Best Papers From Hot Chips 32

Priyanka Raina, *Stanford University, Stanford, CA, 94305, USA*
Cliff Young, *Google, Palo Alto, CA, 94306, USA*

Welcome to our special issue of *IEEE Micro*, which highlights the best presentations from Hot Chips 32, held virtually on August 16–18, 2020. Like many things in 2020, Hot Chips was unprecedented, going virtual for the first time in its history. Presentations were done by video, with a small production team working in a studio and a virtual conference supplemented by Zoom chatrooms. Despite the switch to the virtual format and the challenges to the global economy, attendance was the highest ever, and the technical program was robust, with strong representation from traditional CPU, GPU, and FPGA manufacturers, and strong offerings from startups in both communication networks and neural networks. This issue collects articles derived from the best talks, chosen after the conference by the Program Committee. It is a great time to work in computer architecture, where the combination of approaching limits to Moore's Law and new transformative application areas mean that both incumbent computer architectures and new startups have large contributions to make.

Two of the articles focus on server processors, one from a long-time mainframe manufacturer and one from a potential disruptor of the server business. "IBM's Power10 Processor" describes the newest instance of the POWER architecture from the seminal computing company. Power10 was designed for general-purpose enterprise computing with interconnection across 16 chips in a multiprocessor, with 1-TB/s memory bandwidth per CPU and high-bandwidth links to accelerators including GPUs. By contrast, "Marvell ThunderX3: Next-Generation Arm-Based Server Processor" represents the new wave of ARM-based server-class chips that aim to change the performance and price/performance of the datacenter server market.

Two articles describe chips with significant graphics capability. "The Xbox Series X System Architecture" details the system-on-a-chip that powers Microsoft's latest gaming console, dedicating over two-thirds of its die to the GPU that delivers 4K at 120 frames per second. While graphics remain central to the mission of the article titled "NVIDIA A100 Tensor Core GPU: Performance and Innovation," GPUs have become the default for high performance and programmability in one solution, supporting a huge variety of scientific computing workloads and powering both neural network training and inference.

Bridging between the general-purpose floating-point power of GPUs and the specialized application focus of neural network accelerators, "Manticore: A 4096-Core RISC-V Chiplet Architecture for Ultraefficient Floating-Point Computing" uses the extensibility of the RISC-V ISA to reduce control overheads and energy costs for neural-network workloads.

Our final two articles come from startups, both with networking in their roots. "Pensando Distributed Services Architecture" describes their domain-specific architecture (including chips and programmable software) for building new applications and services within a datacenter network. "Compute Substrate for Software 2.0" explains startup TensTorrent's unique architecture for neural network acceleration, which takes a packet-network-inspired approach to control and flexibility, allowing better support for sparsity, varying numerical precision, and compression than older, more monolithic neural network accelerators.

All of the talks from Hot Chips 32 (https://www.hotchips.org/archives/hc32/) are available at the Hot Chips website. Hot Chips are run by a great set of volunteers, including a sophisticated logistical and marketing team in the Organizing Committee and the wonderful set of academic and professional computer architects on our Program Committee. No other conference has the same focus on production computer systems, presented by their designers, sharing how and why they built their chips. We hope you find this issue, and future Hot Chips, as informative and fun as we have.

0272-1732 © 2021 IEEE
Digital Object Identifier 10.1109/MM.2021.3060294
Date of current version 26 March 2021.