Guest Editors' Introduction: Hardware Accelerators for Data Centers

Mustafa Ozdal
Bilkent University

Debbie Marr
Intel Corp.

Gi-Joon Nam
IBM Research

It is our pleasure to introduce this special issue on hardware accelerators for data centers. Data centers around the world have been expanding and multiplying rapidly in the last decade with increased internet use, online services, compute consolidation, and data analytics. Hardware accelerators are increasingly important architectural components in the context of data center customization to achieve high performance and lower energy. Prominent companies have introduced FPGA/GPU-based platforms for data centers. For example, IBM's Coherent Accelerator Processor Interface and Intel's Quick-Assist Accelerator Abstraction Layer enable the integration of CPUs and FPGAs/GPUs through coherent shared memory. Microsoft built the configurable cloud platform for data centers and demonstrated significant performance improvements for different workloads. In addition to FPGAs and GPUs, application-specific hardware accelerators are being integrated into platforms for widely used workloads such as compression, cryptography, and pattern matching. Google's tensor processing unit is reported to be used to accelerate machine learning (ML) workloads at Google's data centers.

This special issue highlights transformative ideas related to the design and test of energy efficient, high performance, and secure computing technologies via accelerators, particularly tailored for data centers. Through a rigorous peer-review process, five papers out of ten submissions were finally selected for inclusion in this special issue.

In "A Memory Centric Architecture of the Link Assessment Algorithm in Large Graphs," Brugger et al. present a memory-centric optimized hardware architecture that achieves substantially better performance and energy efficiency for Link Assessment algorithm, which is a common big data graph application. With an innovative parallelization and customized DRAM subsystem architecture, the authors achieved an order of magnitude faster and more energy efficient system compared with the conventional existing system.

FPGA accelerators integrated with general-purpose CPUs have brought opportunities to improve the energy efficiency of data center workloads. In "CPU-FPGA Coscheduling for Big Data Applications," Cong et al. conducted a careful case study on one of the most important big data applications for personalized healthcare, *in-memory Samtools* sorting in genomic data processing. They propose a novel dataflow execution model to coordinate the computation between the multithreaded CPU and a high-performance FPGA.

The next article is "Designing for FPGAs in the Cloud" by Tarafdar et al. Due to the compute

Digital Object Identifier 10.1109/MDAT.2017.2779981
Date of current version: 2 February 2018.

capabilities and power efficiency, FPGAs have become a popular accelerator platform for big data applications. By capitalizing on the existing virtual machine model and OpenStack platform, Tarafdar et al. demonstrate that the software level design and test of an FPGA application is possible before committing to actual hardware implementation in data center environments. Thus, their techniques enable the inclusion and provisioning of FPGAs as accelerators in the cloud computing pool.

The security of data in the cloud is one of the major concerns that hold back cloud adoption for IT industries. The fourth article, "FASTEN: An FPGA-Based Secure System for Big Data Processing" by Hong et al., exactly addresses this issue by leveraging the security features in modern FPGAs such as crypto engines and physical unclonable functions. Their proposed system called FASTEN keeps security critical data stored in encrypted form in FPGA programmable logic so that they are not exposed to main memory and secondary storage. Through the performance evaluation of various applications using Hadoop MapReduce on Linux, the authors demonstrated both performance and security advantages over the conventional Hadoop environments.

Convolutional Neural Networks (CNNs) have become some of the most influential innovations in the field of ML, and accelerating CNNs has become a very important task for a variety of applications including computer vision. In the final article, "ZeNA: Zero-Aware Neural Network Accelerator" by Kim et al., the authors make a critical observation that a majority of the kernel weights and input activations in the state-of-the-art CNNs have zero values. They propose a CNN hardware accelerator that exploits this property to achieve significant performance and energy improvements. The need for such accelerator systems is apparent, as CNNs are such popular ML models that are being applied to a wide range of applications.

Furthermore, in this special issue, a survey paper titled "Emerging Accelerator Platforms for Data Centers," by Mustafa Ozdal, is provided to give the readers an overview of the important commercial and academic data center platforms with hardware accelerators.

THIS SPECIAL ISSUE would not have been possible without extensive help from the community. We are grateful to all the authors of the submitted papers for their important contributions to this exciting field, the reviewers for their comprehensive and rigorous reviews of multiple drafts, and the staff members of the *IEEE Design&Test*, including the editor-in-chiefs, administrative, and editorial staff members for helping us develop this special issue. We enjoyed working as the guest editors of this special issue and we are extremely glad that this special issue has finally come to fruition. ■

Mustafa Ozdal is an Assistant Professor with the Computer Engineering Department, Bilkent University, Ankara, Turkey. His research interests include high-performance computing, parallel and heterogeneous computing, computer-aided design algorithms, and hardware/FPGA accelerators for big data applications. He received a PhD in computer science from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2005.

Gi-Joon Nam is a Research Staff Member with the IBM's T.J. Watson Research Center, Yorktown Heights, NY, USA. His research interests include high-performance system architecture, VLSI designs and design methodologies, and hardware accelerator technologies particularly for big data applications. He has a PhD in computer science and engineering from the University of Michigan, Ann Arbor, MI, USA.

Debbie Marr is a Senior Principal Engineer and the Director of the Intel Labs' Accelerator Architecture Lab, Intel Corporation, Hillsboro, OR, USA. Her research team is focused on efficient hardware acceleration techniques to meet the computing needs of machine learning and artificial intelligence algorithm innovation. She has a PhD in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA.

■ Direct questions and comments about this article to Mustafa Ozdal, Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey; e-mail: mustafa.ozdal@cs.bilkent.edu.tr.