2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA 2018)

Los Angeles, California, USA 1-6 June 2018



IEEE Catalog Number: ISBN:

CFP18030-POD 978-1-5386-5985-4

Copyright © 2018 by the Institute of Electrical and Electronics Engineers, Inc. All Rights Reserved

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

*** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.

IEEE Catalog Number: ISBN (Print-On-Demand): ISBN (Online): ISSN: CFP18030-POD 978-1-5386-5985-4 978-1-5386-5984-7 1063-6897

Additional Copies of This Publication Are Available From:

Curran Associates, Inc 57 Morehouse Lane Red Hook, NY 12571 USA Phone: (845) 758-0400 Fax: (845) 758-2633 E-mail: curran@proceedings.com Web: www.proceedings.com



2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture ISCA 2018

Table of Contents

Message from the ISCA 2018 General Co-Chairs .xiv
Message from the ISCA 2018 Program Chair xxi
ISCA 2018 Organizing Committee .xx
ISCA 2018 Program Committee xxi
ISCA 2018 External Reviewers .xxiii
ISCA 2018 Sponsors xxy
ACM Turing Award Lecture Abstract xxvii

Session 1A: Clouds & Datacenters

A Configurable Cloud-Scale DNN Processor for Real-Time AI .1..... Jeremy Fowers (Microsoft), Kalin Ovtcharov (Microsoft), Michael Papamichael (Microsoft), Todd Massengill (Microsoft), Ming Liu (Microsoft), Daniel Lo (Microsoft), Shlomi Alkalay (Microsoft), Michael Haselman (Microsoft), Logan Adams (Microsoft), Mahdi Ghandi (Microsoft), Stephen Heil (Microsoft), Prerak Patel (Microsoft), Adam Sapek (Microsoft), Gabriel Weisz (Microsoft), Lisa Woods (Microsoft), Sitaram Lanka (Microsoft), Steven K. Reinhardt (Microsoft), Adrian M. Caulfield (Microsoft), Eric S. Chung (Microsoft), and Doug Burger (Microsoft)

Virtual Melting Temperature: Managing Server Load to Minimize Cooling Overhead with Phase Change

Materials .15.

Matt Skach (University of Michigan), Manish Arora (Advanced Micro Devices & University of California, San Diego), Dean Tullsen (University of California, San Diego), Lingjia Tang (University of Michigan), and Jason Mars (University of Michigan) FireSim: FPGA-Accelerated Cycle-Exact Scale-Out System Simulation in the Public Cloud .29..... Sagar Karandikar (University of California, Berkeley), Howard Mao (University of California, Berkeley), Donggyu Kim (University of California, Berkeley), David Biancolin (University of California, Berkeley), Alon Amid (University of California, Berkeley), Dayeol Lee (University of California, Berkeley), Nathan Pemberton (University of California, Berkeley), Emmanuel Amaro (University of California, Berkeley), Colin Schmidt (University of California, Berkeley), Aditya Chopra (University of California, Berkeley), Qijing Huang (University of California, Berkeley), Kyle Kovacs (University of California, Berkeley), Borivoje Nikolic (University of California, Berkeley), Randy Katz (University of California, Berkeley), Jonathan Bachrach (University of California, Berkeley), and Krste Asanovic (University of California, Berkeley)

Session 1B: Accelerators for Emerging Apps

PROMISE: An End-to-End Design of a Programmable Mixed-Signal Accelerator for Machine-Learning Algorithms 43.
Prakalp Srivastava (University of Illinois at Urbana-Champaign), Mingu Kang (IBM Thomas J. Watson Research Center), Sujan K. Gonugondla (University of Illinois at Urbana-Champaign), Sungmin Lim (University of Illinois at Urbana-Champaign), Jungwook Choi (IBM Thomas J. Watson Research Center), Vikram Adve (University of Illinois at Urbana-Champaign), Nam Sung Kim (University of Illinois at Urbana-Champaign), and Naresh Shanbhag (University of Illinois at Urbana-Champaign)
Computation Reuse in DNNs by Exploiting Input Similarity .5.7
Marc Riera (Universitat Politecnica de Catalunya), Jose-Maria Arnau
(Universitat Politecnica de Catalunya), and Antonio Gonzalez
(Universitat Politecnica de Catalunya)
GenAx: A Genome Sequencing Accelerator .69.
Daichi Fujiki (University of Michigan), Arun Subramaniyan (University
of Michigan), Tianjun Zhang (University of Michigan), Yu Zeng
(University of Michigan), Reetuparna Das (University of Michigan),
David Blaauw (University of Michigan), and Satish Narayanasamy
(University of Michigan)

Session 2A: Prefetching

Division of Labor: A More Effective Approach to Prefetching .83 Sushant Kondguli (University of Rochester) and Michael Huang (University of Rochester)
Criticality Aware Tiered Cache Hierarchy: A Fundamental Relook at Multi-Level Cache Hierarchies .96
Anant Vithal Nori (Intel Microarchitecture Research Lab), Jayesh Gaur
(Intel Microarchitecture Research Lab), Siddharth Rai (Indian
Institute of Technology Kanpur), Sreenivas Subramoney (Intel
Microarchitecture Research Lab), and Hong Wang (Intel
Microarchitecture Research Lab)

Rethinking Belady's Algorithm to Accommodate Prefetching .110..... Akanksha Jain (University of Texas at Austin) and Calvin Lin (University of Texas at Austin)

Session 2B: Languages & Models

Constructing a Weak Memory Model .124 Sizhuo Zhang (Massachusetts Institute of Technology), Muralidaran Vijayaraghavan (Massachusetts Institute of Technology), Andrew Wright (Massachusetts Institute of Technology), Mehdi Alipour (Uppsala University), and _Arvind (Massachusetts Institute of Technology)
A Hardware Accelerator for Tracing Garbage Collection .138 Martin Maas (University of California, Berkeley), Krste Asanovi (University of California, Berkeley), and John Kubiatowicz (University of California, Berkeley)
Charm: A Language for Closed-Form High-Level Architecture Modeling .152 Weilong Cui (University of California, Santa Barbara), Yongshan Ding (University of Chicago), Deeksha Dangwal (University of California, Santa Barbara), Adam Holmes (University of Chicago), Joseph McMahan (University of California, Santa Barbara), Ali Javadi-Abhari (IBM Research), Georgios Tzimpragos (University of California, Santa Barbara), Frederic Chong (University of Chicago), and Timothy Sherwood (University of California, Santa Barbara)

Session 3A: Virtual Memory

Get Out of the Valley: Power-Efficient Address Mapping for GPUs .166 Yuxi Liu (Ghent University and Peking University), Xia Zhao (Ghent University), Magnus Jahre (Norwegian University of Science and Technology), Zhenlin Wang (Michigan Technological University), Xiaolin Wang (Peking University), Yingwei Luo (Peking University), and Lieven Eeckhout (Ghent University)
Scheduling Page Table Walks for Irregular GPU Applications .180 Seunghee Shin (North Carolina State University), Guilherme Cox (Rutgers University), Mark Oskin (Advanced Micro Devices), Gabriel H. Loh (Advanced Micro Devices), Yan Solihin (North Carolina State University), Abhishek Bhattacharjee (Rutgers University), and Arkaprava Basu (Indian Institute of Science)
SEESAW: Using Superpages to Improve VIPT Caches .193 Mayank Parasar (Georgia Institute of Technology), Abhishek Bhattacharjee (Rutgers University), and Tushar Krishna (Georgia Institute of Technology)
A Case for Richer Cross-Layer Abstractions: Bridging the Semantic Gap with Expressive Memory .207 Nandita Vijaykumar (Carnegie Mellon University), Abhilasha Jain (Carnegie Mellon University), Diptesh Majumdar (Carnegie Mellon University), Kevin Hsieh (Carnegie Mellon University), Gennady Pekhimenko (University of Toronto), Eiman Ebrahimi (NVIDIA), Nastaran Hajinazar (Simon Fraser University), Phillip B. Gibbons (Carnegie Mellon University), and Onur Mutlu (ETH Zurich)

Session 3B: Coherence & Memory Ordering

Non-Speculative Store Coalescing in Total Store Order .221. Alberto Ros (University of Murcia) and Stefanos Kaxiras (Uppsala University)	
Dynamic Memory Dependence Predication .235.	
Zhaoxiang Jin (Michigan Technological University) and Soner Önder	
(Michigan Technological University)	
ProtoGen: Automatically Generating Directory Cache Coherence Protocols from Atomic Specifications . Nicolai Oswald (University of Edinburgh), Vijay Nagarajan (University of Edinburgh), and Daniel J. Sorin (Duke University)	247
Spandex: A Flexible Interface for Efficient Heterogeneous Coherence .261	
Johnathan Alsop (University of Illinois at Urbana-Champaign), Matthew	
Sinclair (University of Illinois at Urbana-Champaign & University of	
Wisconsin-Madison), and Sarita Adve (University of Illinois at	
Urbana-Champaign)	

Session 4A: Emerging Paradigms

Flexon: A Flexible Digital Neuron for Efficient Spiking Neural Network Simulations .275
Dayeol Lee (University of California, Berkeley), Gwangmu Lee (Seoul
National University), Dongup Kwon (Seoul National University), Sunghwa
Lee (Seoul National University), Youngsok Kim (Seoul National
University), and Jangwoo Kim (Seoul National University)
Space-Time Algebra: A Model for Neocortical Computation .289.
James Smith (University of Wisconsin-Madison [Emeritus])
Architecting a Stochastic Computing Unit with Molecular Optical Devices .301
Xiangyu Zhang (Duke University), Ramin Bashizade (Duke University),
Craig LaBoda (Duke University), Chris Dwyer (Parabon Labs), and Alvin

R. Lebeck (Duke University)

Session 4B: Persistence

(Georgia Institute of Technology)

RANA: Towards Efficient Neural Acceleration with Refresh-Optimized Embedded DRAM .340..... Fengbin Tu (Tsinghua University), Weiwei Wu (Tsinghua University), Shouyi Yin (Tsinghua University), Leibo Liu (Tsinghua University), and Shaojun Wei (Tsinghua University)

Session 5A: Emerging Memory 1

Scaling Datacenter Accelerators with Compute-Reuse Architectures .353 Adi Fuchs (Princeton University) and David Wentzlaff (Princeton University)	
Enabling Scientific Computing on Memristive Accelerators .367 Ben Feinberg (University of Rochester), Uday Kumar Reddy Vengalam (University of Rochester) Nathan Whitehair (University of Rochester)	
Shibo Wang (University of Rochester), and Engin Ipek (University of Rochester), Rochester)	
Neural Cache: Bit-Serial In-Cache Acceleration of Deep Neural Networks .383 Charles Eckert (University of Michigan), Xiaowei Wang (University of Michigan), Jingcheng Wang (University of Michigan), Arun Subramaniyan (University of Michigan), Ravi Iyer (Intel Corporation), Dennis Sylvester (University of Michigan), David Blaaauw (University of Michigan), and Reetuparna Das (University of Michigan)	

Session 5B: Storage

FLIN: Enabling Fairness and Enhancing Performance in Modern NVMe Solid State Drives .397
Arash Tavakkol (ETH Zürich), Mohammad Sadrosadati (ETH Zürich),
Saugata Ghose (Carnegie Mellon University), Jeremie Kim (Carnegie
Mellon University & ETH Zürich), Yixin Luo (Carnegie Mellon
University), Yaohua Wang (ETH Zürich & NUDT), Nika Mansouri Ghiasi
(ETH Zürich), Lois Orosa (ETH Zürich & Unicamp), Juan Gómez-Luna (ETH
Zürich), and Onur Mutlu (ETH Zürich & Carnegie Mellon University)
GraFBoost: Using Accelerated Flash Storage for External Graph Analytics .411

Sang-Woo Jun (Massachusetts Institute of Technology), Andy Wright (Massachusetts Institute of Technology), Sizhuo Zhang (Massachusetts Institute of Technology), Shuotao Xu (Massachusetts Institute of Technology), and _ Arvind (Massachusetts Institute of Technology)

2B-SSD: The Case for Dual, Byte- and Block-Addressable Solid-State Drives .425..... Duck-Ho Bae (Samsung Electronics), Insoon Jo (Samsung Electronics), Youra Adel Choi (Samsung Electronics), Joo-Young Hwang (Samsung Electronics), Sangyeun Cho (Samsung Electronics), Dong-Gi Lee (Samsung Electronics), and Jaeheon Jeong (Samsung Electronics)

Session 6A: Emerging Memory 2

Lazy Persistency: A High-Performing and Write-Efficient Software Persistency Technique .439...... Mohammad Alshboul (North Carolina State University), James Tuck (North Carolina State University), and Yan Solihin (North Carolina State University) DHTM: Durable Hardware Transactional Memory .452. Arpit Joshi (University of Edinburgh), Vijay Nagarajan (University of Edinburgh), Marcelo Cintra (Intel), and Stratis Viglas (Google)

Hardware Supported Permission Checks on Persistent Objects for Performance and Programmability .466..... *Tiancong Wang (North Carolina State University), Sakthikumaran Sambasivam (North Carolina State University), and James Tuck (North Carolina State University)*

Session 6B: Controllers & Control Systems

RoboX: An End-to-End Solution to Accelerate Autonomous Control in Robotics .479..... Jacob Sacks (Georgia Institute of Technology), Divya Mahajan (Georgia Institute of Technology), Richard C. Lawson (Georgia Institute of Technology), and Hadi Esmaeilzadeh (University of California, San Diego)

DCS-ctrl: A Fast and Flexible Device-Control Mechanism for Device-Centric Server Architecture .491..... Dongup Kwon (Seoul National University), Jaehyung Ahn (POSTECH), Dongju Chae (POSTECH), Mohammadamin Ajdari (POSTECH), Jaewon Lee (Seoul National University), Suheon Bae (Seoul National University), Youngsok Kim (Seoul National University), and Jangwoo Kim (Seoul National University)

Yukta: Multilayer Resource Controllers to Maximize Efficiency .505...... Raghavendra Pradyumna Pothukuchi (University of Illinois at Urbana-Champaign), Sweta Yamini Pothukuchi (University of Illinois at Urbana-Champaign), Petros Voulgaris (University of Illinois at Urbana-Champaign), and Josep Torrellas (University of Illinois at Urbana-Champaign)

Session 7A: Mobile Platforms

Stitch: Fusible Heterogeneous Accelerators Enmeshed with Many-Core Architecture for Wearables .5.75...... Cheng Tan (National University of Singapore), Manupa Karunaratne (National University of Singapore), Tulika Mitra (National University of Singapore), and Li-Shiuan Peh (National University of Singapore)

Session 7B: Security

Nonblocking Memory Refresh .588 Kate Nguyen (Virginia Tech), Kehan Lyu (Virginia Tech), Xianze Meng (Virginia Tech), Vilas Sridharan (Advanced Micro Devices), and Xun Jian (Virginia Tech)
Practical Memory Safety with REST .600 Kanad Sinha (Columbia University) and Simha Sethumadhavan (Columbia University)
Mitigating Wordline Crosstalk Using Adaptive Trees of Counters .6.12 Seyed Mohammad Seyedzadeh (University of Pittsburgh), Alex K. Jones (University of Pittsburgh), and Rami Melhem (University of Pittsburgh)
Mobilizing the Micro-Ops: Exploiting Context Sensitive Decoding for Security and Energy Efficiency .624 Mohammadkazem Taram (University of California, San Diego), Ashish Venkat (University of California, San Diego), and Dean Tullsen (University of California, San Diego)
 Hiding Intermittent Information Leakage with Architectural Support for Blinking .638 Alric Althoff (University of California, San Diego), Joseph McMahan (University of California, Santa Barbara), Luis Vega (University of Washington), Scott Davidson (University of Washington), Timothy Sherwood (University of California, Santa Barbara), Michael Taylor (University of Washington), and Ryan Kastner (University of California, San Diego)

Session 8A: Machine Learning Systems 1

GANAX: A Unified MIMD-SIMD Acceleration for Generative Adversarial Networks .650...... Amir Yazdanbakhsh (Georgia Institute of Technology), Kambiz Samadi (Qualcomm Technologies), Nam Sung Kim (University of Illinois at Urbana-Champaign), and Hadi Esmaeilzadeh (University of California, San Diego)

SnaPEA: Predictive Early Activation for Reducing Computation in Deep Convolutional Neural Networks .662
Vahideh Akhlaghi (University of California, San Diego), Amir
Yazdanbakhsh (Georgia Institute of Technology), Kambiz Samadi
(Qualcomm Technologies), Rajesh K. Gupta (University of California,
San Diego), and Hadi Esmaeilzadeh (University of California, San
Diego)

UCNN: Exploiting Computational Reuse in Deep Neural Networks via Weight Repetition .6.7.4..... Kartik Hegde (University of Illinois at Urbana-Champaign), Jiyong Yu (University of Illinois at Urbana-Champaign), Rohit Agrawal (University of Illinois at Urbana-Champaign), Mengjia Yan (University of Illinois at Urbana-Champaign), Michael Pellauer (NVIDIA), and Christopher Fletcher (University of Illinois at Urbana-Champaign) Energy-Efficient Neural Network Accelerator Based on Outlier-Aware Low-Precision Computation .688...... Eunhyeok Park (Seoul National University), Dongyoung Kim (Seoul National University), and Sungjoo Yoo (Seoul National University)

Session 8B: Interconnection Networks

Synchronized Progress in Interconnection Networks (SPIN): A New Theory for Deadlock Freedom .699 Aniruddh Ramrakhyani (Georgia Institute of Technology), Paul V. Gratz (Texas A&M University), and Tushar Krishna (Georgia Institute of Technology)
TCEP: Traffic Consolidation for Energy-Proportional High-Radix Networks .7.12 Gwangsun Kim (Arm Research), Hayoung Choi (KAIST), and John Kim (KAIST)
Modular Routing Design for Chiplet-Based Systems .726. Jieming Yin (Advanced Micro Devices), Zhifeng Lin (University of Southern California), Onur Kayiran (Advanced Micro Devices), Matthew Poremba (Advanced Micro Devices), Muhammad Shoaib Bin Altaf (Advanced Micro Devices), Natalie Enright Jerger (University of Toronto), and Gabriel H. Loh (Advanced Micro Devices)
FastTrack: Leveraging Heterogeneous FPGA Wires to Design Low-Cost High-Performance Soft NoCs .739 Nachiket Kapre (University of Waterloo) and Tushar Krishna (Georgia

Institute of Technology)

Session 9A: Machine Learning Systems 2

Prediction Based Execution on Deep Neural Networks 752. <i>Mingcong Song (University of Florida), Jiechen Zhao (University of Florida), Yang Hu (University of Texas at Dallas), Jiaqi Zhang (University of Florida), and Tao Li (University of Florida)</i>
Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Network .764 Hardik Sharma (Georgia Institute of Technology), Jongse Park (Georgia Institute of Technology), Naveen Suda (Arm, Inc.), Liangzhen Lai (Arm, Inc.), Benson Chau (Georgia Institute of Technology), Vikas Chandra (Arm, Inc.), and Hadi Esmaeilzadeh (University of California, San Diego)
Gist: Efficient Data Encoding for Deep Neural Network Training .7.76 Animesh Jain (University of Michigan), Amar Phanishayee (Microsoft Research), Jason Mars (University of Michigan), Lingjia Tang (University of Michigan), and Gennady Pekhimenko (University of Toronto)
The Dark Side of DNN Pruning .790 Reza Yazdani (Universitat Politecnica de Catalunya), Marc Riera (Universitat Politecnica de Catalunya), Jose-Maria Arnau (Universitat Politecnica de Catalunya), and Antonio González (Universitat Politecnica de Catalunya)

Session 9B: GPUs

HetCore: TFET-CMOS Hetero-Device Architecture for CPUs and GPUs .802 Bhargava Gopireddy (University of Illinois at Urbana-Champaign), Dimitrios Skarlatos (University of Illinois at Urbana-Champaign), Wenjuan Zhu (University of Illinois at Urbana-Champaign), and Josep Torrellas (University of Illinois at Urbana-Champaign)
RegMutex: Inter-Warp GPU Register Time-Sharing .816. <i>Farzad Khorasani (Georgia Institute of Technology), Hodjat Asghari</i> <i>Esfeden (University of California Riverside), Amin Farmahini-Farahani</i> <i>(AMD Research), Nuwan Jayasena (AMD Research), and Vivek Sarkar</i> <i>(Georgia Institute of Technology)</i>
The Locality Descriptor: A Holistic Cross-Layer Abstraction to Express Data Locality In GPUs .829 Nandita Vijaykumar (Carnegie Mellon University), Eiman Ebrahimi (NVIDIA), Kevin Hsieh (Carnegie Mellon University), Phillip B. Gibbons (Carnegie Mellon University), and Onur Mutlu (ETH Zurich)
Generic System Calls for GPUs .843 Ján Veselý (Rutgers University), Arkaprava Basu (Indian Institute of Science), Abhishek Bhattacharjee (Rutgers University), Gabriel H. Loh (Advanced Micro Devices), Mark Oskin (University of Washington & Advanced Micro Devices), and Steven K. Reinhardt (Microsoft)

Author Index 857.