

Guest Editors' Introduction

TOP PICKS FROM THE 2016 COMPUTER ARCHITECTURE CONFERENCES



..... It is our pleasure to introduce this year's Top Picks in Computer Architecture. This issue is the culmination of the hard work of the selection committee, which chose from 113 submissions that were published in computer architecture conferences in 2016. We followed the precedent set by last year's co-chairs and encouraged the selection committee members to consider characteristics that make a paper worthy of being a "top pick." Specifically, we asked them to consider whether a paper challenges conventional wisdom, establishes a new area of research, is the definitive "last word" in an established research area, has a high potential for industry impact, and/or is one they would recommend to others to read.

Since the number of papers that could be selected for this Top Picks special issue was limited to 12, we continued the precedent set over the past two years of having the selection committee recognize 12 additional high-quality papers for Honorable Mention. We strongly encourage you to read these papers (see the "Honorable Mentions" sidebar). Before we present the list of articles appearing in this special issue, we will first describe the new review process that we implemented to improve the paper selection process.

Review Process

A selection committee comprising 31 members reviewed all the 113 papers (see the "Selection Committee" sidebar). This year, we tried a different selection process com-

pared to previous years' Top Picks, keeping in mind the constraints and objectives that are unique to Top Picks. The conventional approach to Top Picks selection has largely remained similar to that used in our conferences (for example, four to five reviews per paper and a four-to-six-point grading scale). For Top Picks, the number of papers that can be accepted is fixed (11 to 12), and the selection committee's primary job is to identify the top 12 papers out of all the submitted papers, instead of providing a detailed critique of the technical work and how the paper can be improved. The papers submitted to Top Picks tend to be of much higher (average) quality than the typical paper submitted at our conferences, and in many cases the reviewers are already aware of the work (through prior reviewing, reading the papers, or attending the presentations). Therefore, the time and effort spent reviewing Top Picks papers tends to be less than that spent reviewing the typical conference submissions.

We identified two key areas in which the Top Picks selection process could be improved. First, a small number of reviewers (approximately five) made the decisions for Top Picks. The confidence in selection could be improved significantly by having a larger number of reviews (approximately 10) per paper, especially for the papers that are likely to be discussed at the selection committee meeting. This also ensures that reviewers are more engaged at the meeting and make informed decisions. Second, the selection of Top Picks gets overly influenced by excessively

Aamer Jaleel
Nvidia

Moinuddin Qureshi
Georgia Tech

Honorable Mentions

Paper	Summary
“Exploiting Semantic Commutativity in Hardware Speculation” by Guowei Zhang, Virginia Chiu, and Daniel Sanchez (MICRO 2016)	This paper introduces architectural support to exploit a broad class of commutative updates enabling update-heavy applications to scale to thousands of cores.
“The Computational Sprinting Game” by Songchun Fan, Seyed Majid Zahedi, and Benjamin C. Lee (ASPLOS 2016)	Computational sprinting is a mechanism that supplies extra power for short durations to enhance performance. This paper introduces game theory for allocating shared power between multiple cores.
“PoisonIvy: Safe Speculation for Secure Memory” by Tamara Silbergleit Lehman, Andrew D. Hilton, and Benjamin C. Lee (MICRO 2016)	Integrity verification is a main cause of slowdown in secure memories. PoisonIvy provides a way to enable safe speculation on unverified data by tracking the instructions that consume the unverified data using poisoned bits.
“Data-Centric Execution of Speculative Parallel Programs” by Mark C. Jeffrey, Suvinay Subramanian, Maleen Abeydeera, Joel Emer, and Daniel Sanchez (MICRO 2016)	The authors’ technique enables speculative parallelization (such as thread-level speculation and transactional memory) to scale to thousands of cores. It also makes speculative parallelization as easy to program as sequential programming.
“Efficiently Scaling Out-of-Order Cores for Simultaneous Multithreading” by Faissal M. Sleiman and Thomas F. Wenisch (ISCA 2016)	This paper demonstrates that it is possible to unify in-order and out-of-order issue into a single, integrated, energy-efficient SMT microarchitecture.
“Racer: TSO Consistency via Race Detection” by Alberto Ros and Stefanos Kaxiras (MICRO 2016)	The authors propose a scalable approach to enforce coherence and TSO consistency without directories, timestamps, or software intervention.
“The Anytime Automaton” by Joshua San Miguel and Natalie Enright Jerger (ISCA 2016)	This paper provides a general, safe, and robust approximate computing paradigm that abstracts away the challenge of guaranteeing user acceptability from the system architect.
“Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units” by Siyang Wang, Xiangyu Zhang, Yuxuan Li, Ramin Bashizade, Song Yang, Chris Dwyer, and Alvin R. Lebeck (ISCA 2016)	This paper proposes cross-layer support for probabilistic computing using novel technologies and specialized architectures.
“Stripes: Bit-Serial Deep Neural Network Computing” by Patrick Judd, Jorge Albericio, Tayler Hetherington, Tor M. Aamodt, and Andreas Moshovos (MICRO 2016)	The authors demonstrate that bit-serial computation can lead to high-performance and energy-efficient designs whose performance and accuracy adapts to precision at a fine granularity.
“Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL” by Donggyu Kim, Adam Izraelevitz, Christopher Celio, Hokeun Kim, Brian Zimmer, Yunsup Lee, Jonathan Bachrach, and Krste Asanovic (ISCA 2016)	This paper proposes a sample-based RTL energy modeling methodology for fast and accurate energy evaluation.
“Back to the Future: Leveraging Belady’s Algorithm for Improved Cache Replacement” by Akanksha Jain and Calvin Lin (ISCA 2016)	The authors’ algorithm enhances cache replacement by learning replacement decisions made by Belady. The paper also presents a novel mechanism to efficiently simulate Belady behavior.
“ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars” by Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R. Stanley Williams, and Vivek Srikumar (ISCA 2016)	The authors advance the state of the art in deep network accelerators by an order of magnitude and overcome the challenges of analog-digital conversion with innovative encodings and pipelines suitable for precise and energy-efficient analog acceleration.

Selection Committee

- Tor Aamodt, University of British Columbia
- Alaa Alameldeen, Intel
- Murali Annavaram, University of Southern California
- Todd Austin, University of Michigan
- Chris Batten, Cornell University
- Luis Ceze, University of Washington
- Sandhya Dwarkadas, University of Rochester
- Lieven Eeckhout, Ghent University
- Joel Emer, Nvidia and MIT
- Babak Falsafi, EPFL
- Hyesoon Kim, Georgia Tech
- Nam Sung Kim, University of Illinois at Urbana–Champaign
- Benjamin Lee, Duke University
- Hsien-Hsin Lee, Taiwan Semiconductor Manufacturing Company
- Gabriel Loh, AMD
- Debbie Marr, Intel
- Andreas Moshovos, University of Toronto
- Onur Mutlu, ETH Zurich
- Ravi Nair, IBM
- Milos Prvulovic, Georgia Tech
- Scott Rixner, Rice University
- Eric Rotenberg, North Carolina State University
- Karu Sankaralingam, University of Wisconsin
- Yanos Sazeidas, University of Cyprus
- Simha Sethumadhavan, Columbia University
- Andre Seznec, INRIA
- Dan Sorin, Duke University
- Viji Srinivasan, IBM
- Karin Strauss, Microsoft
- Tom Wenisch, University of Michigan
- Antonia Zhai, University of Minnesota

harsh or generous reviewers, who either give scores at extreme ends or advocate for too few or too many papers from their stack. We wanted to ensure that all reviewers play an equal role in the selection, regardless of their harshness or generosity. For example, we could give all reviewers an equal voice by requiring them to advocate for a fixed number of papers from their stack. We used the data from the past three years' Top Picks meetings to analyze the process for Top Picks and used this data to drive the design of our process. For example, the typical acceptance rate of Top Picks is approximately 10 percent; therefore, if we assign 15 papers to each reviewer, then each reviewer can be expected to have only 1.5 Top Picks papers on average in their stack, and the likelihood of having 5 or more Top Picks papers in the stack would be extremely small.

Based on the data and constraints of Top Picks, we formulated a ranking-based two-phase process. The objective of the first phase was to filter about 35 to 40 papers that would be discussed at the selection committee meeting. The objective of the second phase was to increase the number of reviews per paper to about 10 and ask each reviewer to provide a concrete decision for the assigned paper: whether it should be selected as a Top Picks

or Honorable Mention, or neither. In the first phase, each reviewer was assigned exactly 14 papers and was asked to recommend exactly five papers (Top 5) to the second phase. Each paper received four ratings in this phase. If a paper got three or more ratings of Top 5, it automatically advanced to the second phase. If the paper had two ratings of Top 5, then both positive reviewers had to champion the paper for it to advance to the second phase. Papers with less than two ratings of Top 5 did not advance to the second phase. A total of 38 papers advanced to the second phase, and each such paper got a total of 9 to 10 reviews. In the second phase, each reviewer was assigned an additional seven to eight papers in addition to the four to five papers that survived the first phase. Each reviewer had 12 papers and was asked to place exactly 4 of them into each category: Top Picks, Honorable Mention, and neither.

The selection committee meeting was held in person in Atlanta, Georgia, on 17 December 2016. At the selection committee meeting, the 38 papers were ranked on the basis of the number of Top Picks votes and the average rating the paper received in the second phase. If, after the in-person discussion, 60 percent or more reviewers rated a paper as a Top Pick, then

the paper was selected as a Top Pick. Otherwise, the decision to select the paper as a Top Pick (or Honorable Mention or neither) was made by a committee-wide vote using a simple majority. We observed that the top eight ranked papers all got accepted as Top Picks, and four more papers were selected as Top Picks from the next nine papers. Overall, out of the top 25 papers, all but one was selected as either a Top Pick or an Honorable Mention. Thus, having a large number of reviews per paper reduced the dependency on the in-person discussion. Coincidentally, the day before the selection committee meeting there was a hurricane, which caused many flights to be canceled, and 4 of the 31 selection committee members were unable to attend the meeting. However, having 9 to 10 reviewers per paper still ensured that there were at least eight reviewers present for each paper discussed at the selection committee meeting, resulting in a robust and high-confidence process, even with a relatively high rate of absentees. Given the unique constraints and objectives of Top Picks, we hope that such a process with a larger number of reviews per paper and a process that is robust to variation in generosity levels of reviewers (for example, ranking papers into fixed-sized bins) will be useful for future Top Picks selection committees as well.

Selected Papers

With the slowing down of conventional means for improving performance, the architecture community has been investigating accelerators to improve performance and energy efficiency. This was evident in the emergence of a large number of papers on accelerators appearing throughout the architecture conferences in 2016. Given the emphasis on accelerators, it is no surprise that more than half of the articles in this issue focus on architecting accelerators. Memory system and energy considerations are two other areas from which the Top Picks papers were selected.

Accelerators

Data movement is a primary factor that determines the energy efficiency and effectiveness of accelerators. “Using Dataflow to

Optimize Energy Efficiency of Deep Neural Network Accelerators” by Yu-Hsin Chen and his colleagues describes a spatial architecture that optimizes the dataflow for energy efficiency. This article also has an insightful framework for classifying different accelerators based on access patterns.

“The Memristive Boltzmann Machines” by Mahdi Nazm Bojnordi and Engin Ipek proposes a memory-centric hardware accelerator for combinatorial optimization and deep learning that leverages in-situ computing of bit-line computation in memristive arrays to eliminate the need for exchanging data among the memory arrays and the computational units.

The concept of using analog computing for efficient computation is also explored by Yipeng Huang and colleagues in “Analog Computing in a Modern Context: A Linear Algebra Accelerator Case Study.” The authors try to address the typical challenges faced by analog computing, such as limited problem size, limited dynamic range, and precision.

In contrast to the first three articles, which use domain-specific acceleration, “Domain Specialization Is Generally Unnecessary For Accelerators” by Tony Nowatzki and his colleagues focuses on retaining the programmability of accelerators while maintaining their energy efficiency. The authors use an architecture that has a large number of tiny cores with key building blocks typically required for accelerators and configure these cores intelligently based on the domain requirement.

Large-Scale Accelerators

The next three articles look at enhancing the scalability of accelerators so that they can handle larger problem sizes and cater to varying problem domains. The article “Configurable Clouds” by Adrian Caulfield and his colleagues describes a cloud-scale acceleration architecture that can connect different accelerator nodes within a datacenter using a high-speed FPGA fabric that lets the system accelerate a wide variety of applications and has been deployed in Microsoft datacenters.

In “Specializing a Planet’s Computation: ASIC Clouds,” Moein Khazraee and his colleagues target scale-out workloads comprising many independent but similar jobs, often on

behalf of many users. This architecture shows a way to make ASIC usage more economical, because different users can potentially share the cost of fabricating a given ASIC, rather than each design team incurring the cost of fabricating the ASIC.

“DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric” by Mingyu Gao and his colleagues describes a way to increase the size of FPGA fabrics at low cost by using DRAM instead of SRAM for the storage inside the FPGA, thereby enabling a high-density and low-power reconfigurable fabric.

Memory and Storage Systems

Memory systems continue to be important in determining the performance and efficiency of computer systems. This issue features three articles that focus on improving memory and storage systems. “Agile Paging for Efficient Memory Virtualization” by Jayneel Gandhi and his colleagues addresses the performance overhead of virtual memory in virtualized environments by getting the best of both worlds: nested paging and shadow paging.

Virtual address translation can sometimes affect the correctness of memory consistency models. Daniel Lustig and his colleagues address this problem in their article, “Transistency Models: Memory Ordering at the Hardware–OS Interface.” The authors propose to rigorously integrate memory consistency models and address translation at the microarchitecture and operating system levels.

Moving on to the storage domain, in “Toward a DNA-Based Archival Storage System,” James Bornholt and his colleagues demonstrate DNA-based storage architected as a key-value store. Their design enables random access and is equipped with error correction capability to handle the imperfections of the read and write process. As the demand for cheap storage continues to increase, such alternative technologies have the potential to provide a major breakthrough in storage capability.

Energy Considerations

The final two articles are related to optimizing energy or operating under low energy budgets. Modern processors are provisioned with a timing margin to protect against tem-

perature inversion. In the article “T_i-states: Power Management in Active Timing Margin Processors,” Yazhou Zu and his colleagues show how actively monitoring the temperature on the chip and dynamically reducing this timing margin can result in significant energy savings.

Energy harvesting systems represent an extreme end of energy-constrained computing in which the system performs computing only when the harvested energy is present. One challenge in such systems is to provide debugging functionality for software, because system failure could happen due to either lack of energy or incorrect code. “An Energy-Aware Debugger for Intermittently Powered Systems” by Alexei Colin and his colleagues describes a hardware–software debugger for an intermittent energy-harvesting system that can allow software verification to proceed without getting interference from the energy-harvesting circuit.

We hope you enjoy reading these articles and that you will explore both the original conference versions and the Honorable Mention papers. We welcome your feedback on this special issue and any suggestions for next year’s Top Picks issue. MICRO

Acknowledgments

We thank Lieven Eeckhout for providing support and direction as we tried out the new paper selection process. Lieven also handled the papers that were conflicted with both co-chairs. We also thank the selection committee co-chairs for the past three Top Picks issues (Gabe Loh, Babak Falsafi, Luis Ceze, Karin Strauss, Milo Martin, and Dan Sorin) for providing the review statistics from their editions of Top Picks and for answering our questions. We thank Vinson Young for handling the submission website and Prashant Nair and Jian Huang for facilitating the process at the selection committee meeting. We owe a huge thanks to our fantastic selection committee, which not only diligently reviewed all the papers but also were supportive of the new review process. Furthermore, the selection committee members spent a day attending the in-person meeting in Atlanta, fairly close to the holiday season. Finally, we

thank all the authors who submitted their work for consideration to this Top Picks issue and the authors of the selected papers for producing the final versions of their papers for this issue.

Aamer Jaleel is a principal research scientist at Nvidia. Contact him at ajaleel@nvidia.com.

Moinuddin Qureshi is an associate professor in the School of Electrical and Computer Engineering at Georgia Tech. Contact him at moin@ece.gatech.edu.

myCS Read your subscriptions through the myCS publications portal at <http://mycs.computer.org>.


Recognizing Excellence in High-Performance Computing

Nominations are Solicited for the


**SEYMOUR CRAY
SIDNEY FERNBACH
& KEN KENNEDY AWARDS**

Deadline: 1 July 2017
All nomination details available at awards.computer.org


SEYMOUR CRAY COMPUTER ENGINEERING AWARD
The Seymour Cray Award is awarded to recognize innovative contributions to high-performance computing systems that best exemplify the creative spirit demonstrated by Seymour Cray. The award consists of a crystal memento and honorarium of US\$10,000.



SIDNEY FERNBACH MEMORIAL AWARD
The award, which consists of a certificate and a US\$2,000 honorarium, is presented annually to an individual for "an outstanding contribution in the application of high-performance computers using innovative approaches."



ACM/IEEE-CS KEN KENNEDY AWARD
A certificate and US\$5,000 honorarium are awarded jointly by the ACM and the IEEE Computer Society for outstanding contributions to programmability or productivity in high-performance computing together with significant community service or mentoring contributions.



IEEE computer society Association for Computing Machinery IEEE