

2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2016)

**Taipei, Taiwan
15 – 19 October 2016**



**IEEE Catalog Number: CFP16071-POD
ISBN: 978-1-5090-3509-0**

**Copyright © 2016 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 publication is a representation of what appears in the IEEE Digital Libraries. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP16071-POD
ISBN (Print-On-Demand):	978-1-5090-3509-0
ISBN (Online):	978-1-5090-3508-3

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

CURRAN ASSOCIATES INC.
proceedings
.com

Program

Session 1a: Microarchitecture

Dictionary Sharing: An Efficient Cache Compression Scheme for Compressed Caches

Biswabandan Panda (INRIA), André Seznec (INRIA).....1

Perceptron Learning for Reuse Prediction, Elvira Teran (Texas A&M University), Zhe Wang (Intel), Daniel A. Jiménez (Texas A&M University).....13

pTask: A Smart Prefetching Scheme for OS Intensive Applications, Prathmesh Kallurkar (Indian Institute of Technology, New Delhi), Smruti R. Sarangi (Indian Institute of Technology, New Delhi).....25

Register Sharing for Equality Prediction, Arthur Perais (INRIA/IRISA), Fernando A. Endo (INRIA/IRISA), André Seznec (INRIA/IRISA).....37

Data-Centric Execution of Speculative Parallel Programs, Mark C. Jeffrey (MIT), Suvinay Subramanian (MIT), Maleen Abeydeera (MIT), Joel Emer (NVIDIA/MIT), Daniel Sanchez (MIT).....49

Session 1b: Cloud & Storage

SABRes: Atomic Object Reads for In-Memory Rack-Scale Computing, Alexandros Daglis (EPFL), Dmitrii Ustiugov (EPFL), Stanko Novaković (EPFL), Edouard Bugnion (EPFL), Babak Falsafi (EPFL), Boris Grot (University of Edinburgh).....62

A Cloud-Scale Acceleration Architecture, Adrian M. Caulfield (Microsoft), Eric S. Chung (Microsoft), Andrew Putnam (Microsoft), Hari Angepat (Microsoft), Jeremy Fowers (Microsoft), Michael Haselman (Microsoft), Stephen Heil (Microsoft), Matt Humphrey (Microsoft), Puneet Kaur (Microsoft), Joo-Young Kim (Microsoft), Daniel Lo (Microsoft), Todd Massengill (Microsoft), Kalin Ovtcharov (Microsoft), Michael Papamichael (Microsoft), Lisa Woods (Microsoft), Sitaram Lanka (Microsoft), Derek Chiou (Microsoft), Doug Burger (Microsoft).....75

Towards Efficient Server Architecture for Virtualized Network Function Deployment:

Implications and Implementations, Yang Hu (University of Florida), and Tao Li (University of Florida).....88

Bridging the I/O Performance Gap for Big Data Workloads: A New NVDIMM-based Approach

Renhai Chen (The Hong Kong Polytechnic University), Zili Shao (The Hong Kong Polytechnic University), Tao Li (University of Florida).....100

NeSC: Self-Virtualizing Nested Storage Controller, Yonatan Gottesman (Technion-Israel Institute of Technology), Yoav Etsion (Technion-Israel Institute of Technology).....112

Session 2a: GPU

MIMD Synchronization on SIMT Architectures, Ahmed ElTantawy (University of British Columbia), Tor M. Aamodt (University of British Columbia).....124

Efficient Kernel Synthesis for Performance Portable Programming, Li-Wen Chang (University of Illinois at Urbana-Champaign), Izzat El Hajj (University of Illinois at Urbana-Champaign), Christopher Rodrigues (Huawei), Juan Gómez-Luna (University of Córdoba), Wen-mei Hwu (University of Illinois at Urbana-Champaign).....138

KLAP: Kernel Launch Aggregation and Promotion for Optimizing Dynamic Parallelism, Izzat El Hajj (University of Illinois at Urbana-Champaign), Juan Gómez-Luna (University of Córdoba), Cheng Li (University of Illinois at Urbana-Champaign), Li-Wen Chang (University of Illinois at Urbana-Champaign), Dejan Milojicic (Hewlett-Packard), Wen-mei Hwu (University of Illinois at Urbana-Champaign).....151

Cache-Emulated Register File: An Integrated On-Chip Memory Architecture for High

Performance GPGUs, Naifeng Jing (Shanghai Jiao Tong University), Jianfei Wang (Shanghai Jiao Tong University), Fengfeng Fan (Shanghai Jiao Tong University), Wenkang Yu (Shanghai Jiao Tong University), Li Jiang (Shanghai Jiao Tong University), Chao Li (Shanghai Jiao Tong University), Xiaoyao Liang (Shanghai Jiao Tong University).....163

Zorua: A Holistic Approach to Resource Virtualization in GPUs, Nandita Vijaykumar (Carnegie Mellon University), Kevin Hsieh (Carnegie Mellon University), Gennady Pekhimenko (Carnegie Mellon University), Samira Khan (University of Virginia), Ashish Shrestha (Carnegie Mellon University), Saugata Ghose (Carnegie Mellon University), Adwait Jog (College of William and Mary), Phillip B. Gibbons (Carnegie Mellon University), Onur Mutlu (ETH Zürich and Carnegie Mellon University).....175

GRAPE: Minimizing Energy for GPU Applications with Performance Requirements, Muhammad Husni Santrijaji (Surya University & University of Chicago), Henry Hoffmann (University of Chicago).....189

Session 2b: Neural Networks

From High-Level Deep Neural Models to FPGAs, Hardik Sharma (Georgia Institute of Technology), Jongse Park (Georgia Institute of Technology), Divya Mahajan (Georgia Institute of Technology), Emmanuel Amaro (Georgia Institute of Technology), Joon Kyung Kim (Georgia Institute of Technology), Chenkai Shao (Georgia Institute of Technology), Asit Mishra (Intel), Hadi Esmailzadeh (Georgia Institute of Technology).....202

vDNN: Virtualized Deep Neural Networks for Scalable, Memory-Efficient Neural Network Design, Minsoo Rhu (NVIDIA), Natalia Gimelshein (NVIDIA), Jason Clemons (NVIDIA), Arslan Zulfiqar (NVIDIA), Stephen W. Keckler (NVIDIA).....214

Stripes: Bit-Serial Deep Neural Network Computing, Patrick Judd (University of Toronto), Jorge Albericio (University of Toronto), Tayler Hetherington (University of British Columbia), Tor M. Aamodt (University of British Columbia), Andreas Moshovos (University of Toronto).....227

Cambricon-X: An Accelerator for Sparse Neural Networks, Shijin Zhang (Chinese Academy of Sciences), Zidong Du (Chinese Academy of Sciences), Lei Zhang (Chinese Academy of Sciences), Huiying Lan (Chinese Academy of Sciences), Shaoli Liu (Chinese Academy of Sciences), Ling Li (Chinese Academy of Sciences), Qi Guo (Chinese Academy of Sciences), Tianshi Chen (Chinese Academy of Sciences), Yunji Chen (Chinese Academy of Sciences).....239

NEUTRAMS: Neural Network Transformation and Co-design under Neuromorphic Hardware Constraints, Yu Ji (Tsinghua University), Youhui Zhang (Tsinghua University), ShuangChen Li (University of California, Santa Barbara), Ping Chi (University of California, Santa Barbara), CiHang Jiang (Tsinghua University), Peng Qu (Tsinghua University), Yuan Xie (University of California, Santa Barbara), WenGuang Chen (Tsinghua University).....251

Fused-Layer CNN Accelerators, Manoj Alwani (Stony Brook University), Han Chen (Stony Brook University), Michael Ferdman (Stony Brook University), Peter Milder (Stony Brook University).....264

Session 3a: Compilation & Memory

Continuous Shape Shifting: Enabling Loop Co-optimization via Near-Free Dynamic Code Rewriting, Animesh Jain (University of Michigan, Ann Arbor), Michael A. Laurenzano (University of Michigan, Ann Arbor), Lingjia Tang (University of Michigan, Ann Arbor), Jason Mars (University of Michigan, Ann Arbor).....276

CrystalBall: Statically Analyzing Runtime Behavior via Deep Sequence Learning, Stephen Zekany (University of Michigan, Ann Arbor), Daniel Rings (University of Michigan, Ann Arbor), Nathan Harada (University of Michigan, Ann Arbor), Michael A. Laurenzano (University of Michigan, Ann Arbor; Clinc), Lingjia Tang (University of Michigan, Ann Arbor; Clinc), Jason Mars (University of Michigan, Ann Arbor; Clinc).....288

Low-Cost Soft Error Resilience with Unified Data Verification and Fine-Grained Recovery for Acoustic Sensor Based Detection, Qingrui Liu (Virginia Tech), Changhee Jung (Virginia Tech, Blacksburg), Dongyoon Lee (Virginia Tech, Blacksburg), Devesh Tiwari (Oak Ridge National Lab).....300

Lazy Release Consistency for GPUs, Johnathan Alsop (University of Illinois at Urbana-Champaign), Marc S. Orr (University of Wisconsin - Madison and AMD), Bradford M. Beckmann (AMD), David A. Wood (University of Wisconsin - Madison and AMD).....312

Improving Energy Efficiency of DRAM by Exploiting Half Page Row Access, Heonjae Ha (Stanford University), Ardavan Pedram (Stanford University and Movidius), Stephen Richardson (Stanford University), Shahar Kvatinsky (Technion-Israel Institute of Technology), Mark Horowitz (Stanford University).....325

Session 3b: Interconnect

OSCAR: Orchestrating STT-RAM Cache Traffic for Heterogeneous CPU-GPU Architectures, Jia Zhan (University of California, Santa Barbara), Onur Kayiran (Advanced Micro Devices), Gabriel H. Loh (Advanced Micro Devices), Chita R. Das (The Pennsylvania State University), Yuan Xie (University of California, Santa Barbara).....337

A Unified Memory Network Architecture for In-Memory Computing in Commodity Servers, Jia Zhan (University of California, Santa Barbara), Itir Akgun (University of California, Santa Barbara), Jishen Zhao (University of California, Santa Cruz), Al Davis (HP), Paolo Faraboschi (HP), Yuangang Wang (Huawei), Yuan Xie (University of California, Santa Barbara).....350

Contention-based Congestion Management in Large-Scale Networks, Gwangsun Kim (KAIST), Changhyun Kim (KAIST), Jiyun Jeong (KAIST), Mike Parker (Intel), John Kim (KAIST).....364

Dynamic Error Mitigation in NoCs using Intelligent Prediction Techniques, Dominic DiTomaso (Ohio University), Travis Boraten (Ohio University), Avinash Kodi (Ohio University), Ahmed Louri (George Washington University).....377

Reducing Data Movement Energy via Online Data Clustering and Encoding, Shibo Wang (University of Rochester), Engin Ipek (University of Rochester).....389

Session 4a: Multicore

Racer: TSO Consistency via Race Detection, Alberto Ros (Universidad de Murcia), Stefanos Kaxiras (Uppsala Universitet).....402

Exploiting Semantic Commutativity in Hardware Speculation, Guowei Zhang (MIT), Virginia Chiu (MIT), Daniel Sanchez (MIT).....415

CANDY: Enabling Coherent DRAM Caches for Multi-Node Systems, Chiachen Chou (Georgia Institute of Technology), Aamer Jaleel (NVIDIA), Moinuddin K. Qureshi (Georgia Institute of Technology).....427

C³D: Mitigating the NUMA Bottleneck via Coherent DRAM Caches, Cheng-Chieh Huang (University of Edinburgh), Rakesh Kumar (University of Edinburgh), Marco Elver (University of Edinburgh), Boris Grot (University of Edinburgh), Vijay Nagarajan (University of Edinburgh).....440

Session 4b: Security

Quantifying and Improving the Efficiency of Hardware-based Mobile Malware Detectors, Mikhail Kazdagli (UT Austin), Vijay Janapa Reddi (UT Austin), Mohit Tiwari (UT Austin).....452

PoisonIvy: Safe Speculation for Secure Memory, Tamara Silbergleit Lehman (Duke University), Andrew D. Hilton (Duke University), Benjamin C. Lee (Duke University).....465

ReplayConfusion: Detecting Cache-based Covert Channel Attacks Using Record and Replay, Mengjia Yan (University of Illinois at Urbana Champaign), Yasser Shalabi (University of Illinois at Urbana Champaign), Josep Torrellas (University of Illinois at Urbana Champaign).....478

Jump Over ASLR: Attacking Branch Predictors to Bypass ASLR, Dmitry Evtyushkin (Binghamton University), Dmitry Ponomarev (Binghamton University), Nael Abu-Ghazaleh (University of California, Riverside).....492

Session 5a: Approximate Computing

Concise Loads and Stores: The Case for an Asymmetric Compute-Memory Architecture for Approximation, Animesh Jain (University of Michigan), Parker Hill (University of Michigan), Shih-Chieh Lin (University of Michigan), Muneeb Khan (Uppsala University), Md E. Haque (University of Michigan), Michael A. Laurenzano (University of Michigan), Scott Mahlke (University of Michigan), Lingjia Tang (University of Michigan), Jason Mars (University of Michigan).....505

Approxilyzer: Towards A Systematic Framework for Instruction-Level Approximate Computing and its Application to Hardware Resiliency, Radha Venkatagiri (University of Illinois at Urbana Champaign), Abdulrahman Mahmoud (University of Illinois at Urbana Champaign), Siva Kumar Sastry Hari (NVIDIA), Sarita V. Adve (University of Illinois at Urbana Champaign).....518

The Bunker Cache for Spatio-Value Approximation, Joshua San Miguel (University of Toronto), Jorge Albericio (University of Toronto), Natalie Enright Jerger (University of Toronto), Aamer Jaleel (NVIDIA).....532

Session 5b: Accelerators 1

HARE: Hardware Accelerator for Regular Expressions, Vaibhav Gogte (University of Michigan), Aasheesh Kolli (University of Michigan), Michael J. Cafarella (University of Michigan), Loris D'Antoni (University of Wisconsin-Madison), Thomas F. Wenisch (University of Michigan).....544

The Microarchitecture of a Real-time Robot Motion Planning Accelerator, Sean Murray (Duke University), Will Floyd-Jones (Duke University), Ying Qi (Duke University), George Konidaris (Duke University), Daniel J. Sorin (Duke University).....556

Efficient Data Supply for Hardware Accelerators with Prefetching and Access/Execute Decoupling, Tao Chen (Cornell University), G. Edward Suh (Cornell University).....568

Session 6a: Accelerators 2

An Ultra Low-Power Hardware Accelerator for Automatic Speech Recognition, Reza Yazdani Aminabadi (Universitat Politècnica de Catalunya), Albert Segura (Universitat Politècnica de Catalunya), Jose-Maria Arnau (Universitat Politècnica de Catalunya), Antonio Gonzalez (Universitat Politècnica de Catalunya).....580

Co-Designing Accelerators and SoC Interfaces using gem5-Aladdin, Yakun Sophia Shao (NVIDIA), Sam (Likun) Xi (Harvard University), Vijayalakshmi Srinivasan (IBM), Gu-Yeon Wei (Harvard University), David Brooks (Harvard University).....592

CHAINSAW: Von-Neumann Accelerators to Leverage Fused Instruction Chains, Amirali Sharifan (Simon Fraser University), Snehasish Kumar (Simon Fraser University), Apala Guha (Simon Fraser University), Arrvindh Shriraman (Simon Fraser University).....604

Chameleon: Versatile and Practical Near-DRAM Acceleration Architecture for Large Memory Systems, Hadi Asghari-Moghaddam (University of Illinois at Urbana-Champaign), Young Hoon Son (Seoul National University), Jung Ho Ahn (Seoul National University), Nam Sung Kim (University of Illinois at Urbana-Champaign).....618

Session 6b: Mobile & Power Mgmt

A Patch Memory System For Image Processing and Computer Vision, Jason Clemons (NVIDIA), Chih-Chi Cheng (Qualcomm), Turi Frosio (NVIDIA), Daniel Johnson (NVIDIA), Steve W. Keckler (NVIDIA).....631

Evaluating Programmable Architectures for Imaging and Vision Applications, Artem Vasilyev (Stanford University), Nikhil Bhagdikar (Stanford University), Ardavan Pedram (Stanford University and Movidius), Stephen Richardson (Stanford University), Shahar Kvatinsky (Technion), Mark Horowitz (Stanford University).....644

Redefining QoS and Customizing the Power Management Policy to Satisfy Individual Mobile Users, Kaige Yan (University of Houston), Xingyao Zhang (University of Houston), Jingweijia Tan (University of Houston), Xin Fu (University of Houston).....657

Snatch: Opportunistically Reassigning Power Allocation between Processor and Memory in 3D Stacks, Dimitrios Skarlatos (University of Illinois at Urbana-Champaign), Renji Thomas (Ohio State University), Aditya Agrawal (NVIDIA), Shibin Qin (University of Illinois at Urbana-Champaign), Robert Pilawa-Podgurski (University of Illinois at Urbana-Champaign), Ulya R. Karpuzcu (University of Minnesota, Twin Cities), Radu Teodorescu (Ohio State University), Nam Sung Kim (University of Illinois at Urbana-Champaign), Josep Torrellas (University of Illinois at Urbana-Champaign).....669

Tj-states: Processor Power Management in the Temperature Inversion Region, Yazhou Zu (University of Texas at Austin), Wei Huang (AMD), Indrani Paul (AMD), Vijay Janapa Reddi (University of Texas at Austin).....681

Session 7: Best Paper Candidates

Graphicionado: A High-Performance and Energy-Efficient Accelerator for Graph Analytics, Tae Jun Ham (Princeton University), Lisa Wu (University of California, Berkeley), Narayanan Sundaram (Intel), Nadathur Satish (Intel), Margaret Martonosi (Princeton University).....694

Improving Bank-Level Parallelism for Irregular Applications, Xulong Tang (Pennsylvania State University, University Park), Mahmut Kandemir (Pennsylvania State University, University Park), Praveen Yedlapalli (VMware), Jagadish Kotra (Pennsylvania State University, University Park).....707

Delegated Persist Ordering, Aasheesh Kolli (University of Michigan), Jeff Rosen (Snowflake Computing), Stephan Diestelhorst (ARM), Ali Saidi (ARM), Steven Pelley (Snowflake Computing), Sihang Liu (University of Michigan), Peter M. Chen (University of Michigan), Thomas F. Wenisch (University of Michigan).....719

Spectral Profiling: Observer-Effect-Free Profiling by Monitoring EM Emanations, Nader Sehatbakhsh (Georgia Institute of Technology, Atlanta), Alireza Nazari (Georgia Institute of Technology, Atlanta), Alenka Zajic (Georgia Institute of Technology, Atlanta), Milos Prvulovic (Georgia Institute of Technology, Atlanta).....732

Path Confidence based Lookahead Prefetching, Jinchun Kim (Texas A&M University), Seth H. Pugsley (Intel), Paul V. Gratz (Texas A&M University), A. L. Narasimha Reddy (Texas A&M University), Chris Wilkerson (Intel), Zeshan Chishti (Intel).....743

Continuous Runahead: Transparent Hardware Acceleration for Memory Intensive Workloads, Milad Hashemi (The University of Texas at Austin), Onur Mutlu (ETH Zürich), Yale N. Patt (The University of Texas at Austin).....755