FPGA Computing

Maya Gokhale, *Lawrence Livermore National Laboratory, Livermore, CA, 94550, USA*
Lesley Shannon, *Simon Fraser University, Burnaby, BC, V5A 1S6, Canada*

Welcome to *IEEE Micro*'s Special Issue on FPGA Computing. FPGAs have become ubiquitous in designing, evaluating, and accelerating heterogeneous computing architectures. From early uses, such as accelerating network packet processing, signal processing, and logic emulation, FPGAs are now additionally used across the computing fabric in IoT devices, memory and storage systems, and as compute node accelerators. With the profusion of System-on-Chip (SoC) designs, FPGAs with embedded processors offer unique opportunities for microarchitectural innovation in closely integrating custom logic with the cache hierarchy and CPU cores.

This special issue of *IEEE Micro* explores cutting-edge research on topics that relate to FPGAs in computing. Our call for papers invited industry and academic researchers to consider these research challenges and resulted in a total of 21 submissions. We want to thank both the submitting authors and the numerous reviewers who participated in this process to create, identify, and improve the accepted submissions through our review process, which identified seven submissions for inclusion in this special issue. The selected articles cover a wide range of topics from heterogeneous computing system design to custom compute solutions for application acceleration to emulation.

Designing heterogeneous systems comprising one or more processors with one or more hardware accelerators is a very active research area. Key research questions include how to properly integrate the hardware accelerators into the compute platform's hierarchy. The article "Accelerator integration for open-source SoC design" describes a design flow that facilitates the integration of hardware accelerators and software in heterogeneous systems for SoC designs.

Additionally, researchers are interested in how to convert existing solutions purely described in software to a combination of hardware and software. As designers move computing loads to the edge to reduce data movement and energy consumption, there are questions as to which components of an application are the most appropriate to migrate from software to hardware for acceleration. As the number of software designers outnumbers the number of hardware designers by about an order of magnitude, the ability to automatically translate a software binary into hardware is extremely attractive as presented in the article "A binary translation framework for automated hardware generation."

THIS SPECIAL ISSUE OF IEEE MICRO EXPLORES CUTTING-EDGE RESEARCH ON TOPICS THAT RELATE TO FPGAs IN COMPUTING.

FPGAs have long been used to create custom compute solutions for high-performance applications that leverage their customization, opportunities for parallel processing, and deep pipelines. The article "Accelerating phylogenetics using FPGAs in the cloud" describes a proof-of-concept, custom, FPGA-based computing solution that is able to process phylogenetic analyses by nearly 10× faster than via a CPU using AVX2 extensions.

With the growth of AI and machine learning applications, FPGAs have been considered for CNN and DNN applications, which are often mapped to GPUs. Lower precision CNNs can reduce computational complexity and the bandwidth requirements on a platform, but they may also lead to unacceptable accuracy degradation. The article "High-performance mixed-low-precision CNN inference accelerator on FPGA" proposes mixed low-precision representations and computations to improve performance and throughput without unacceptably degrading accuracy.

The need for energy efficiency in data centers has led to the growing use of FPGAs for applications dominated by data movement. Combining FPGA computation with attached high-bandwidth memory as a near

0272-1732 © 2021 IEEE
Digital Object Identifier 10.1109/MM.2021.3088975
Date of current version 2 July 2021.

memory offload engine to traditional compute servers offers the opportunity to reduce data movement while simultaneously performing compute operations in power-efficient FPGAs. This approach is demonstrated in the article "FPGA-based near-memory acceleration of modern data-intensive applications" with two kernel applications for genomic analysis and weather prediction.

THE SELECTED ARTICLES COVER A WIDE RANGE OF TOPICS FROM HETEROGENEOUS COMPUTING SYSTEM DESIGN TO CUSTOM COMPUTE SOLUTIONS FOR APPLICATION ACCELERATION TO EMULATION.

Quantum computing represents another key research area as researchers consider alternate compute models for solving challenging search-space problems. As quantum computing becomes a more active area of research, it is important to be able to emulate the computations that they might perform. The article "FPGA-accelerated quantum computing emulation and quantum key distillation" provides a survey of current work in using FPGA to accelerate quantum computing emulation that leverages their ability to provide deep pipelined parallelism in conjunction with custom-precision operations.

Our final special issue article considers the challenging area of system emulation. Verifying increasingly complex computing platforms is a significant challenge in the chip design process. The authors of "Accessible, FPGA resource-optimized simulation of multiclock systems in FireSim" provide an open-source, FPGA-accelerated, cloud-hosted hardware emulation platform that tries to balance the need for accurate emulation against faster and cheaper prototyping solutions. Toward this goal, FireSim has been updated with two new capabilities in this article: multicycle resource optimizations and support for multiclock systems.

In closing, we would again like to thank all those who submitted manuscripts for consideration in this special issue and all of the reviewers who helped us select and guide authors in how to improve the articles you see here. We would also like to acknowledge the time and efforts of Lizy Kurian John to help us navigate this process and create this special issue. We hope you thoroughly enjoy reading the articles herein.

MAYA GOKHALE is currently a distinguished Member of Technical Staff at the Lawrence Livermore National Laboratory, Livermore, CA, USA. She contributes to open source projects on near memory computing emulation and SystemC synthesis tools based on LLVM. She is a coauthor of more than 130 technical publications. Her career spans research conducted in academia, industry, and national laboratories. Her current research interests include data-intensive architectures, near-memory computing, and reconfigurable computing. Gokhale received a Ph.D. degree in computer science from the University of Pennsylvania, Philadelphia, PA, USA. She was the co-recipient of an R&D 100 Award for a C-to-FPGA compiler, four patents related to memory architectures for embedded processors, reconfigurable computing architectures, and cybersecurity. She is a member of Phi Beta Kappa and a Fellow of IEEE. Contact her at maya@llnl.gov.

LESLEY SHANNON is currently with the Natural Sciences and Engineering Research Council (NSERC) Chair for Women in Science and Engineering for the BC/Yukon Region. She is a professional engineer. She is also a Professor and Chair for the Computer Engineering Option with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. Her current research interests include heterogeneous computing system design, reconfigurable computing, reconfigurable microfluidics, soft-processor design and verification, and application-specific architectures for a wide range of application areas including machine learning, aerospace, biomedical systems, and multimedia applications. Shannon received a Ph.D. degree in computer engineering from the University of Toronto, Toronto, ON, Canada. Her research team has created the open-source RISC-V Taiga processor architecture to support tightly integrated custom instructions and loosely integrated hardware accelerators while ensuring high performance for metrics including instruction per cycle and performance per resource usage. She is a member of IEEE. Contact her at lshannon@ensc.sfu.ca.