# Special Issue on Top Picks From the 2021 Computer Architecture Conferences

Sudhanva Gurumurthi, *Advanced Micro Devices, Inc., Austin, TX, 78753, USA*
Radu Teodorescu, *The Ohio State University, Columbus, OH, 43210, USA*

It is our pleasure to introduce the *IEEE Micro* Special Issue on Top Picks from the 2021 Computer Architecture Conferences. This special issue includes 12 articles chosen by a Selection Committee as being the most significant research articles in computer architecture in 2021 in terms of their novelty and potential for long-term impact.

## SELECTION PROCESS

The selection process was carried out by a committee of experts that we assembled. We served as the Co-Chairs of this Selection Committee. The committee consisted of 43 members, 23% of whom were women, 28% junior, 21% from industry, and 23% from outside the United States. The PC meeting was held virtually via Zoom in January 2022.

To embrace the interdisciplinary nature of computer architecture research today, we broadened the definition of a "computer architecture conference" to include both the traditional conferences in the field [International Symposium on Computer Architecture (ISCA)/International Symposium on Microarchitecture (MICRO)/International Symposium on High Performance Computer Architecture (HPCA)/International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)] and other top conferences in related areas where a computer architecture relevant paper is published. We welcomed submissions from both the main and industry tracks of these conferences.

We received in total 109 submissions. We used the cloud-based HotCRP platform for managing the submissions, reviews, running the PC meeting, and for communication with the authors and the selection committee members. In keeping with Top Picks of prior years, we followed a two-phase reviewing process. In the first round, four reviews were requested for each submission. The review deadline was then followed by an online discussion period to select those submissions that advanced to the second reviewing round. A discussion lead was assigned for each submission to initiate and manage the discussion and bring the reviewers to a consensus on whether an article should advance to the second round or not. In total, 65 submissions advanced to the second round. Each of these submissions received four additional reviews, after which a second online discussion was held prior to the PC meeting and the set of articles that were candidates for discussion at the PC meeting were identified. Discussion candidates had to have at least one reviewer willing to champion the article.

The articles were clustered into three groups. The discussion order within each group was based on the aggregate merit scores normalized to factor in review bias. This normalization algorithm compares how each reviewer's score differs from the mean score of the papers they reviewed and applies an adjustment if the reviewer's scores are consistently different from other reviewers of the same papers. This normalization algorithm was used for Top Picks last year and at other major computer architecture conferences. We used two Zoom breakout rooms for handling conflicts of interest: 1) a Discussion Room for all committee members not conflicted with a specific paper; and 2) a Conflict Room for committee members who have a conflict. For each paper that was discussed, the discussion lead summarized the work and the champion made the case for why it should be a Top Pick. We then asked if any of the other reviewers had anything to add and allowed the committee to ask questions. After this, we had the reviewers of the paper vote whether it should be a Top Pick and used a simple majority to determine the outcome. If the vote was tied, we had a committee-wide vote for Top Pick and again used a simple majority or a tie to determine the winner. If the article was not selected as a Top Pick, we followed a similar workflow to determine whether the article should be chosen as

0272-1732 © 2022 IEEE
Digital Object Identifier 10.1109/MM.2022.3173823
Date of current version 30 June 2022.

an Honorable Mention. The PC meeting ended when the Selection Committee identified 12 articles as Top Picks (a hard limit imposed by *IEEE Micro*) and another 12 Honorable Mention papers.

We then shifted roles from being Selection Committee co-chairs to *IEEE Micro* guest editors. Each of the 12 Top Picks articles was invited to submit a 6,000-word manuscript. Each manuscript was required to have at least 30% new content compared to the conference version of the paper and was again peer reviewed. The guest editors chose the reviewer for each Top Pick from the Selection Committee. After the peer review, editor recommendation, and Editor-in-Chief decision steps, the manuscripts went through the IEEE production process to be published in this special issue.

We now briefly summarize the papers selected as Top Picks and Honorable Mentions.

## SELECTED PAPERS

### Data Center and Cloud

As more computing moves to the cloud and the scale of the data center infrastructures to support user and workload demands grow, research and innovation are required to meet these needs efficiently and sustainably. There are four Top Picks articles devoted to this topic. "Overclocking in Immersion-Cooled Datacenters" is an article from multiple groups in Microsoft and 3M that was originally published in ISCA. This article makes the case for two-phase immersion cooling and overclocking to meet cloud workload demands and reduce cost. "Warehouse-Scale Video Acceleration" is an article from Google that was originally published in ASPLOS. This article describes a video acceleration system, which includes a new video coding unit accelerator, and insights from deploying the system at scale. "Practical and Scalable ML-Driven Cloud Performance Debugging With Sage" is an article from Cornell University and Google that was originally published in ASPLOS. This article addresses the problem of performance debugging of cloud microservices and presents a machine learning based system to identify the root causes of QoS violations that has been tested at scale. "Chasing Carbon: The Elusive Environmental Footprint of Computing" is an article from co-authors at Meta, Samsung, Harvard University, Soongsil University, and Arizona State University that was originally published in HPCA. This article examines the carbon emissions from computing end-to-end, points out hardware manufacturing to be an area where sustainability improvements are required, and discusses ways to achieve this goal.

### Security and Reliability

Hardware security has been a key concern in recent years, with efforts to understand and mitigate vulnerabilities and implement security mechanisms in an efficient manner. The reliability challenges from technology scaling, the scale of modern compute infrastructures, and the push toward functional safety (e.g., for automotive) have made resiliency a key research need. There are three Top Picks articles that address these issues. "Maya: Using Formal Control to Obfuscate Power Side Channels" is an article from co-authors at Yale University, the University of Illinois at Urbana-Champaign, and the University of Nevada, Reno, that was published in ISCA. This article uses formal control with a combination of DVFS, idle cycles, and a balloon task to distort power consumption and mislead attackers. "An Architecture to Accelerate Computation on Encrypted Data" is an article from MIT and the University of Michigan, Ann Arbor, that was published in MICRO. This article presents the design and evaluation of a programmable fully homomorphic encryption accelerator and shows that it can provide large speedups over software implementations. "Characterizing and Mitigating Soft Errors in GPU DRAM" is an article from NVIDIA that was published at MICRO. This article presents DRAM soft error insights and best practices for accelerated testing from a neutron beam-testing campaign and proposes several ECC options that provide byte and pin error correction while attaining high detection coverage for HBM2.

### Emerging Applications and Architectures

Our community is also exploring new architectures optimized for emerging applications, as well as platforms for new application domains. In the "The Laplace Microarchitecture for Tracking Data Uncertainty," originally published in MICRO by authors from Signaloid and the University of Cambridge, a new architecture is proposed that is able to track and compute on probability distributions associated with architectural state. The Laplace architecture speeds up applications with probabilistic data by multiple orders of magnitude. "Systematically Understanding Graph Accelerator Dimensions and the Value of Hardware Flexibility," published by researchers from the University of California, Los Angeles, in ISCA, addresses the highly fragmented space of graph accelerator architectures. It showcases unified taxonomy of graph algorithms and develops a template architecture that is flexible across multiple algorithm variants, while being able to integrate specialized features. "ILLIXR: An Open Testbed to Enable Extended Reality Systems

Research," developed at the University of Illinois at Urbana-Champaign and published in the IEEE International Symposium on Workload Characterization (IISWC), presents an open-source extended really test bed aimed at spurring innovation across the system stack for the emerging augmented reality/virtual reality applications.

## Memory Models and Optimizations

Emerging memory technologies are leading to a reconsideration of consistency models, while vector architectures present a new opportunity for runahead execution. "Distributed Data Persistency" presented by researchers from the University of Illinois at Urbana-Champaign in MICRO, considers the impact of nonvolatile memory on future distributed systems. It proposes the joint treatment of memory persistency and data consistency in such systems, and considers the tradeoffs between weak/strong persistency and consistency, and the implications on distributed applications. In "Vector Runahead for Indirect Memory Accesses," published in ISCA, authors from Ghent University, University of Edinburgh, and University of Cambridge present a fresh look at the

## HONORABLE MENTIONS

| Article Title | Conference | Summary |
|---|---|---|
| GPS: A Global Publish-Subscribe Model for Multi-GPU Memory Management | MICRO 2021 | This article presents a multi-GPU memory management technique that uses proactive data transfers. |
| Logical Abstractions for Noisy Variational Quantum Algorithm Simulation | ASPLOS 2021 | This article presents a quantum circuit simulation toolchain for variational algorithms. |
| A Hardware Accelerator for Protocol Buffers | MICRO 2021 | This article uses insights from scale to design a hardware accelerator for protocol buffers that is integrated with a RISC-V SoC. |
| SquiggleFilter: An Accelerator for Portable Virus Detection | MICRO 2021 | This article presents a hardware accelerator for the MinION handheld nanopore sequencer. |
| Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology | ISCA 2021 | This article presents the architecture and implementation of a DRAM-based PIM that uses an HBM-compatible interface. |
| Dagger: Efficient and Fast RPCs in Cloud Microservices With Near-Memory Reconfigurable NICs | ASPLOS 2021 | This article quantifies the networking needs of cloud microservices, the overheads current network stacks add to them, and demonstrates how programmable hardware acceleration can improve their performance. |
| A Hierarchical Neural Model of Data Prefetching | ASPLOS 2021 | This article presents a neural network-based prefetcher using a hierarchical model that can learn address correlations, increasing prefetching accuracy. |
| Cheetah: Optimizing and Accelerating Homomorphic Encryption for Private Inference | HPCA 2021 | This article combines algorithmic optimizations with hardware acceleration to improve the performance of homomorphic encryption for private inference. |
| Exploiting Long-Distance Interactions and Tolerating Atom Loss in Neutral Atom Quantum Architectures | ISCA 2021 | This article presents hardware and compiler methods to increase quantum system resilience to atom loss, reducing total computation time. |
| Superconducting Computing With Alternating Logic Elements | ISCA 2021 | This article demonstrated how general-purpose computing can be implemented efficiently in superconducting logic. |
| Robomorphic Computing: A Design Methodology for Domain-Specific Accelerators Parameterized by Robot Morphology | ASPLOS 2021 | This article introduces robomorphic computing, a general methodology for the codesign of domain-specific hardware accelerator architectures for robotic applications, based on high-level physical robot structure. |
| ITSLF: Inter-Thread Store-to-Load Forwarding in Simultaneous Multithreading | MICRO 2021 | This article presents a solution for multithreaded store-to-load forwarding with the goal of improving the performance of synchronization-intensive applications. |

## SELECTION COMMITTEE

| | | |
|---|---|---|
| Akanksha | Jain | Google |
| Akshitha | Sriraman | Carnegie Mellon University / Google |
| Alaa | Alameldeen | Simon Fraser University |
| Alexandra | Jimborean | Universidad de Murcia |
| Amro | Awad | North Carolina State University |
| Anand | Sivasubramaniam | Penn State University |
| Antonia | Zhai | University of Minnesota |
| Antonio | Gonzalez | Polytechnic University of Catalonia |
| Arkaprava | Basu | Indian Institute of Science |
| Ashish | Venkat | University of Virginia |
| Christopher | Batten | Cornell University |
| Daniel A. | Jiménez | Texas A&M University |
| Devesh | Tiwari | Northeastern University |
| Dimitrios | Skarlatos | Carnegie Mellon University |
| Gabriel | Loh | AMD Research |
| Hubertus | Franke | IBM Research |
| James C. | Hoe | Carnegie Mellon University |
| Jishen | Zhao | UCSD |
| John | Carter | IBM |
| Jun | Yang | University of Pittsburgh |
| Kaitlin | Smith | University of Chicago |
| Karu | Sankaralingam | University of Wisconsin |
| Koji | Inoue | Kyushu University |
| Lieven | Eeckhout | Ghent University |
| Nam Sung | Kim | UIUC |
| Natalie | Enright Jerger | University of Toronto |
| Nikos | Hardavellas | Northwestern University |
| Nuwan | Jayasena | AMD Research |
| Onur | Mutlu | ETH Zurich |
| Parthasarathy | Ranganathan | Google |
| Prashant J. | Nair | University of British Columbia |
| Rajeev | Balasubramonian | University of Utah |
| Ravishankar | Iyer | Intel |
| Samira | Khan | University of Virginia |
| Shubu | Mukherjee | SiFive |
| Simha | Sethumadhavan | Columbia University |
| Swamit | Tannu | University of Wisconsin-Madison |
| Tor | Aamodt | University of British Columbia |
| Tushar | Krishna | Georgia Tech |
| Ulya | Karpuzcu | University Minnesota |
| Vijay | Janapa Reddi | Harvard/UT Austin/ Google |
| Viji | Srinivasan | IBM |
| Yuan | Xie | Univ. of California Santa Barbara |

idea of runahead execution. This article highlights that memory accesses within loops repeat on a similar control flow path. This can be exploited to execute multiple iterations simultaneously using vector units for the purpose of runahead prefetching.

## ACKNOWLEDGMENTS

The authors would like to thank Lizy Kurian John (*IEEE Micro* Editor-In-Chief) for inviting them to serve as Selection Committee co-chairs and guest editors for the Special Issue on Top Picks from the 2021 Computer Architecture Conferences and for her guidance and support throughout the process. The authors also thank the Selection Committee members for all their hard work, especially through several key reviewing milestones that overlapped with the winter holidays. Kristin Barber and Saikat Majumdar helped gear up for the PC meeting and ensured that the virtual meeting ran smoothly. Jishen Zhao and Aamer Jaleel provided the Zoom scripts for running the virtual PC meeting. Daniel Jimenez provided valuable information and guidance from his experience serving as the Selection Committee Chair and Guest Editor for the Special Issue on Top Picks last year. Simha Sethumadhavan kindly agreed to serve as a backup PC Chair for the virtual PC meeting, if it ever became necessary. The authors also thank IEEE for providing funding to use the cloud-based HotCRP platform. Finally, the authors also thank all the authors who submitted to Top Picks this year and for their excellent contributions to the field of computer architecture.

**SUDHANVA GURUMURTHI** is a principal member of the technical staff at Advanced Micro Devices, Inc., Austin, TX, 78753, USA, where he leads advanced development in reliability, availability, and serviceability (RAS). Additionally, he serves on the Dean's Advisory Council of the College of Science and Engineering, Texas State University. Gurumurthi received a Ph.D. degree in computer science and engineering from Penn State. He is a senior member of IEEE. Contact him at Sudhanva.Gurumurthi@amd.com.

**RADU TEODORESCU** is a professor with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210, USA, where he leads the Computer Architecture Research Lab. His research interest is in computer architecture with a focus on energy-efficient computing, security, and reliability. Teodorescu received a Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign. He is a member of IEEE. Contact him at teodores@cse.ohio-state.edu.