Guest Editor's Introduction

The 2019 Top Picks in Computer Architecture

Hyesoon Kim
Georgia Institute of Technology

It is my pleasure to introduce the "2019 Top Picks in Computer Architecture." This annual publication presents 12 articles selected from major computer architecture conferences of the year. The 12 papers are recognized for their importance, mainly the long-term impact and influence on the industry and other researchers. The selection committee members put enormous effort into picking the papers. We asked what the criteria should be for the top picks, and then we tried to answer that question by looking for significant improvement over previous work, establishing a new area.

As in prior years, only 12 articles could be selected to appear in this special issue. The selection committee chose 14 additional high-quality articles to be recognized as honorable mentions. I strongly encourage you to read these articles (see the "Honorable Mentions" sidebar).

REVIEW PROCESS

This year's review process built on previous years' selection processes. Authors submitted a three-page document that contained a two-page summary of the article and one page of supporting arguments for long-term impact and influence on other researchers and industry. Twenty eight selection committee members (see the "Selection Committee" sidebar) read the three-page documentations along with the original conference papers (single-blind review process by nature). In keeping with the successful two-round ranking-based review process of the past several years, the PC members first categorized each article as either a top pick, an honorable mention, or not a top pick. They also ranked the articles. After the first round of reviews, all PC members participated in online discussions to decide which articles should move to the second round. In the first round, all the articles were assigned at least four reviewers, and in the second round, the articles had at least four additional reviewers.

This year, as we expanded our research areas into special accelerators that rely on emerging technologies, we found it particularly challenging to ensure all reviewers understood the underlying technologies. Because the selection process is concerned more with the impact of the work rather than evaluating its technical accuracy, technical expertise is less critical than for main conference reviews. Nonetheless, when several papers cover similar topics, it is also important to identify those worthy of nomination based on technical merits. To overcome the limitations on available expertise, we increased the number of reviewers for such emerging technology based papers.

Digital Object Identifier 10.1109/MM.2020.2992834
Date of current version 22 May 2020.

Published by the IEEE Computer Society

SELECTION COMMITTEE

- Arka Basu, Indian Institute of Science (IISc)
- Babak Falsafi, École Polytechnique Fédérale de Lausanne (EPFL)
- Boris Grot, University of Edinburgh
- Christopher Fletcher, University of Illinois Urbana-Champaign (UIUC)
- Daniel Jiménez, Texas A&M University
- Dmitry Ponomarev, SUNY Binghamton
- Edward Suh, Cornell University
- Gennady Pekhimenko, University of Toronto
- Jangwoo Kim, Seoul National University
- Jayasena Nuwan, AMD
- Jishen Zhao, University of California San Diego
- John Kim, KAIST
- Jose Joao, Arm Research
- Josep Torrellas, University of Illinois Urbana-Champaign (UIUC)
- Lisa Wu Wills, Duke University
- Mike O'Connor, NVIDIA/UT-Austin
- Mohit Tiwari, UT Austin
- Onur Mutlu, ETH Zurich/CMU
- Parthasarathy Ranganathan, Google
- Rajeev Balasubramonian, University of Utah
- Ravi Iyer, Intel
- Reetuparna Das, University of Michigan
- Thomas Wenisch, University of Michigan/Google
- Tor Aamodt, The University of British Columbia
- Tushar Krishna, Georgia Institute of Technology
- Ulya Karpuzcu, University of Minnesota
- Vijay Janapa Reddi, Harvard University
- Yunji Chen, Institute of Computing Technology Chinese Academy of Sciences (ICT–CAS)

PC MEETING

The in-person (now, we need to differentiate in-person versus virtual!) PC meeting occurred on January 10, 2020 on the campus of Georgia Institute of Technology, Atlanta, GA, USA. Of the 28 PC members from three continents, 24 attended the meeting and 4 could not attend due to last minute emergencies. The discussion order was loosely determined by the articles' overall score and rank. Articles with similar topics were grouped together so as to provide more consistent evaluations and effective discussions.

The PC meeting was conducted in two phases in order to minimize the influence of the discussion order. In the first phase, we made preliminary decisions about the articles' outcomes. In the second phase, we adjusted the results; e.g., by deselecting articles from the top pick pool or rescuing articles from the honorable mention pool.

Because voting was often the key instrument in deciding the outcome of the articles, we instituted a rigorous set of voting procedures. First, only reviewers of the article voted on whether the article was a top pick or an honorable mention. If the vote was above 60% among reviewers, the article became a candidate for top picks or honorable mentions (unless the vote was unanimous for top pick, in which case the article was finalized as a top pick in the first phase). If not, the vote went to all PC members excluding those with conflicts. If the vote was above 50%, then the article became a candidate for top picks or honorable mentions. Since some articles had many conflicts, for each article, we precounted the number of votes needed to top 50%. All vote results during the meeting were recorded and later used to decide the order of discussions in the second phase. This two-phase mechanism was critical because when PC members voted, they wanted to see all the articles before making final decisions.

SELECTED ARTICLES

Cloud and Accelerators

The importance of cloud computing and accelerators continues to grow in the architecture community. The challenges of evaluating cloud and edge systems are presented in "Unveiling the Hardware and Software Implications of Microservices in Cloud and Edge Systems" by Gan *et al*. The article presents a new open source benchmark suite for cloud microservices and evaluations on real systems to guide future architecture designs. Accelerating DNN is an ongoing topic in architecture

HONORABLE MENTIONS

- "The Accelerator Wall: Limits of Chip Specialization," *by Adi Fuchs and David Wentzlaff* (HPCA 2019).
- "TensorDIMM: A Practical Near-Memory Processing Architecture for Embeddings and Tensor Operations in Deep Learning," *by Youngeun Kwon, Yunjae Lee, and Minsoo Rhu* (MICRO 2019).
- "ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs," *by Fei Gao, Georgios Tziantzioulis, and David Wentzlaff* (MICRO 2019).
- "NDA: Preventing Speculative Execution Attacks at Their Source," *by Ofir Weisse, Ian Neal, Kevin Loughlin, Thomas F. Wenisch, and Baris Kasikci* (MICRO 2019).
- "Janus: Optimizing Memory and Storage Support for Non-Volatile Memory System," *by Sihang Liu, Korakit Seemakhupt, Gennady Pekhimenko, Aasheesh Kolli, and Samira Khan*, (ISCA 2019).
- "D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput," *by Jeremie S. Kim, Minesh Patel, Hasan Hassan, Lois Orosa, and Onur Mutlu* (HPCA 2019).
- "Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture," *by Yakun Sophia Shao, Jason Clemons, Rangharajan Venkatesan, Brian Zimmer, Matthew Fojtik, Nan Jiang, Ben Keller, Alicia Klinefelter, Nathaniel Pinckney, Priyanka Raina, Stephen G. Tell, Yanqing Zhang, William J. Dally, Joel Emer, C. Thomas Gray, Brucek Khailany, and Stephen W. Keckler* (MICRO 2019).
- "Buffets: An Efficient and Composable Storage Idiom for Explicit Decoupled Data Orchestration," *by Michael Pellauer, Yakun Sophia Shao, Jason Clemons, Neal Crago, Kartik Hegde, Rangharajan Venkatesan, Stephen W. Keckler, Christopher W. Fletcher, and Joel Emer* (ASPLOS 2019).
- "ExTensor: An Accelerator for Sparse Tensor Algebra," *by Kartik Hegde, Hadi Asghari-Moghaddam, Michael Pellauer, Neal Crago, Aamer Jaleel, Edgar Solomonik, Joel Emer, and Christopher Fletcher* (MICRO 2019).
- "A Formal Analysis of the NVIDIA PTX Memory Consistency Model," *by Daniel Lustig, Sameer Sahasrabuddhe, and Olivier Giroux* (ASPLOS 2019).
- "SpecShield: Shielding Speculative Data from Microarchitectural Covert Channels," *by Kristin Barber, Anys Bacha, Li Zhou, Yinqian Zhang, and Radu Teodorescu* (PACT 2019).
- "Designing Vertical Processors in Monolithic 3D," *by Bhargava Gopireddy and Josep Torrellas* (ISCA 2019).
- "CIDR: A Cost-Effective In-Line Data Reduction System for Terabit-per-Second Scale SSD Arrays," *by Mohammadamin Ajdari, Pyeongsu Park, Joonsung Kim, Dongup Kwon, Jangwoo Kim* (HPCA 2019).
- "Practical Byte-Granular Memory Blacklisting using Califorms," *by Hiroshi Sasaki, Miguel A. Arroyo, M. Tarek Ibn Ziad, Koustubha Bhat, Kanad Sinha, and Simha Sethumadhavan* (MICRO 2019).

conferences, as shown in "MAESTRO: A Data-Centric Approach to Understand Reuse, Performance, and Hardware Cost of DNN Mappings" by Kwon *et al*. This article presents an analytical cost-benefit analysis framework that considers the cost of DNN mapping tradeoffs in terms of data reuse. Accelerating virtual reality becomes more important, especially in the current environment. The article "Energy-Efficient Video Processing for Virtual Reality" by Leng *et al.* seeks to improve energy efficiency from the architecture support based on the characterizations and evaluations of VR prototypes.

Accelerations From Understanding Applications

Understanding target application characteristics generates synergetic architectural improvements. A generalized acceleration framework for irregular workloads is presented in "Towards General-Purpose Acceleration: Finding Structure in Irregularity" by Dadu *et al*. The article "Varifocal Storage: Dynamic Multiresolution Data Storage" by Hu *et al.* proposes a dynamic multiresolution storage system to help approximate computing. Understanding applications can also be extended to warehouse-scale computer (WSC) applications. A profiling and code

analysis tool to allow identification of critical code segments and then proposing solutions to improve the performance of WSC is presented in "AsmDB: Understanding and Mitigating Front-End Stalls in Warehouse-Scale Computers" by Nagendra *et al*.

Quantum Computing

2019 was the year that quantum computing architecture/compilers became one of the mainstream computer architecture research topics. Two articles were selected to represent the research challenges in quantum computing. The article "Extending the Frontier of Quantum Computers With Qutrits" by Gokhale *et al.* presents the use of three-level qutrits and also evaluates the system-level impact of qutrits-based quantum computing. The article also provides good background materials for quantum computing. A measurement-based full-stack characterization of basic quantum computing applications on real systems is presented in "Architecting Noisy Intermediate-Scale Quantum Computers: A Real-System Study" by Murali *et al*.

Security and Privacy

As a reflection of new computing design requirements from security and privacy, three articles on security and one article on privacy are presented in this issue. The article "Speculative Taint Tracking (STT): A Comprehensive Protection for Speculatively Accessed Data" by Yu *et al.* provides a framework that tracks flow of speculative instructions through covert channels. The article "MicroScope: Enabling Microarchitectural Replay Attacks" by Skarlatos *et al.* presents a means of replaying code by forcing microarchitectural reply based on address translations. ISA extensions design methodology to improve security while considering performance is presented in "Creating Foundations for Secure Microarchitectures With Data-Oblivious ISA Extensions" by Yu *et al*. Finally, the article "Trace Wringing for Program Trace Privacy" by Dangwal *et al.* presents a compression method to generate traces that can limit the information leak, and memory trace writing for cache simulation is shown as an example.

CONCLUSION

I hope you will enjoy reading these articles, and I encourage you to explore the full conference versions of both the Top Pick and Honorable Mention selections. The authors made significant efforts to write a version that can be read by a broad audience.

ACKNOWLEDGMENTS

I would like to thank Lizy Kurian John, the Editor-in-Chief of *IEEE Micro*, for providing support and guidance at every stage of issue preparation. I would also like to thank Vijay J. Redidi for handling papers that I had conflicts with and J. Zho for handling papers that both Vijay and I had conflicts with. I also thank the previous year's guest editor, S. Dwarkadas, for her input. Attending her Top Picks meeting last year was very helpful in preparing for this year's meeting. I thank M. Qureshi for providing valuable input during the discussions, and the submission chair Y. Kim, R. Hadidi and the volunteer students J. Lee, and B. Asgari who helped organize PC meetings. I thank all the PC members who have diligently read the articles and put enormous effort into selecting the finalists. Finally, I thank all the authors who have submitted their work.

Hyesoon Kim is an Associate Professor with the School of Computer Science, Georgia Institute of Technology. Her research areas include the intersection of computer architectures and compilers, with an emphasis on heterogeneous architectures such as GPUs and accelerators. She is a recipient of NSF Career Award and is a member of Micro Hall of Fame. She is a senior member of IEEE. Contact her at hyesoon@cc.gatech.edu.