TOP PICKS FROM THE 2012 COMPUTER ARCHITECTURE CONFERENCES

Babak Falsafi
EPFL

Gabriel H. Loh
Advanced Micro Devices

In recent years, information technology (IT) has been witnessing major transformations placing unprecedented demands on efficient processing, communication, and storage of data. These transformations are driven by both the slowdown in energy efficiency in modern circuits and fabrication technologies and by the emergence of "big data" requiring not only higher efficiency in manipulating data but also making reliability and security first-class design constraints. With diminishing returns in circuit efficiency, computer architecture innovation is now once more at the core of these transformations and central to the continued growth of IT. We are pleased to bring to you the Top Picks issue this year with 11 articles selected by the program committee and deemed as the best of computer architecture innovation presented in conferences.

Choosing nearly a dozen papers from a highly selective group of submissions that have appeared in flagship conferences is not an easy feat. Our community has also grown to encompass a variety of research areas in response to the emerging design challenges requiring expertise across a diverse spectrum of areas. Nevertheless, with the help of 26 distinguished members of our community, we identified 11 papers out of 78 submissions to be included in this year's issue.

The selection process

We asked authors to submit a three-page summary highlighting the novelty and impact of their work, together with a copy of the original conference paper. There were no constraints regarding the publication venue. As advised by Editor in Chief Erik Altman, and the new submission guidelines set in place in 2011, all conference papers in 2012 with the exception of guest editor papers were eligible for submission. Last year's guest editors were also allowed to submit their 2011 conference papers, as these were not considered last year according to the new guidelines. These guidelines are compatible with our conference submission guidelines and have been put in place to better match our community practices and improve the fairness of the selection process.

This year we had a total of 78 submissions, with 33 percent from ISCA, 27 percent from MICRO, 17 percent from HPCA, and 8 percent from ASPLOS. We also had a smaller representation from CGO, DSN, ICCD, ISPASS, NOCS, PACT, and PLDI. With the help of our selection committee

Published by the IEEE Computer Society

The Selection Committee

- Tor Aamodt, University of British Columbia/Stanford University
- Shubu Mukherjee, Cavium Networks
- Doug Burger, Microsoft Research
- Onur Mutlu, Carnegie Mellon University
- Luis Ceze, University of Washington
- Moin Qureshi, Georgia Institute of Technology
- Sandhya Dwarkadas, University of Rochester
- Parthasarathy Ranganathan, HP Labs
- Lieven Eeckhout, Ghent University
- Karu Sankaralingam, University of Wisconsin
- Natalie Enright Jerger, University of Toronto
- Yanos Sazeides, University of Cyprus
- Kim Hazelwood, Google
- Simha Sethumadhavan, Columbia University
- Hillery Hunter, IBM
- Per Stenström, Chalmers University of Technology
- Wen-Mei Hwu, University of Illinois at Urbana-Champaign
- Michael Taylor, University of California, San Diego
- Aamer Jaleel, Intel
- Josep Torrellas, University of Illinois at Urbana-Champaign
- Christos Kozyrakis, Stanford University
- T.N. Vijaykumar, Purdue University
- Hsien-Hsin Lee, Georgia Institute of Technology/Taiwan Semiconductor Manufacturing Company
- Thomas Wenisch, University of Michigan
- Andreas Moshovos, University of Toronto
- Yuan Xie, Pennsylvania State University/Advanced Micro Devices

(see the "Selection Committee" sidebar) and Boris Grot of EPFL—who both hosted the submission website and helped run a physical program committee meeting—we completed a total of four reviews per paper for all but two lower-ranked papers, which received three reviews.

We designed a review form that included six evaluation criteria, and for each evaluation criteria we included a detailed description of how to assign a score. The criteria were potential impact, novelty of solution or insights, solution feasibility/practicality, soundness of ideas, soundness of evaluation, and reviewer expertise.

We had a physical meeting in Chicago on 12 January 2013 with all PC members present. We discussed 41 papers in total, from which we accepted 11 papers for this issue. The overall ranking in which papers were discussed took a weighted average of the scores given in all criteria, in particular accounting for variations in reviewer expertise and overall scoring generosity (similar to averaging techniques used by Onur Mutlu in the MICRO 2012 conference).

The 2012 Top Picks articles

For this year's Top Picks issue, we include 11 articles that demonstrate the breadth and depth of ongoing computer architecture research (see the "Top Picks of 2012" sidebar). These articles fall into three themes: energy and efficiency, safety and security, and parallelism and memory.

Energy and efficiency

The first theme directly targets energy-efficiency improvements in modern platforms with contributions that target better performance with existing power budgets to exploit the tradeoff in computational precision and variations in demands for quality of service to save energy. In "Designing for Responsiveness with Computational Sprinting," Arun Raghavan et al. propose to incorporate phase-change materials directly into a processor die that provides a limited amount of thermal capacitance. This additional thermal headroom allows for short bursts, or *sprints*, of computation without the risk of overheating and damaging the processor.

Many important classes of workloads in light of the emergence of Big Data (such as machine learning and media processing) are probabilistic in nature with acceptable approximate rather than precise outcomes. Conventional general-purpose processors for such workloads would lead to overly precise results, wasting both computational resources and energy. In "Neural Acceleration for General-Purpose Approximate Programs," Hadi Esmaeilzadeh et al. describe a framework from compiler to hardware for replacing certain regions of code with invocations to neural-processing accelerators that generate approximately correct results, but at significant energy and latency savings.

Modern servers provide limited levels of energy proportionality and as such waste energy during periods of low server activity and

Top Picks of 2012

Energy and efficiency

- ''Designing for Responsiveness with Computational Sprinting'' by Arun Raghavan, Yixin Luo, Anuj Chandawalla, Marios Papaefthymiou, Kevin P. Pipe, Thomas F. Wenisch, and Milo M. K. Martin
- ''Neural Acceleration for General-Purpose Approximate Programs'' by Hadi Esmaeilzadeh, Adrian Sampson, Luis Ceze, and Doug Burger
- ''Scaling the Energy Proportionality Wall with KnightShift'' by Daniel Wong and Murali Annavaram

Safety and security

- ''Hardware-Enforced Comprehensive Memory Safety'' by Santosh Nagarakatte, Milo M. K. Martin, and Steve Zdancewic
- ''Inspection-Resistant Memory Architectures'' by Jonathan Kaveh Valamehr, Melissa Chase, Seny Kamara, Andrew Putnam, Daniel Shumow, Vinod Vaikuntanathan, and Timothy Sherwood
- ''Relyzer: Application Resiliency Analyzer for Transient Faults'' by Siva Kumar Sastry Hari, Sarita V. Adve, Helia Naeimi, and Pradeep Ramachandran
- ''A Quantitative, Experimental Approach to Measuring Side-Channel Security'' by John Demme, Robert Martin, Adam Waksman, and Simha Sethumadhavan

Parallelism and memory

- ''Cache-Conscious Thread Scheduling for Massively Multithreaded Processors'' by Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt
- ''Parallel Block Vectors: Collection, Analysis, and Uses'' by Melanie Kambadur, Kui Tang, and Martha A. Kim
- ''A Safety-First Approach to Memory Models'' by Abhayendra Singh, Satish Narayanasamy, Daniel Marino, Todd Millstein, and Madanlal Musuvathi
- ''Programmable DDRx Controllers'' by Mahdi Nazm Bojnordi and Engin Ipek

request rate. In ''Scaling the Energy Proportionality Wall with KnightShift,'' Daniel Wong and Murali Annavaram present a case for server-level heterogeneity through energy-efficient server proxies during low-activity periods while maximizing performance at moderate to peak loads.

Safety and security

As IT systems scale to cope with ever-increasing demands, security and safety take center stage as computer system design challenges. The second theme highlights the central role computer architecture plays in protecting computation and data. Many security vulnerabilities arise from simple memory management bugs. In ''Hardware-Enforced Comprehensive Memory Safety,'' Santosh Nagarakatte, Milo M. K. Martin, and Steve Zdancewic propose hardware to generate and automatically check unique identifiers for pointers to catch illegal memory accesses that would otherwise lead to a security vulnerability.

In ''Inspection-Resistant Memory Architectures,'' Jonathan Kaveh Valamehr et al. tackle the difficult problem of securing a memory system from direct physical inspection (for example, electron microscopes). They introduce a hybrid scheme combining concepts from secret sharing and coding to create a robust memory system.

Transient faults, if undetected, can also lead to silent data corruptions (SDCs) and comprise both computational outcome and data integrity. In ''Relyzer: Application Resiliency Analyzer for Transient Faults,'' Siva Kumar Sastry Hari et al. provide a new approach that carefully analyzes an application to select a few fault-injection sites that continue to provide high levels of fault coverage while reducing the required detailed simulation time to identify SDCs by orders of magnitude.

A class of security attacks on computer systems exploits program execution side effects. In ''A Quantitative, Experimental Approach to Measuring Side-Channel Security,'' John Demme et al. introduce a quantitative methodology for measuring the potential for information leakage through side effects. Analysis of a system's side-channel vulnerability factor enables designers to better make tradeoffs between performance and security.

Parallelism and memory

Finally, the third theme focuses on parallelism, memory systems, and the interaction of the two to tackle efficiency in modern

IT platforms. In "Cache-Conscious Thread Scheduling for Massively Multithreaded Processors," Timothy G. Rogers, Mike O'Connor, and Tor M. Aamodt examine the problem of cache interference among a GPU's numerous threads of computation. Rather than optimizing the cache-replacement policy, this article proposes new locality-aware GPU scheduling to shape the cache traffic to make better use of the GPU's caches.

To make good use of multicore processors, programmers must understand the relationship between their parallel code and the actual concurrency it exposes at runtime. The article "Parallel Block Vectors: Collection, Analysis, and Uses" by Melanie Kambadur, Kui Tang, and Martha A. Kim presents a new concept of parallel block vectors that provides a means of mapping from a running program's concurrency phases back to the original source code.

Prior work has argued that relaxed memory consistency models might not be necessary to achieve high performance. These designs offered sequential consistency through relaxing memory speculatively in hardware. In "A Safety-First Approach to Memory Models," Abhayendra Singh et al. observe that thread-local data and read-only memory regions do not need hardware support for sequential consistency. They present a hardware design that exploits this observation to provide efficient support for sequential consistency without hardware speculation.

Finally, "Programmable DDRx Controllers" by Mahdi Nazm Bojnordi and Engin Ipek tackles the problem of building customizable controllers for modern DDR memory chips. By introducing programmability into the memory controller design, this work presents a new approach that enables more flexible and efficient memory systems.

We are pleased to have brought you this year's edition of Top Picks. We hope that you enjoy reading these articles as well as their original conference versions, and we welcome any feedback you may have on this issue. MICRO

Babak Falsafi is a professor of computer and communication sciences at EPFL, and the founding director of EcoCloud, an interdisciplinary research center targeting robust, economic, and environmentally friendly cloud technologies. Falsafi has a PhD in computer science from the University of Wisconsin—Madison. He is a fellow of IEEE.

Gabriel H. Loh is a fellow design engineer at AMD Research, the research and advanced development lab for Advanced Micro Devices. His research interests include computer architecture, processor microarchitecture, memory systems, emerging technologies, and 3D die stacking. Loh has a PhD in computer science from Yale University. He is a senior member of IEEE.