Guest Editor's Introduction

Top Picks in Computer Architecture from Conferences in 2018

Sandhya Dwarkadas
University of Rochester

It is my pleasure to introduce the *top picks* selection of papers in computer architecture that were published in 2018. This is the 16th year of publication of what has now become a tradition in the computer architecture community. Each year, the MICRO Top Picks selection committee, which I chaired for this year, selects 12 papers for this special issue, to highlight for the *IEEE Micro* readership, a sample of the papers published in 2018 that have architectural contributions of potentially high impact and significance.

SELECTION COMMITTEE AND REVIEW PROCESS

This year's selection committee, listed below, had 32 members. Authors were asked to submit the original conference paper along with a two-page summary of its key contributions and a one-page statement outlining its potential for future impact and reason for selection as a Top Pick. Authors were allowed to submit papers they published in any 2018 conference as long as the architectural contribution was deemed to be significant and was clearly identified, with the exception of the prior year's selection committee chair, who is allowed to submit papers coauthored in 2017 due to their conflict-based ineligibility last year. The authors of 123 papers chose to submit their papers for consideration as a top pick.

As in the last two years, the committee followed a two-phase review process, with each paper receiving four reviews in the first phase and an additional five reviews if the paper was advanced to the second phase. In addition to categorizing each paper as a "Top Pick," "Honorable Mention," or "Not a Top Pick," reviewers were asked to order the papers in their individual review stack during each review phase. Papers advanced to the second phase based on (extensive in some cases) online discussion, requiring two advocates among the reviewers, but also considering factors that weighted whether or not reviewers considered the paper a Top Pick or Honorable Mention against individual reviewer scoring preferences reflected by the rank order of the paper within their pile. Committee members were encouraged to and exercised their right to reconsider decisions and advance

Digital Object Identifier 10.1109/MM.2019.2911751
Date of current version 8 May 2019.

Published by the IEEE Computer Society

additional papers during the second round of reviewing. Fifty-one papers advanced to the second phase.

The selection committee meeting was held in Rochester, NY, USA, on January 11, 2019, with all but three committee members physically present. As luck would have it, weather in Rochester was not a factor! All 51 papers were considered for discussion, with the discussion order using a combination of average rank order, top pick rating, and grouping by topic or request. The hours of online discussion followed by the day-long discussion and hard work by the committee culminated in the selection of the 12 papers in this special issue as Top Picks. An additional 11 papers were deemed worthy of Honorable Mention in providing notable contributions that *IEEE Micro* readers might consider seeking out. Ultimately, papers with at least a two-third majority recommending "Top Pick" or "Honorable Mention" received at least an honorable mention designation.

As in the last two years, the committee followed a two-phase review process, with each paper receiving four reviews in the first phase and an additional five reviews if the paper was advanced to the second phase.

TOPICS AMONG THE SELECTED ARTICLES

Not surprisingly, ***application-specific acceleration*** continues to be one of the dominant themes in the selected articles, with deep neural networks, genomics, and garbage collection being the focus of the top picks. "Neural Cache: Bit-Serial In-Cache Acceleration of Deep Neural Networks," presented at the ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018), shows how SRAM caches may be augmented with in-place support for arithmetic operations to accelerate DNNs. "A Configurable Cloud-Scale DNN Processor for Real-Time AI," presented at the ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018), describes Microsoft's Project Brainwave Neural Processing Unit, using FPGAs and a SIMD ISA to dispatch millions of operations from a single instruction. "Darwin: A Genomics Co-processor Provides up to 15,000 Acceleration on Long Read Assembly," presented at the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2018), teaches important lessons (with impressive results) in hardware-algorithm codesign in the context of the important and computationally intensive problem of sequence alignment in genomics. "A Hardware Accelerator for Tracing Garbage Collection," presented at the ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018), shows the importance of designing hardware support for garbage collection (in a manner similar to that for page table walks) that is located close to the memory controller.

Additionally, papers targeting continuous computer vision (highlighting the need to consider the entire application rather than constituent kernels in the design of application-specific accelerators); programmability via system call support on GPUs; the need to optimize for data flow via communication network configurability in designing accelerators for deep neural networks; the programmability of mixed-signal accelerators for machine learning; and error correction that avoids accumulating errors in analog memristor-based *in situ* neural network acceleration received honorable mention.

In the space of ***system and processor design***, "Composable Building Blocks to Open Up Processor Design," presented at the 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2018), describes a processor design methodology with potential to ease the design of complex hardware systems, particularly the design of modular systems in which hardware modules can be exchanged. "FireSim: FPGA-Accelerated Cycle-Exact Scale-Out System Simulation in the Public Cloud," presented at the ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018), shows how FPGA-accelerated simulation of silicon-proven RTL designs on a distributed network of publicly available FPGA resources can be used to enable cycle-exact microarchitectural simulation of scale-out clusters.

Additionally, papers that presented tools for the automatic generation of directory-based

Table 1. Selection committee.

Arrvindh Shriraman	Simon Fraser University
Babak Falsafi	EPFL
Benjamin Lee	Duke University
Boris Grot	University of Edinburgh
Brad Beckmann	AMD
Chris Wilkerson	Nvidia
Christina Delimitrou	Cornell University
Daniel Sanchez	MIT
David Albonesi	Cornell University
Doug Burger	Microsoft
Hsien-Hsin Lee	Facebook
Hyesoon Kim	Georgia Tech.
Jaejin Lee	Seoul National University
Lisa Hsu	Microsoft
Lixin Zhang	Institute of Computing Technology/ Chinese Academy of Sciences
Margaret Martonosi	Princeton University
Mohit Tiwari	UT Austin
Moin Qureshi	Georgia Tech.
Natalie Enright Jerger	University of Toronto
Parthasarathy Ranganathan	Google
Per Stenstrom	Chalmers University
R. Govindarajan	Indian Institute of Science
Rajeev Balasubramonian	University of Utah
Ravi Rajwar	Intel
Reetuparna Das	University of Michigan
Russ Joseph	Northwestern University
Sandhya Dwarkadas (chair)	University of Rochester
Shubu Mukherjee	Marvell
Stefanos Kaxiras	Uppsala University
Ulya Karpuzcu	University of Minnesota/ Brown University
Yuan Xie	UCSB
Yuhao Zhu	University of Rochester

coherence protocols that include transient states from the specification of its stable states and for the automated all-program verification of axiomatic microarchitectural memory ordering against ISA-level memory consistency specification, received honorable mention.

Another theme in the 2018 publications and selected "top picks" is ***security***, particularly related to ***side channel*** attacks and defenses: exposing new security vulnerabilities, proposing mitigations/defenses to protect against vulnerabilities, or developing new architectural designs to ensure, or tools to verify, that processor designs do not expose vulnerabilities. "Foreshadow: Extracting the Keys to the Intel SGX Kingdom With Transient Out-of-Order Execution," presented at the 27th USENIX Conference on Security (SEC 2018), shows how virtual memory support can be circumvented in Intel's SGX security extensions by allowing speculative access to data that has not been mapped. "Mobilizing the Micro-Ops: Exploiting Context Sensitive Decoding for Security and Energy Efficiency," presented at the ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018), uses dynamically modifiable microcode as a security defense mechanism by, for example, inserting decoy micro-ops or memory fences to thwart side channel attacks. "CheckMate: Automated Synthesis of Hardware Exploits and Security Litmus Tests," presented at the 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2018), shows how microarchitecturally happens-before graphs of microarchitecture operations used to uncover memory consistency violations can also be extended to uncover potential side channel information leakage via an automated tool for evaluating hardware susceptibility to formally specified classes of security exploits, enabling design-time hardening.

Additionally, solutions to improve the performance of verifying data integrity by dual-purposing message authentication codes as error-correcting codes, and defenses against hardware speculation attacks by making load speculation invisible in the cache hierarchy, received honorable mention.

Finally, 2018 publications also presented several innovations in the space of ***memory and network infrastructure***. As byte-addressable persistent memories become a reality, "Language-Level Persistency," presented at the ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA 2017), proposes

extending language-level memory models to include guarantees on the order of persistent writes and makes the case for the co-design of language-level and ISA-level persistency models. "Nonblocking Memory Refresh," presented at the ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018), offers a simple solution to DRAM refresh stalls (refreshing only a portion of a memory block at a time with sufficient redundancy to reconstruct the refreshing/unreadable data) that can be implemented in the memory controller and in the DDR interface and can truly make the refresh process invisible to the processor. "Synchronized Progress in Interconnection Networks (SPIN): A New Theory for Deadlock Freedom," presented at the ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018), turns solutions for deadlock in routing from one of overprovisioning or routing restriction to one of topologically independent coordinated orchestration.

Additionally, work on dynamic control of ***energy storage capacity*** in battery-less systems to simultaneously meet the needs of energy-capacity-constrained and temporally constrained tasks, along with techniques to manage ***latency and energy*** that combine the benefits of control theory and machine learning for dynamic configuration management, received honorable mention.

We hope the readers enjoy this compendium! The authors have spent considerable effort to clearly identify the key takeaways from their work for a broad audience. Further detailed information may be found in their original publications.

TOP PICKS

1. C. Eckert *et al.*, "Neural cache: Bit-serial in-cache acceleration of deep neural networks," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00040.
2. J. Fowers *et al.*, "A configurable cloud-scale DNN processor for real-time AI," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00012.
3. Y. Turakhia, G. Bejerano, and W. J. Dally, "Darwin: A genomics co-processor provides up to 15,000 acceleration on long read assembly," in *Proc. 23rd Int. Conf. Architect. Support Programm. Lang. Oper. Syst.*, 2018, DOI: 10.1145/3296957.3173193.
4. M. Maas, K. Asanovic, and J. Kubiatowicz, "A hardware accelerator for tracing garbage collection," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00022.
5. S. Zhang, A. Wright, T. Bourgeat, and Arvind, "Composable building blocks to open up processor design," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2018, DOI: 10.1109/MICRO.2018.00015.
6. S. Karandikar *et al.*, "FireSim: FPGA-accelerated cycle-exact scale-out system simulation in the public cloud," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00014.
7. J. Van Bulck *et al.*, "Foreshadow: Extracting the keys to the intel SGX kingdom with transient out-of-order execution," in *Proc. 27th USENIX Conf. Security Symp.*, 2018, pp. 991–1008.
8. M. Taram, A. Venkat, and D. M. Tullsen, "Mobilizing the Micro-Ops: Exploiting context sensitive decoding for security and energy efficiency," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00058.
9. C. Trippel, D. Lustig, and M. Martonosi, "CheckMate: Automated synthesis of hardware exploits and security litmus tests," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2018, DOI: 10.1109/MICRO.2018.00081.
10. A. Kolli *et al.*, "Language-level persistency," in *Proc. 44th Int. Symp. Comput. Architecture*, 2017, DOI: 10.1145/3079856.3080229.
11. K. Nguyen, K. Lyu, X. Meng, V. Sridharan, and X. Jian, "Nonblocking memory refresh," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00055.
12. A. Ramrakhyani, P. V. Gratz, and T. Krishna, "Synchronized progress in interconnection networks (SPIN): A new theory for deadlock freedom," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00064.

HONORABLE MENTIONS

1. Y. Zhu, A. Samajdar, M. Mattina, and P. Whatmough, "Euphrates: Algorithm-SoC Co-design for low-power mobile continuous vision," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00052.
2. J. Vesely, A. Basu, A. Bhattacharjee, G. H. Loh, M. Oskin, and S. K. Reinhardt, "Generic system calls for GPUs," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00075.
3. H. Kwon, A. Samajdar, T. Krishna, "MAERI: Enabling flexible dataflow mapping over DNN accelerators via reconfigurable interconnects," in *Proc. 23rd Int. Conf. Architect. Support Programm. Lang. Oper. Syst.*, 2018, DOI: 10.1145/3296957.3173176.
4. P. Srivastava *et al.*, "Promise: An end-to-end design of a programmable mixed-signal accelerator for machine-learning algorithms," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00015.
5. B. Feinberg, S. Wang, and E. Ipek, "Making memristive neural network accelerators reliable," in *Proc. IEEE Int. Symp. High Performance Comput. Architecture*, 2018, DOI: 10.1109/HPCA.2018.00015.
6. N. Oswald, V. Nagarajan, and D. J. Sorin, "ProtoGen: Automatically generating directory cache coherence protocols from atomic specifications," in *Proc. ACM/IEEE Annu. Int. Symp. Comput. Architecture*, 2018, DOI: 10.1109/ISCA.2018.00030.
7. Y. A. Manerkar, D. Lustig, M. Martonosi, and A. Gupta, "PipeProof: Automated memory consistency proofs for microarchitectural specifications," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2018, DOI: 10.1109/MICRO.2018.00069.
8. G. Saileshwar, P. J. Nair, P. Ramrakhyani, W. Elssaser, and M. K. Qureshi, "Synergy: Rethinking secure-memory design for error-correcting memories," in *Proc. IEEE Int. Symp. High Performance Comput. Architecture*, 2018, DOI: 10.1109/HPCA.2018.00046.
9. M. Yan, J. Choi, D. Skarlatos, A. Morrison, C. W. Fletcher, and J. Torrellas, "InvisiSpec: Making speculative execution invisible in the cache hierarchy," in *Proc. 51st Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2018, DOI: 10.1109/MICRO.2018.00042.
10. A. Colin, E. Ruppel, and B. Lucia, "A reconfigurable energy storage architecture for energy-harvesting devices," in *Proc. 23rd Int. Conf. Architect. Support Programm. Lang. Oper. Syst.*, 2018, DOI: 10.1145/3296957.3173210.
11. N. Mishra, C. Imes, J. D. Lafferty, and H. Hoffmann, "Caloree: Learning control for predictable latency and low energy," in *Proc. 23rd Int. Conf. Architect. Support Programm. Lang. Oper. Syst.*, 2018, DOI: 10.1145/3296957.3173184.

ACKNOWLEDGMENTS

In addition to the authors and the selection committee, I would like to thank David Costello, Kristi Kongmany, and James Roche at the University of Rochester for website support, administrative support, and technical support during the PC meeting, respectively; and Joanna Gojlik from IEEE for her help and patience in putting the special issue together. Margaret Martonosi handled submissions with which I had a conflict. Lieven Eeckhout as Editor-in-Chief during 2018, as well as past selection committee chairs Tom Wenisch and Moin Qureshi, provided invaluable advice on the review process. Lizy John, as the 2019 Editor-in-Chief, was an enormous help with the shepherding of the final versions of the selected Top Picks papers.

Sandhya Dwarkadas is the Albert Arendt Hopeman professor of engineering, and professor and chair of computer science with a secondary appointment in electrical and computer engineering at the University of Rochester. Her research interests include computer architecture and parallel and distributed systems. She has a BS from the Indian Institute of Technology, and an MS and a PhD from Rice University. She is a Fellow of the IEEE and ACM, and a member of the CRA-W Board and Steering Committee. Contact her at sandhya@cs.rochester.edu.