cloud can also be used to connect the FPGAs to each other without a limitation on the number of FPGAs directly connected. With this design, the FPGA can be used as a coprocessor (linked to the CPU through PCI) or as a network accelerator (in front of the NIC), with the new resource being available through the regular network and without any of the limitations of the previous design. The design makes the FPGA available to applications, as well as to the cloud infrastructure, widening the range of potential uses.

Ibex—An Intelligent Storage Engine with Support for Advanced SQL Off-loading

Louis Woods, Zsolt István, Gustavo Alonso

Proceedings of the VLDB Endowment 7(11), 2014

<http://www.vldb.org/pvldb/vol7/p963-woods.pdf>

YourSQL: A High-Performance Database System Leveraging In-storage Computing

Insoon Jo, Duck-Ho Bae, Andre S. Yoon, Jeong-Uk Kang, Sangyeun Cho, Daniel DG Lee, Jaeheon Jeong
Proceedings of the VLDB Endowment 9(12), 2016
<http://www.vldb.org/pvldb/vol9/p924-jo.pdf>

These two papers illustrate an oft-heard debate around FPGAs. If the functionality provided in the FPGA is so important, can it not be embedded in an ASIC or a dedicated component for even higher performance and power efficiency? The first paper shows how to extend the database MySQL with an SSD+FPGA-based storage engine that can be used to offload queries or parts of queries near the storage. The result is much-reduced data movement from storage to the database engine, in addition to significant performance gains.

The second paper uses an identical database scenario and configuration but replaces the FPGA with the processor already available in the SSD (solid-state drive) device. Doing so avoids the data transfer from the SSD to the FPGA, which is now reduced to reading the data from storage into the processor used to manage the SSD.

As these two papers illustrate, the efficiency advantages derived from using a specialized processor must be balanced with the ability to repurpose the accelerator, a discussion that mirrors the steps taken by Microsoft designers toward refining the architecture of Catapult to increase the number of potential use cases. In a cloud setting, database applications would greatly benefit from an SSD capable of processing queries. All other applications, however, cannot

do much with it, a typical tradeoff between specialization (i.e., performance) and generality (i.e., flexibility of use) common in FPGA designs.

LOOKING AHEAD

FPGAs are slowly leaving the niche space they have occupied for decades (e.g., circuit design, customized acceleration, and network management) and are now becoming processing elements in their own right. This is a fascinating phase where different architectures and applications are being tested and deployed. As FPGAs are redesigned to use the latest technologies, it is reasonable to expect they will offer larger capacity, higher clock rates, higher memory bandwidth, and more functionality, and become available in off-the-shelf configurations suitable for data centers. How it all develops will be fascinating to watch in the coming years.

Gustavo Alonso is a professor of computer science at ETH Zürich, Switzerland, where he is a member of the Systems Group (www.systems.ethz.ch). His recent research includes multicore architectures, data appliances, cloud computing, and hardware acceleration, with the main goal of adapting system software (operating systems, databases, and middleware) to modern hardware platforms. He has M.S. and Ph.D. degrees from the University of California at Santa Barbara and was at IBM Almaden Research Center before joining ETH. He is a Fellow of the ACM and of the IEEE.

Copyright © 2018 held by owner/author. Publication rights licensed to ACM.