Hot Chips 29

Lieven Eeckhout
Ghent University

It is my great pleasure to welcome you to the 2018 Hot Chips special issue. Hot Chips is the annual conference typically held in August in Cupertino, California where new high-performance microprocessor designs are presented from industry, research labs, and academia. It is always exciting to see the latest blast of cutting-edge designs presented at the conference. *IEEE Micro* traditionally devotes its March/April issue to Hot Chips by featuring half a dozen or so articles on state-of-the-art microprocessors. I wholeheartedly thank all the authors for having taken the time in spite of their busy schedules to contribute these articles from the Hot Chips 29 conference to *IEEE Micro*.

Eric Chung and colleagues describe Project Brainwave, Microsoft's FPGA-empowered cloud infrastructure for accelerated deep neural network (DNN) models in real time. A fabric of high-performance FPGAs is directly attached to the datacenter network on which microservices can be called by CPU software in the datacenter. The Brainwave system pins pre-trained DNN models in high-bandwidth on-chip memories across multiple FPGAs, and it features a precision-adaptable soft processor to offer low-latency hardware microservices.

Jeff Dean, David Patterson, and Cliff Young—based on their past experience working on two generations of Google's Tensor Processing Unit (TPU)—present a roadmap for computer architects focused on machine-learning problems (deep learning, in particular). The authors underline six issues that will impact hardware design for machine learning, as well as pitfalls and fallacies for machine-learning hardware.

Scott Davidson and colleagues present the Celerity open-source SoC consisting of 500+ cores in 16-nm Fin field-effect transistor (FinFET) technology. Celerity consists of three tiers: a general-purpose tier comprised of RISC-V processors, a massively parallel tier consisting of a RISC-V tiled manycore array, and a specialization tier for neural-network acceleration. This massive chip was designed in an impressive 9 months by a small academic team.

Jack Choquette, Olivier Giroux, and Denis Foley describe Volta, Nvidia's latest flagship GPU. The article describes various performance and programmability features, including NVLink enhancements to improve CPU-to-GPU and GPU-to-GPU communication; a redesigned streaming multiprocessor (SM) for improved performance and efficiency; and Tensor cores attached to the SMs for enhanced performance for training deep-learning neural networks.

John Sell presents the gaming SoC that sits at the heart of the Microsoft Xbox One X entertainment console. It features 7 billion transistors and comprises a CPU and GPU connected to a high-bandwidth DDR5 main memory system, leading to a substantial increase in graphics performance over preceding Xbox One and Xbox One S systems.

Brendan Farley, John McGrath, and Christophe Erdmann from Xilinx describe the first product to integrate RF data converters into an FPGA SoC, called RFSoC. The chip is fabricated in 16-nm FinFET technology and supports radio frequencies of up to 4 GHz while including a state-of-the-art FPGA, allowing for an easy adaptation of the RFSoC system to various communication standards.

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

In addition to these six Hot Chips articles, this issue also features an article on automotive computing by Shanker Shreejith and Suhaib A. Fahmy. They describe a smart network interface that incorporates computation in the data path to enable more features in the network layer to offload tasks from the engine control unit (ECU) processor.

Finally, we also have a Micro Economics column by Shane Greenstein on how "technology outsiders" succeed when, and because, they started to look less like outsiders. The column goes through a number of examples in recent computer history to illustrate this point.

With that, I would like to conclude and wish you a happy reading, as always.

ABOUT THE AUTHOR

Lieven Eeckhout is a professor in the Department of Electronics and Information Systems at Ghent University. Contact him at lieven.eeckhout@ugent.be.

Hot Chips is the annual conference typically held in August in Cupertino, California where new high-performance microprocessor designs are presented from industry, research labs, and academia. It is always exciting to see the latest blast of cutting-edge designs presented at the conference.