2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2018)

Fukuoka, Japan 20-24 October 2018

Pages 1-493



IEEE Catalog Number: ISBN:

978-1-5386-6241-0

CFP18071-POD

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:
 CFP18071-POD

 ISBN (Print-On-Demand):
 978-1-5386-6241-0

 ISBN (Online):
 978-1-5386-6240-3

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



2018 51st Annual IEEE/ACM International Symposium on Microarchitecture MICRO 2018

Table of Contents

Message from the General Chairs xvi Message from the Program Chairs xvii Organization Committee xx Program Committee xxi Reviewers xxiv		
1A: Accelerators		
Exploiting Locality in Graph Analytics through Hardware-Accelerated Traversal Scheduling .1		
Cambricon-S: Addressing Irregularity in Sparse Neural Networks through A Cooperative Software/Hardware Approach .1.5		
CSE: Parallel Finite State Machines with Convergence Set Enumeration .29		
Inter-Thread Communication in Multithreaded, Reconfigurable Coarse-Grain Arrays .42		

An Architectural Framework for Accelerating Dynamic Parallel Algorithms on Reconfigurable Hardware 55.		
Tao Chen (Cornell University), Shreesha Srinath (Cornell University), Christopher Batten (Cornell University), and G. Edward Suh (Cornell University)		
1B: Microarchitecture		
Composable Building Blocks to Open up Processor Design .68. Sizhuo Zhang (Massachusetts Institute of Technology), Andrew Wright (Massachusetts Institute of Technology), Thomas Bourgeat (Massachusetts Institute of Technology), and Arvind Arvind (Massachusetts Institute of Technology)		
Performance Improvement by Prioritizing the Issue of the Instructions in Unconfident Branch Slices .82		
The Superfluous Load Queue .95		
Architectural Support for Probabilistic Branches .1.08. Almutaz Adileh (Ghent University), David J. Lilja (University of Minnesota), and Lieven Eeckhout (Ghent University)		
STRAIGHT: Hazardless Processor Architecture Without Register Renaming .1.21		
2 A: ML Accelerators		
Diffy: a Déjà vu-Free Differential Deep Neural Network Accelerator 1.34		
Beyond the Memory Wall: A Case for Memory-Centric HPC System for Deep Learning .1.48		
Towards Memory Friendly Long-Short Term Memory Networks (LSTMs) on Mobile GPUs .1.62		

A Network-Centric Hardware/Algorithm Co-Design to Accelerate Distributed Training of Deep Neural Networks .1.75.... Youjie Li (University of Illinois at Urbana-Champaign, United States), Jongse Park (Georgia Institute of Technology, United States), Mohammad Alian (University of Illinois at Urbana-Champaign, United States), Yifan Yuan (University of Illinois at Urbana-Champaign, United States), Zheng Qu (Tsinghua University, China), Peitian Pan (Shanghai Jiao Tong University, China), Ren Wang (Intel Corporation, United States), Alexander Schwing (University of Illinois at Urbana-Champaign, United States), Hadi Esmaeilzadeh (University of California San Diego, United States), and Nam Sung Kim (University of Illinois at Urbana-Champaign, United States) PermDNN: Efficient Compressed DNN Architecture with Permuted Diagonal Matrices .189..... Chunhua Deng (City University of New York), Siyu Liao (City University of New York), Yi Xie (City University of New York), Keshab K. Parhi (University of Minnesota, Twin Cities), Xuehai Qian (University of Southern California), and Bo Yuan (City University of New York) 2B: Compilers and Programming Languages Rethinking the Memory Hierarchy for Modern Languages .203.... Po-An Tsai (Massachusetts İnstitute of Technology), Yee Ling Gan (Massachusetts Institute of Technology), and Daniel Sanchez (Massachusetts Institute of Technology) Harmonizing Speculative and Non-Speculative Execution in Architectures for Ordered Parallelism 217 Mark C. Jeffrey (Massachusetts Institute of Technology), Victor A. Ying (Massachusetts Institute of Technology), Suvinay Subramanian (Massachusetts Institute of Technology), Hyun Ryong Lee (Massachusetts Institute of Technology), Joel Emer (NVIDIA/MIT), and Daniel Sanchez (Massachusetts Institute of Technology) Sampler: PMU-Based Sampling to Detect Memory Errors Latent in Production Software .231..... Sam Silvestro (University of Texas at San Antonio), Hongyu Liu (University of Texas at San Antonio), Tong Zhang (Virginia Tech), Changhee Jung (Virginia Tech), Dongyoon Lee (Virginia Tech), and Tongping Liu (University of Texas at San Antonio) TAPAS: Generating Parallel Accelerators from Parallel Programs .245..... Steven Margerm (Simon Fraser University), Amirali Sharifian (Simon Fraser University), Apala Guha (Simon Fraser University), Arrvindh Shriraman (Simon Fraser University), and Gilles Pokam (Intel Corporation) iDO: Compiler-Directed Failure Atomicity for Nonvolatile Memory .258....... Qingrui Liu (Virginia Tech), Joseph Izraelevitz (University of Rochester), Se Kwon Lee (UNIST), Michael L. Scott (University of Rochester), Sam H. Noh (UNIST), and Changhee Jung (Virginia Tech)

3A: Memory Systems - I

Scalable Distributed Last-Level TLBs Using Low-Latency Interconnects .271		
Duplicon Cache: Mitigating Off-Chip Memory Bank and Bank Group Conflicts Via Data Duplication .285.		
Ben Lin (University of Texas at Austin), Michael B. Healy (IBM T.J. Watson Research Center), Rustam Miftakhutdinov (Intel Corporation), Philip G. Emma (IBM T.J. Watson Research Center), and Yale Patt (University of Texas at Austin)		
Reducing DRAM Latency via Charge-Level-Aware Look-Ahead Partial Restoration .298		
CABLE: A CAche-Based Link Encoder for Bandwidth-Starved Manycores .312 Tri Nguyen (Princeton University, USA), Adi Fuchs (Princeton University, USA), and David Wentzlaff (Princeton University, USA)		
Attaché: Towards Ideal Memory Compression by Mitigating Metadata Bandwidth Overheads .326 Seokin Hong (IBM Thomas J. Watson Research Center), Prashant Jayaprakash Nair (IBM Thomas J. Watson Research Center), Bulent Abali (IBM Thomas J. Watson Research Center), Alper Buyuktosunoglu (IBM Thomas J. Watson Research Center), Kyu-Hyoun Kim (IBM Thomas J. Watson Research Center), and Michael Healy (IBM Thomas J. Watson Research Center)		
3B: GPGPU/GPU		
Combining HW/SW Mechanisms to Improve NUMA Performance of Multi-GPU Systems .339		
Neighborhood-Aware Address Translation for Irregular GPU Applications .3.52		
FineReg: Fine-Grained Register File Management for Augmenting GPU Throughput .36.4		

In-Register Parameter Caching for Dynamic Neural Nets with Virtual Persistent Processor Specialization 377 Farzad Khorasani (Tesla, Inc.), Hodjat Asghari Esfeden (University of California Riverside), Nael Abu-Ghazaleh (University of California Riverside), and Vivek Sarkar (Georgia Institute of Technology) Voltage-Stacked GPUs: A Control Theory Driven Cross-Layer Solution for Practical Voltage Stacking in GPUs .390. An Zou (Washington University in St. Louis), Jingwen Leng (Shanghai Jiao Tong University), Xin He (Washington University in St. Louis), Yazhou Zu (University of Texas at Austin), Christopher D. Gill (Washington University in St. Louis), Vijay Janapa Reddi (Harvard University. The University of Texas at Austin), and Xuan Zhang (Washington University in St. Louis) 4A: Security - I Osiris: A Low-Cost Mechanism to Enable Restoration of Secure Non-Volatile Memories .403...... Mao Ye (University of Central Florida), Clayton Hughes (Sandia National Laboratories), and Amro Awad (University of Central Florida) Morphable Counters: Enabling Compact Integrity Trees For Low-Overhead Secure Memories .416... Gururaj Saileshwar (Georgia Institute of Technology, USA), Prashant Nair (IBM Research, USA), Prakash Ramrakhyani (Arm Research, USA), Wendy Elsasser (Arm Research, USA), Jose Joao (Arm Research, USA), and Moinuddin Qureshi (Georgia Institute of Technology, USA) InvisiSpec: Making Speculative Execution Invisible in the Cache Hierarchy .428...... Mengjia Yan (University of Illinois at Urbana-Champaign, USA), Jiho Choi (University of Illinois at Urbana-Champaign, USA), Dimitrios Skarlatos (University of Illinois at Urbana-Champaign, USA), Adam Morrison (Tel Aviv University, Israel), Christopher Fletcher (University of Illinois at Urbana-Champaign, USA), and Josep Torrellas (University of Illinois at Urbana-Champaign, USA) Improving the Performance and Endurance of Encrypted Non-Volatile Main Memory through Deduplicating Writes .442..... Pengfei Zuo (Huazhong University of Science and Technology, China), Yu Hua (Huazhong University of Science and Technology, China), Ming Zhao

4B: Storage Systems & Technologies

of Science and Technology, China)

(Arizona State University, USA), Wen Zhou (Huazhong University of

Science and Technology, China), and Yuncheng Guo (Huazhong University

Amber*: Enabling Precise Full-System Simulation with Detailed Modeling of All SSD Resources 469 Donghyun Gouk (Yonsei University, Republic of Korea), Miryeong Kwon (Yonsei University, Republic of Korea), Jie Zhang (Yonsei University, Republic of Korea), Sungioon Koh (Yonsei University, Republic of Korea), Wonil Choi (Yonsei University and Pennsylvania State University), Nam Sung Kim (University of Illinois Urbana-Champaign, USA), Mahmut Kandemir (Pennsylvania State University, USA), and Myoungsoo Jung (Yonsei University, Republic of Korea) Invalid Data-Aware Coding to Enhance the Read Performance of High-Density Flash Memories .482 Wonil Choi (Pennsylvania State University, United States), Myoungsoo Jung (Yonsei University, Korea), and Mahmut Kandemir (Pennsylvania State University, United States) Persistence Parallelism Optimization: A Holistic Approach from Memory Bus to RDMA Network .4.94. Xing Hu (University of California, Santa Barbara, United States), Matheus Ogleari (University of California, Santa Cruz, United States), Jishen Zhao (University of California, San Diego, United States), Shuangchen Li (University of California, Santa Barbara, United States), Abanti Basak (University of California, Santa Barbara, United States), and Yuan Xie (University of California, Santa Barbara, United States) 5A: Memory Systems - II PiCL: A Software-Transparent, Persistent Cache Log for Nonvolatile Main Memory .507...... Tri Nguyen (Princeton University, USA) and David Wentzlaff (Princeton University, USA) Efficient Hardware-Assisted Logging with Asynchronous and Direct-Update for Persistent Memory 520 Jungi Jeong (KAIST, South Korea), Chang Hyun Park (KAIST, South Korea), Jaehyuk Huh (KAIST, South Korea), and Seungryoul Maeng (KAIST, South Korea) CHAMELEON: A Dynamically Reconfigurable Heterogeneous Memory System .533..... Jagadish B. Kotra (AMD Research), Haibo Zhang (The Pennsylvania State University), Alaa R. Alameldeen (Intel Labs), Chris Wilkerson (NVIDIA Corporation Ltd), and Mahmut T. Kandemir (The Pennsylvania State University) Compresso: Pragmatic Main Memory Compression .546. Esha Choukse (University of Texas at Austin), Mattan Erez (University of Texas at Austin), and Alaa R. Alameldeen (Intel Labs) Farewell My Shared LLC! A Case for Private Die-Stacked DRAM Caches for Servers .559..... Amna Shahab (University of Edinburgh), Mingcan Zhu (University of Edinburgh), Artemiy Margaritov (University of Edinburgh), and Boris Grot (University of Edinburgh)

5B: Measurement, Modeling, and Simulation

Leveraging CPU Electromagnetic Emanations for Voltage Noise Characterization .573
RpStacks-MT: A High-Throughput Design Evaluation Methodology for Multi-Core Processors .586 Hanhwi Jang (POSTECH, South Korea), Jae-Eon Jo (POSTECH, South Korea), Jaewon Lee (Seoul National University, South Korea), and Jangwoo Kim (Seoul National University, South Korea)
The EH Model: Early Design Space Exploration of Intermittent Processor Architectures .6.00 Joshua San Miguel (University of Wisconsin - Madison), Karthik Ganesan (University of Toronto), Mario Badr (University of Toronto), Chunqiu Xia (University of Toronto), Rose Li (University of Toronto), Hsuan Hsiao (University of Toronto), and Natalie Enright Jerger (University of Toronto)
CounterMiner: Mining Big Performance Data from Hardware Counters .613
Taming the Killer Microsecond .627
6A: Near Memory Computing
Adaptive Scheduling for Systems with Asymmetric Memory Hierarchies .64.1
Processing-in-Memory for Energy-Efficient Neural Network Training: A Heterogeneous Approach 655
LerGAN: A Zero-Free, Low Data Movement and PIM-Based GAN Architecture .669

6B: Near Memory Computing

	ploring and Optimizing Chipkill-Correct for Persistent Memory Based on High-Density /RAMs .710
	Da Zhang (Virginia Tech), Vilas Sridharan (Advanced Micro Devices (AMD)), and Xun Jian (Virginia Tech)
Co	omprehensive Evaluation of Supply Voltage Underscaling in FPGA on-Chip Memories .724 Behzad Salami (Barcelona Supercomputing Center (BSC)), Osman S. Unsal (Barcelona Supercomputing Center (BSC)), and Adrian Cristal Kestelman (Barcelona Supercomputing Center (BSC))
Erı	ror Correlation Prediction in Lockstep Processors for Safety-Critical Systems .7:3.7 Emre Ozer (Arm Ltd.), Balaji Venu (Arm Ltd.), Xabier Iturbe (Arm Ltd.), Shidhartha Das (Arm Ltd.), Spyros Lyberis (Arm Ltd.), John Biggs (Arm Ltd.), Peter Harrod (Arm Ltd.), and John Penton (Arm Ltd.)
Fa	ult Site Pruning for Practical Reliability Analysis of GPGPU Applications .749. Bin Nie (College of William and Mary, USA), Lishan Yang (College of William and Mary, USA), Adwait Jog (College of William and Mary, USA), and Evgenia Smirni (College of William and Mary, USA)
Sw	vapCodes: Error Codes for Hardware-Software Cooperative GPU Pipeline Error Detection .7.62 Michael B. Sullivan (NVIDIA), Siva Kumar Sastry Hari (NVIDIA), Brian Zimmer (NVIDIA), Timothy Tsai (NVIDIA), and Stephen W. Keckler (NVIDIA)

7: Best Papers

CEASER: Mitigating Conflict-Based Cache Attacks via Encrypted-Address and Remapping .7.75......

Moinuddin K. Qureshi (Georgia Institute of Technology)

Application-Transparent Near-Memory Processing Architecture with Memory Channel Network .802. Mohammad Alian (University of Illinois Urbana-Champaign), Seung Won Min (University of Illinois Urbana-Champaign), Hadi Asgharimoghaddam (University of Illinois Urbana-Champaign), Ashutosh Dhar (University of Illinois Urbana-Champaign), Dong Kai Wang (University of Illinois Urbana-Champaign), Thomas Roewer (IBM Research), Adam McPadden (IBM Systems), Oliver O'Halloran (IBM Systems), Deming Chen (University of Illinois Urbana-Champaign), Jinjun Xiong (IBM Research), Daehoon Kim (DGIST), Wen-mei Hwu (University of Illinois Urbana-Champaign), and Nam Sung Kim (University of Illinois Urbana-Champaign)

End-to-End Automated Exploit Generation for Validating the Security of Processor Designs .815......

Rui Zhang (University of North Carolina at Chapel Hill), Calvin

Deutschbein (University of North Carolina at Chapel Hill), Peng Huang

(Johns Hopkins University), and Cynthia Sturton (University of North

Carolina at Chapel Hill)

8A: Non-Conventional Architectures

8B: Mobile and Embedded Architectures

EMPROF: Memory Profiling Via EM-Emanation in IoT and Hand-Held Devices .881..... Moumita Dey (Georgia Institute of Technology, USA), Alireza Nazari (Georgia Institute of Technology, USA), Alenka Zajic (Georgia Institute of Technology, USA), and Milos Prvulovic (Georgia Institute of Technology, USA) MAVBench: Micro Aerial Vehicle Benchmarking .894..... Behzad Boroujerdian (University of Texas, Austin), Hasan Genc (University of California, Berkeley), Srivatsan Krishnan (Harvard University), Wenzhi Cui (University of Texas, Austin), Aleksandra Faust (Google Brain), and Vijay Reddi (Harvard University) 9A: Domain-Specific Architectures Architectural Support for Efficient Large-Scale Automata Processing .908..... Hongyuan Liu (College of William & Mary), Mohamed Ibrahim (College of William & Mary), Onur Kayiran (Advanced Micro Devices, Inc.), Sreepathi Pai (University of Rochester), and Adwait Jog (College of William & Mary) ASPEN: A Scalable In-SRAM Architecture for Pushdown Automata .921..... Kevin Angstadt (University of Michigan, USA), Arun Subramaniyan (University of Michigan, USA), Elaheh Sadredini (University of Virginia, USA), Reza Rahimi (University of Virginia, USA), Kevin Skadron (University of Virginia, USA), Westley Weimer (University of Michigan, USA), and Reetuparna Das (University of Michigan) Morph: Flexible Acceleration for 3D CNN-Based Video Understanding .933..... Kartik Hegde (University of Illinois at Urbana-Champaign), Rohit Agrawal (University of Illinois at Urbana-Champaign), Yulun Yao (University of Illinois at Urbana-Champaign), and Christopher W Fletcher (University of Illinois at Urbana-Champaign) 9B: Security-II CheckMate: Automated Synthesis of Hardware Exploits and Security Litmus Tests .9.47..... Caroline Trippel (Princeton University), Daniel Lustig (NVIDIA), and Margaret Martonosi (Princeton University) Shadow Block: Accelerating ORAM Accesses with Data Duplication .961...... Xian Zhang (Peking University), Guangyu Sun (Peking University), Peichen Xie (Peking University), Chao Zhang (Peking University), Yannan Liu (The Chinese University of Hong Kong), Lingxiao Wei (The Chinese University of Hong Kong), Qiang Xu (The Chinese University of Hong Kong), and Chun Jason Xue (City University of Hong Kong) DAWG: A Defense Against Cache Timing Attacks in Speculative Execution Processors .974...... Vladimir Kiriansky (Massachusetts Institute of Technology, USA), Ilia Lebedev (Massachusetts Institute of Technology, USA), Saman Amarasinghe (Massachusetts Institute of Technology, USA), Srinivas Devadas (Massachusetts Institute of Technology, USA), and Joel Emer (NVIDIA / Massachusetts Institute of Technology, USA)

Author Index aga