

CPUs, GPUs, and More From Hot Chips'32

Lizy Kurian John , The University of Texas, Austin, TX, 78712, USA



Welcome to the March/April 2021 issue of *IEEE Micro*. This Issue features selected articles from the Hot Chips'32 Symposium, held virtually in August 2020. COVID-19 forced Hot Chips'32 to be a virtual event, however the chips are more interesting and powerful than ever! Whether it is graphics acceleration or sheer increase in traditional compute capability, the chip design arena has become hotter than ever. A lot of money is pouring into designing both special purpose and general purpose chips. *IEEE Micro* is pleased to present seven selected articles based on the presentations at the Hot Chips Symposium for our readers. Priyanka Raina of Stanford University and Cliff Young of Google served as guest editors for this special issue. They have compiled an excellent selection of articles on emerging chips and systems from the Symposium, including articles on IBM Power10, Marvell Thunder X3, Xbox Series X, NVIDIA A100, Manticore 4096 core chiplet, Pensando Distributed Service Architecture, and TensTorrent Compute substrate for Software 2.0. Please read the Guest Editors' Introduction to get a preview of the seven articles. Thanks to the editors, authors, and reviewers who worked hard to put this issue together.

In addition to the aforementioned seven Hot Chips articles, there are two General Interest articles. The first one is on the design process for Google's TPUv2 and TPUv3 training chips, written by Norrie *et al.* from Google. This was also originally presented at Hot Chips, however because the authors included one of the guest editors, this was considered as a General Interest article, and the review process was separately coordinated. The article describes Google's approach to machine learning (ML) hardware, and provides details on the scalar computation unit, the vector computation unit, the matrix computation units, the memory system, the interconnect, and the floor plan of TPU v2. The enhancements in TPU v3 compared to

TPU v2 are discussed next and the performance is compared using roofline plots.

The second General Interest article is "Klessydra-T: Designing Vector Coprocessors for Multithreaded Edge-Computing Cores" by Cheikh *et al.* of Sapienza University of Rome. This work addresses the introduction of coprocessor acceleration in interleaved multi-threading (IMT) cores for extreme edge computing. Specifically, it explores possible alternatives to implement vector coprocessing units in RISC-V cores, showing the synergy between IMT and data-level parallelism in edge-computing applications.

*WHETHER IT IS GRAPHICS
ACCELERATION OR SHEER INCREASE
IN TRADITIONAL COMPUTE
CAPABILITY, THE CHIP DESIGN ARENA
HAS BECOME HOTTER THAN EVER.
A LOT OF MONEY IS POURING INTO
DESIGNING BOTH SPECIAL PURPOSE
AND GENERAL PURPOSE CHIPS.*

This issue also features three department articles. Michael Mattioli of Goldman Sachs who joined the *IEEE Micro* Editorial Board as a new Department Editor has coauthored an article with Atte Lahtiranta on the hidden capabilities of video game consoles. The authors describe nonvideo game capabilities of video game consoles such as web browsing, video conferencing, and audio/video/document content creation. They posit that video game consoles are ideal for enterprise deployment, with security features and defense against a wide range of threats. The Microsoft Xbox Series X architecture is described with emphasis on the security processor. The security processor is housed in the security complex with crypto engine, random number generator, secure RAM, secure ROM, security fuse bank, and side-channel monitors. Please read the article to appreciate the less-discussed features of video game consoles.

This issue additionally includes an Award article from Luiz Andre Barroso of Google who received the 2020 Eckert-Mauchly Award. In his article, the author provides a brief history of warehouse-scale computing. He describes the progression of datacenter computing as it evolved during the last two decades, and also describes his personal journey as a computer engineer. He concludes the article acknowledging how the pandemic has made many to realize the importance of computing technology and cloud-based services, and how these have allowed us to continue to work and live.

Another article presented in this issue is a Micro Economics column by Shane Greenstein, "The Economics of Confrontational Conversation," discussing how confrontational conversations are commonplace on the internet. Greenstein focuses on the economics relevant to such confrontational conversations. One economic factor is that it is inexpensive to host terabytes of data. Additionally simple focal platforms attract more users, which then attracts more apps and content. The third relevant economic fact is that the mechanisms to address confrontation, whether human moderation or algorithmic processes, will be expensive.

MANY THEMATIC SPECIAL ISSUES ARE PLANNED FOR THE REMAINDER OF 2021. THEMES INCLUDE QUANTUM COMPUTING, FPGA COMPUTING, IN-MEMORY COMPUTING, AND SMART AGRICULTURE.

There have been some additions/enhancements to the *IEEE Micro* Editorial Board this month. Dr. Vijaykrishnan Narayanan of Penn State has been promoted to the Associate Editor-in-Chief of *IEEE Micro*. Michael Mattioli of Goldman Sachs will serve as a Department Editor for Security and Product Reviews. Prof. Guido Araujo of the University of Campinas (Brazil) joins the Editorial Board this month as an Associate Editor. I look forward to working with all of them and bringing to you an even more interesting reading experience.

Let me also provide an overview of what to expect in upcoming issues. The May/June issue will be the popular "Top Picks" Special Issue which presents the best of the best from articles in computer architecture conferences in 2020. Prof. Daniel Jimenez of Texas A&M University and a selection committee from industry and academia have selected 12 papers from more than 100 articles that were submitted in response to the Top Picks call for papers. Readers can look forward to an amazing collection of excellent articles in May/June.

Many thematic special issues are planned for the remainder of 2021. Themes include quantum computing, FPGA computing, in-memory computing, and smart agriculture. The July/August issue will be a Special Issue on Quantum Computing, guest edited by Ulya Karpuzcu of the University of Minnesota. The FPGA Computing Special Issue will be guest edited by Maya Gokhale of Lawrence Livermore Laboratories and Lesley Shannon of Simon Fraser University, Canada. The In-Memory Computing Special Issue will be guest edited by Reetuparna Das of the University of Michigan. We also have an open call for smart agriculture, focusing on the use of artificial intelligence and IoTs in agriculture, guest edited by Neeraj Kumar of Thapar University and Sudip Misra of IIT Kharagpur.

We invite readers to submit to these Special Issues. Please find the open calls at:

- › <https://www.computer.org/digital-library/magazines/mi/call-for-papers-special-issue-on-processing-in-memory>.
- › <https://www.computer.org/digital-library/magazines/mi/call-for-papers-special-issue-on-ai-edge-and-iot-for-smart-agriculture>.

IEEE Micro is interested in submissions on any aspect of chip/system design or architecture.

Hope you enjoy the articles presented in this issue. Happy reading!

LIZY KURIAN JOHN is a Cullen Trust for Higher Education Endowed Professor with the Electrical and Computer Engineering Department, University of Texas at Austin, Austin, TX, USA. Contact her at ljohn@ece.utexas.edu.