

2018 IEEE International Symposium on High Performance Computer Architecture (HPCA 2018)

**Vienna, Austria
24-28 February 2018**



**IEEE Catalog Number: CFP18013-POD
ISBN: 978-1-5386-3660-2**

**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:	CFP18013-POD
ISBN (Print-On-Demand):	978-1-5386-3660-2
ISBN (Online):	978-1-5386-3659-6
ISSN:	1530-0897

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

2018 IEEE International Symposium on High Performance Computer Architecture HPCA 2018

Table of Contents

Message from the General Chair	xiii
Message from the Program Chair	xiv
Program Committee	xvi
Organizing Committee	xviii
External Review Committee	xix
Industry Track Committee	xxii

Session 1: Best Paper Session

Amdahl's Law in the Datacenter Era: A Market for Fair Processor Allocation	1
<i>Seyed Majid Zahedi (Duke University), Qiuyun Llull (Duke University), and Benjamin C. Lee (Duke University)</i>	
iNPG: Accelerating Critical Section Access with In-network Packet Generation for NoC Based Many-Cores	15
<i>Yuan Yao (KTH Royal Institute of Technology) and Zhonghai Lu (KTH Royal Institute of Technology)</i>	
Enabling Efficient Network Service Function Chain Deployment on Heterogeneous Server Platform	27
<i>Yang Hu (The University of Texas at Dallas) and Tao Li (University of Florida)</i>	
Reducing Data Transfer Energy by Exploiting Similarity within a Data Transaction	40
<i>Donghyuk Lee (NVIDIA), Mike O'Connor (NVIDIA), and Niladrish Chatterjee (NVIDIA)</i>	

Session 2A: Architecture for Neural Network

Making Memristive Neural Network Accelerators Reliable	52
<i>Ben Feinberg (University of Rochester), Shibo Wang (University of Rochester), and Engin Ipek (University of Rochester)</i>	
Towards Efficient Microarchitectural Design for Accelerating Unsupervised GAN-Based Deep Learning	66
<i>Mingcong Song (University of Florida), Jiaqi Zhang (University of Florida), Huixiang Chen (University of Florida), and Tao Li (University of Florida)</i>	

Compressing DMA Engine: Leveraging Activation Sparsity for Training Deep Neural Networks	78
<i>Minsoo Rhu (POSTECH), Mike O'Connor (NVIDIA), Niladrish Chatterjee (NVIDIA), Jeff Pool (NVIDIA), Youngeun Kwon (POSTECH), and Stephen W. Keckler (NVIDIA)</i>	
In-Situ AI: Towards Autonomous and Incremental Deep Learning for IoT Systems	92
<i>Mingcong Song (University of Florida), Kan Zhong (Chongqing University), Jiaqi Zhang (University of Florida), Yang Hu (University of Texas at Dallas), Duo Liu (Chongqing University), Weigong Zhang (Capital Normal University), Jing Wang (Capital Normal University), and Tao Li (University of Florida)</i>	

Session 2B: Cache and Memory

KPart: A Hybrid Cache Partitioning-Sharing Technique for Commodity Multicores	104
<i>Nosayba El-Sayed (MIT), Anurag Mukkara (MIT), Po-An Tsai (MIT), Harshad Kasture (Oracle Labs), Xiaosong Ma (Qatar Computing Research Institute), and Daniel Sanchez (MIT)</i>	
SIPT: Speculatively Indexed, Physically Tagged Caches	118
<i>Tianhao Zheng (The University of Texas at Austin), Haishan Zhu (The University of Texas at Austin), and Mattan Erez (The University of Texas at Austin)</i>	
Domino Temporal Data Prefetcher	131
<i>Mohammad Bakhshalipour (Sharif University of Technology), Pejman Lotfi-Kamran (Institute for Research in Fundamental Sciences (IPM)), and Hamid Sarbazi-Azad (Sharif University of Technology)</i>	
ProFess: A Probabilistic Hybrid Main Memory Management Framework for High Performance and Fairness ..	143
<i>Dmitry Knyagin (Chalmers University of Technology), Vassilis Papaefstathiou (FORTH-ICS), and Per Stenstrom (Chalmers University of Technology)</i>	

Session 3A: Security

RCoal: Mitigating GPU Timing Attack via Subwarp-Based Randomized Coalescing Techniques	156
<i>Gurunath Kadam (College of William and Mary), Danfeng Zhang (Penn State University), and Adwait Jog (College of William and Mary)</i>	
Are Coherence Protocol States Vulnerable to Information Leakage?	168
<i>Fan Yao (George Washington University), Milos Doroslovacki (George Washington University), and Guru Venkataramani (George Washington University)</i>	
Record-Replay Architecture as a General Security Framework	180
<i>Yasser Shalabi (University of Illinois at Urbana Champaign), Mengjia Yan (University of Illinois at Urbana-Champaign), Nima Honarmand (Stony Brook University), Ruby B. Lee (Princeton University), and Josep Torrellas (University of Illinois at Urbana-Champaign)</i>	

The DRAM Latency PUF: Quickly Evaluating Physical Unclonable Functions by Exploiting the Latency-Reliability Tradeoff in Modern Commodity DRAM Devices	194
<i>Jeremie S. Kim (Carnegie Mellon University), Minesh Patel (ETH Zürich), Hasan Hassan (ETH Zürich), and Onur Mutlu (ETH Zürich)</i>	

Session 3B: GPU Cache and Memory

Accelerate GPU Concurrent Kernel Execution by Mitigating Memory Pipeline Stalls	208
<i>Hongwen Dai (North Carolina State University), Zhen Lin (North Carolina State University), Chao Li (North Carolina State University), Chen Zhao (Xi'an Jiaotong University), Fei Wang (Xi'an Jiaotong University), Nanning Zheng (Xi'an Jiaotong University), and Huiyang Zhou (North Carolina State University)</i>	
LATTE-CC: Latency Tolerance Aware Adaptive Cache Compression Management for Energy Efficient GPUs	221
<i>Akhil Arunkumar (Arizona State University), Shin-Ying Lee (Arizona State University), Vignesh Soundararajan (Arizona State University), and Carole-Jean Wu (Arizona State University)</i>	
High-Performance GPU Transactional Memory via Eager Conflict Detection	235
<i>Xiaowei Ren (The University of British Columbia) and Mieszko Lis (The University of British Columbia)</i>	
Efficient and Fair Multi-programming in GPUs via Effective Bandwidth Management	247
<i>Haonan Wang (College of William and Mary), Fan Luo (College of William and Mary), Mohamed Ibrahim (College of William and Mary), Onur Kayiran (Advanced Micro Devices), and Adwait Jog (College of William and Mary)</i>	

Session 4A: Microarchitecture and Benchmark

A Novel Register Renaming Technique for Out-of-Order Processors	259
<i>Hamid Tabani (Universitat Politècnica de Catalunya), Jose-Maria Arnau (Universitat Politècnica de Catalunya), Jordi Tubella (Universitat Politècnica de Catalunya), and Antonio Gonzalez (Universitat Politècnica de Catalunya)</i>	
Wait of a Decade: Did SPEC CPU 2017 Broaden the Performance Horizon?	271
<i>Reena Panda (The University of Texas at Austin), Shuang Song (The University of Texas at Austin), Joseph Dean (The University of Texas at Austin), and Lizy K. John (The University of Texas at Austin)</i>	
Architectural Support for Task Dependence Management with Flexible Software Scheduling	283
<i>Emilio Castillo (Barcelona Supercomputing Center), Lluc Alvarez (Barcelona Supercomputing Center), Miquel Moreto (Barcelona Supercomputing Center), Marc Casas (Barcelona Supercomputing Center), Enrique Vallejo (Universidad de Cantabria), Jose Luis Bosque (Universidad de Cantabria), Ramon Beivide (Universidad de Cantabria), and Mateo Valero (Barcelona Supercomputing Center)</i>	
GDP: Using Dataflow Properties to Accurately Estimate Interference-Free Performance at Runtime	296
<i>Magnus Jahre (Norwegian University of Science and Technology) and Lieven Eeckhout (Ghent University)</i>	

Session 4B: Persistent and NVM Memory

Crash Consistency in Encrypted Non-volatile Main Memory Systems	310
<i>Sihang Liu (University of Virginia), Aasheesh Kolli (VMware Research and Pennsylvania State University), Jinglei Ren (Microsoft Research), and Samira Khan (University of Virginia)</i>	
Adaptive Memory Fusion: Towards Transparent, Agile Integration of Persistent Memory	324
<i>Dongliang Xue (Shanghai Jiao Tong University), Chao Li (Shanghai Jiao Tong University), Linpeng Huang (Shanghai Jiao Tong University), Chentao Wu (Shanghai Jiao Tong University), and Tianyou Li (Intel Asia Pacific R&D Co.)</i>	
Steal but No Force: Efficient Hardware Undo+Redo Logging for Persistent Memory Systems	336
<i>Matheus Almeida Ogleari (University of California), Ethan L. Miller (University of California), and Jishen Zhao (University of California)</i>	
Enabling Fine-Grain Restricted Coset Coding Through Word-Level Compression for PCM	350
<i>Seyed Mohammad Seyedzadeh (University of Pittsburgh), Alex Jones (University of Pittsburgh), and Rami Melhem (University of Pittsburgh)</i>	

Session 5A: GPU

Perception-Oriented 3D Rendering Approximation for Modern Graphics Processors	362
<i>Chenhao Xie (University of Houston), Xin Fu (University of Houston), and Shuaiwen Song (Pacific Northwest National Lab)</i>	
Warp Scheduling for Fine-Grained Synchronization	375
<i>Ahmed ElTantawy (Huawei) and Tor M. Aamodt (University of British Columbia)</i>	
WIR: Warp Instruction Reuse to Minimize Repeated Computations in GPUs	389
<i>Keunsoo Kim (Yonsei University) and Won Woo Ro (Yonsei University)</i>	
G-TSC: Timestamp Based Coherence for GPUs	403
<i>Abdulaziz Tabbakh (University of Southern California), Xuehai Qian (University of Southern California), and Murali Annavaram (University of Southern California)</i>	

Session 5B: Secure Memory

D-ORAM: Path-ORAM Delegation for Low Execution Interference on Cloud Servers with Untrusted Memory	416
<i>Rujia Wang (University of Pittsburgh), Youtao Zhang (University of Pittsburgh), and Jun Yang (University of Pittsburgh)</i>	
Secure DIMM: Moving ORAM Primitives Closer to Memory	428
<i>Ali Shafiee (University of Utah), Rajeev Balasubramonian (University of Utah), Mohit Tiwari (University of Texas), and Feifei Li (University of Utah)</i>	

Comprehensive VM Protection Against Untrusted Hypervisor Through Retrofitted AMD Memory Encryption	441
<i>Yuming Wu (Institute of Parallel and Distributed Systems), Yutao Liu (Institute of Parallel and Distributed Systems), Ruifeng Liu (Institute of Parallel and Distributed Systems), Haibo Chen (Institute of Parallel and Distributed Systems), Binyu Zang (Institute of Parallel and Distributed Systems), and Haibing Guan (Institute of Parallel and Distributed Systems)</i>	
SYNERGY: Rethinking Secure-Memory Design for Error-Correcting Memories	454
<i>Gururaj Saileshwar (Georgia Institute of Technology), Prashant J. Nair (IBM Research), Prakash Ramrakhiani (ARM Research), Wendy Elsasser (ARM Research), and Moinuddin K. Qureshi (Georgia Institute of Technology)</i>	

Session 6A: Novel Architecture

A Case for Packageless Processors	466
<i>Saptadeep Pal (University of California), Daniel Petrisko (University of Illinois at Urbana-Champaign), Adeel A. Bajwa (University of California), Puneet Gupta (University of California), Subramanian S. Iyer (University of California), and Rakesh Kumar (University of Illinois at Urbana-Champaign)</i>	
Extending the Power-Efficiency and Performance of Photonic Interconnects for Heterogeneous Multicores with Machine Learning	480
<i>Scott Van Winkle (Ohio University), Avinash Karanth Kodi (Ohio University), Razvan Bunescu (Ohio University), and Ahmed Louri (George Washington University)</i>	
Routerless Network-on-Chip	492
<i>Fawaz Alazemi (Oregon State University), Arash AziziMazreah (Oregon State University), Bella Bose (Oregon State University), and Lizhong Chen (Oregon State University)</i>	
HeatWatch: Improving 3D NAND Flash Memory Device Reliability by Exploiting Self-Recovery and Temperature Awareness	504
<i>Yixin Luo (Carnegie Mellon University), Saugata Ghose (Carnegie Mellon University), Yu Cai (Seagate Technology), Erich F. Haratsch (Seagate Technology), and Onur Mutlu (ETH Zürich)</i>	

Session 6B: In-memory Computing

RC-NVM: Enabling Symmetric Row and Column Memory Accesses for In-memory Databases	518
<i>Peng Wang (Peking University), Shuo Li (National University of Defense Technology, China), Guangyu Sun (Peking University), Xiaoyang Wang (Peking University), Yiran Chen (Duke University), Hai (Helen) Li (Duke University), Jason Cong (University of California, Los Angeles), Nong Xiao (National University of Defense Technology, China), and Tao Zhang (Pennsylvania State University)</i>	
GraphR: Accelerating Graph Processing Using ReRAM	531
<i>Linghao Song (Duke University), Youwei Zhuo (University of Southern California), Xuehai Qian (University of Southern California), Hai Li (Duke University), and Yiran Chen (Duke University)</i>	

GraphP: Reducing Communication for PIM-Based Graph Processing with Efficient Data Partition	544
<i>Mingxing Zhang (Tsinghua University), Youwei Zhuo (University of Southern California), Chao Wang (University of Southern California), Mingyu Gao (Stanford University), Yongwei Wu (Tsinghua University), Kang Chen (Tsinghua University), Christos Kozyrakis (Stanford University), and Xuehai Qian (University of Southern California)</i>	
PM3: Power Modeling and Power Management for Processing-in-Memory	558
<i>Chao Zhang (Peking University), Tong Meng (Peking University), and Guangyu Sun (Peking University)</i>	

Session 7A: Industry Track

Don't Correct the Tags in a Cache, Just Check Their Hamming Distance from the Lookup Tag	571
<i>Alex Gendler (Intel), Arkady Bramnik (Intel), Ariel Szapiro (Intel), and Yiannakis Sazeides (University of Cyprus)</i>	
Reliability-Aware Data Placement for Heterogeneous Memory Architecture	583
<i>Manish Gupta (UCSD), Vilas Sridharan (AMD), David Roberts (AMD), Andreas Prodromou (UCSD), Ashish Venkat (UCSD), Dean Tullsen (UCSD), and Rajesh Gupta (UCSD)</i>	
SmarCo: An Efficient Many-Core Processor for High-Throughput Applications in Datacenters	596
<i>Dongrui Fan (Institute of Computing Technology), Wenming Li (Institute of Computing Technology), Xiaochun Ye (Institute of Computing Technology), Da Wang (Institute of Computing Technology), Hao Zhang (Institute of Computing Technology), Zhimin Tang (Institute of Computing Technology), and Ninghui Sun (Institute of Computing Technology)</i>	
Lost in Abstraction: Pitfalls of Analyzing GPUs at the Intermediate Language Level	608
<i>Anthony Gutierrez (Advanced Micro Devices), Bradford M. Beckmann (Advanced Micro Devices), Alexandru Dutu (Advanced Micro Devices), Joseph Gross (Advanced Micro Devices), Michael LeBeane (Advanced Micro Devices), John Kalamatianos (Advanced Micro Devices), Onur Kayiran (Advanced Micro Devices), Matthew Poremba (Advanced Micro Devices), Brandon Potter (Advanced Micro Devices), Sooraj Puthoor (Advanced Micro Devices), Matthew D. Sinclair (Advanced Micro Devices), Mark Wyse (Advanced Micro Devices), Jieming Yin (Advanced Micro Devices), Xianwei Zhang (Advanced Micro Devices), Akshay Jain (Purdue University), and Timothy Rogers (Purdue University)</i>	

Session 7B: Best of CAL

Session 8A: Industry Track (Applications)

Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective	620
<i>Kim Hazelwood (Facebook), Sarah Bird (Facebook), David Brooks (Facebook), Soumith Chintala (Facebook), Utku Diril (Facebook), Dmytro Dzhulgakov (Facebook), Mohamed Fawzy (Facebook), Bill Jia (Facebook), Yangqing Jia (Facebook), Aditya Kalro (Facebook), James Law (Facebook), Kevin Lee (Facebook), Jason Lu (Facebook), Pieter Noordhuis (Facebook), Misha Smelyanskiy (Facebook), Liang Xiong (Facebook), and Xiaodong Wang (Facebook)</i>	
Amdahl's Law in Big Data Analytics: Alive and Kicking in TPCx-BB (BigBench)	630
<i>Daniel Richins (The University of Texas at Austin), Tahrina Ahmed (Stanford University), Russell Clapp (Intel), and Vijay Janapa Reddi (Google)</i>	
Memory Hierarchy for Web Search	643
<i>Grant Ayers (Stanford University), Jung Ho Ahn (Seoul National University), Christos Kozyrakis (Stanford University), and Parthasarathy Ranganathan (Google)</i>	
Characterizing Resource Sensitivity of Database Workloads	657
<i>Rathijit Sen (Microsoft Corporation) and Karthik Ramachandra (Microsoft Corporation)</i>	

Session 8B: Memory

ERUCA: Efficient DRAM Resource Utilization and Resource Conflict Avoidance for Memory System Parallelism	670
<i>Sangkug Lym (The University of Texas at Austin), Heonjae Ha (Stanford University), Yongkee Kwon (The University of Texas at Austin), Chun-kai Chang (The University of Texas at Austin), Jungrae Kim (Microsoft), and Matta Erez (The University of Texas at Austin)</i>	
DUO: Exposing On-Chip Redundancy to Rank-Level ECC for High Reliability	683
<i>Seong-Lyong Gong (UT Austin), Jungrae Kim (Microsoft), Sangkug Lym (UT Austin), Michael Sullivan (NVIDIA), Howard David (Huawei), and Mattan Erez (UT Austin)</i>	
Memory System Design for Ultra Low Power, Computationally Error Resilient Processor Microarchitectures	696
<i>Sriseshan Srikanth (Georgia Institute of Technology), Paul G. Rabbat (Intel Corporation), Eric R. Hein (Georgia Institute of Technology), Bobin Deng (Georgia Institute of Technology), Thomas M. Conte (Georgia Institute of Technology), Erik DeBenedictis (Sandia National Laboratories), Jeanine Cook (Sandia National Laboratories), and Michael P. Frank (Sandia National Laboratories)</i>	
NACHOS: Software-Driven Hardware-Assisted Memory Disambiguation for Accelerators	710
<i>Naveen Vedula (Simon Fraser University), Arrvindh Shriraman (Simon Fraser University), Snehasish Kumar (Simon Fraser University), and William N Sumner (Simon Fraser University)</i>	

Session 9A: Accelerators

OuterSPACE: An Outer Product Based Sparse Matrix Multiplication Accelerator	724
<i>Subhankar Pal (University of Michigan), Jonathan Beaumont (University of Michigan), Dong-Hyeon Park (University of Michigan), Aporva Amarnath (University of Michigan), Siying Feng (University of Michigan), Chaitali Chakrabarti (Arizona State University), Hun-Seok Kim (University of Michigan), David Blaauw (University of Michigan), Trevor Mudge (University of Michigan), and Ronald Dreslinski (University of Michigan)</i>	
Searching for Potential gRNA Off-Target Sites for CRISPR/Cas9 Using Automata Processing Across Different Platforms	737
<i>Chunkun Bo (University of Virginia), Vinh Dang (University of Virginia), Elaheh Sadredini (University of Virginia), and Kevin Skadron (University of Virginia)</i>	
Characterizing and Mitigating Output Reporting Bottlenecks in Spatial Automata Processing Architectures	749
<i>Jack Wadden (University of Virginia), Kevin Angstadt (University of Michigan), and Kevin Skadron (University of Virginia)</i>	

Session 9B: Power

Power and Energy Characterization of an Open Source 25-Core Manycore Processor	762
<i>Michael McKeown (Princeton University), Alexey Lavrov (Princeton University), Mohammad Shahradd (Princeton University), Paul J. Jackson (Princeton University), Yaosheng Fu (Princeton University), Jonathan Balkind (Princeton University), Tri M. Nguyen (Princeton University), Katie Lim (Princeton University), Yanqi Zhou (Princeton University), and David Wentzlaff (Princeton University)</i>	
A Spot Capacity Market to Increase Power Infrastructure Utilization in Multi-tenant Data Centers	776
<i>Mohammad Islam (University of California), Xiaoqi Ren (California Institute of Technology), Shaolei Ren (University of California), and Adam Wierman (California Institute of Technology)</i>	
GPGPU Power Modeling for Multi-domain Voltage-Frequency Scaling	789
<i>João Guerreiro (INESC-ID), Aleksandar Ilic (INESC-ID), Nuno Roma (INESC-ID), and Pedro Tomás (INESC-ID)</i>	
Author Index	801