Hardware Acceleration

Martha Kim
Columbia University

Yakun Sophia Shao
Architecture Research Group, NVIDIA

John Hennessy and David Patterson welcomed the end of Moore's Law in their Turing Award lecture in ISCA 2018 and predicted that this end will bring about a new golden age of computer architecture. One of the key aspects of this golden age is the wide adoption of domain-specific hardware accelerators.

It is an exciting era for hardware acceleration. Lack of device scaling for general-purpose processors has motivated the need for hardware specialization in virtually every computing system, from mobile processors to desktops to data centers. As a result, almost all major semiconductor vendors and cloud service providers have chips that include accelerators, large or small, for a variety of applications. In addition, electronic-design-automation vendors are introducing new high-level synthesis tools to lower the design barriers to acceleration.

We are excited to share our passion for hardware acceleration with many authors who answered our call for papers. The review process selected five peer-reviewed articles based on reviewers' feedback. We also invited a paper from Wen-Mei Hwu and Sanjay Patel, guest editors of the 2008 IEEE Micro Special Issue on Accelerator Architecture, to offer a historical perspective on this topic.

In "Accelerator Architectures—A Ten Year Retrospective," Wen-Mei Hwu and Sanjay Patel reflect back to a time when it was unclear if and how GPUs and FPGAs would take hold in this domain. Ten years later, the questions are not about if, but how. Hwu and Patel reflect on the wide adoption of GPU and FPGA computing that happened in the last decade and discuss the emergence of application-specific accelerators, especially in the domain of machine learning. In addition, as educators, they also highlight the importance of education to accelerator adoption.

In "Optimized Solution to Accelerate in Hardware an Intra H.264/SVC Video Encoder," Ronaldo Husemann, Valter Roesler, and Altamiro Amadeu Susin present a video compression accelerator, which smooths the variability in encoding speed that typically results from varying information content in videos.

Three papers focus on system-level challenges associated with hardware and software ecosystems required by accelerators:

Naif Tarafdar and colleagues from the University of Toronto present "Galapagos: A Full Stack Approach to FPGA Integration in the Cloud." This paper describes abstractions to improve design and configuration productivity of clusters of FPGAs supporting cloud services.

Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna of Georgia Tech present a configurable interconnect with the bandwidth to support the many communication flows, regular and irregular, found in DNN accelerators in "A Communication-Centric Approach for Designing Flexible DNN Accelerators." The gains in the interconnect are realized in higher processing element usage and efficiency.

Published by the IEEE Computer Society
0272-1732/18/$33.00 ©2018 IEEE

Davide Giri, Paolo Mantovani, and Luca P. Carloni of Columbia University present "Accelerators & Coherence: An SoC Perspective." This analysis of the interaction between accelerators and memory demonstrates the necessity of full-system accelerator evaluation, in this case illuminating consequences of design time coherence choices.

The final paper, from the Barcelona Supercomputing Center, strikes a cautionary note. Sergi Alcaide, Leonidas Kosmidis, Hamid Tabani, Carles Hernandez, Jaume Abella, and Francisco J. Cazorla outline "Safety-Related Challenges and Opportunities for GPUs in the Automotive Domain." This paper highlights fundamental incompatibilities between GPUs and automotive safety requirements, two domains that, at their inception, were not expected to meet.

We would like to thank all those who submitted manuscripts for this special issue and all the reviewers who helped us select the articles. Many thanks also to Wen-Mei Hwu and Sanjay Patel for their unique perspectives on this topic. Finally, a special thanks to Lieven Eeckhout for his guidance throughout the process. We hope you enjoy reading this issue.

ABOUT THE AUTHORS

Martha Kim is an associate professor of computer science at Columbia University, where she leads the ARCADE Lab. Kim's research interests are in computer architecture, parallel programming, compilers, and low-power computing. She has a PhD in computer science and engineering from the University of Washington. Contact her at martha@cs.columbia.edu.

Yakun Sophia Shao is a senior research scientist in the Architecture Research Group, NVIDIA. Her research interests include specialized architecture, machine learning accelerators, and agile hardware design methodology. She has a PhD from Harvard University. Contact her at sshao@nvidia.com.