

2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA 2017)

**Toronto, Ontario, Canada
24-28 June 2017**



**IEEE Catalog Number: CFP17030-POD
ISBN: 978-1-5090-5901-0**

**Copyright © 2017, Association for Computing Machinery (ACM)
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:	CFP17030-POD
ISBN (Print-On-Demand):	978-1-5090-5901-0
ISBN (Online):	978-1-4503-4892-8
ISSN:	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

CURRAN ASSOCIATES INC.
proceedings
.com

TABLE OF CONTENTS

IN-DATACENTER PERFORMANCE ANALYSIS OF A TENSOR PROCESSING UNIT	1
<i>Norman P. Jouppi ; Cliff Young ; Nishant Patil ; David Patterson ; Gaurav Agrawal ; Raminder Bajwa ; Sarah Bates ; Suresh Bhatia ; Nan Boden ; Al Borchers ; Rick Boyle ; Pierre-luc Cantin ; Clifford Chao ; Chris Clark ; Jeremy Coriell ; Mike Daley ; Matt Dau ; Jeffrey Dean ; Ben Gelb ; Tara Vazir Ghaemmaghami ; Rajendra Gottipati ; William Gulland ; Robert Hagmann ; C. Richard Ho ; Doug Hogberg ; John Hu ; Robert Hundt ; Dan Hurt ; Julian Ibarz ; Aaron Jaffey ; Alek Jaworski ; Alexander Kaplan ; Harshit Khaitan ; Daniel Killebrew ; Andy Koch ; Naveen Kumar ; Steve Lacy ; James Laudon ; James Law ; Diemthu Le ; Chris Leary ; Zhuyuan Liu ; Kyle Lucke ; Alan Lundin ; Gordon MacKean ; Adriana Maggiore ; Maire Mahony ; Kieran Miller ; Rahul Nagarajan ; Ravi Narayanaswami ; Ray Ni ; Kathy Nix ; Thomas Norrie ; Mark Omernick ; Narayana Penukonda ; Andy Phelps ; Jonathan Ross ; Matt Ross ; Amir Salek ; Emad Samadiani ; Chris Severn ; Gregory Sizikov ; Matthew Snelham ; Jed Souter ; Dan Steinberg ; Andy Swing ; Mercedes Tan ; Gregory Thorson ; Bo Tian ; Horia Toma ; Erick Tuttle ; Vijay Vasudevan ; Richard Walter ; Walter Wang ; Eric Wilcox ; Doe Hyun Yoon</i>	
SCALEDEEP: A SCALABLE COMPUTE ARCHITECTURE FOR LEARNING AND EVALUATING DEEP NETWORKS	13
<i>Swagath Venkataramani ; Ashish Ranjan ; Subarno Banerjee ; Dipankar Das ; Sasikanth Avancha ; Ashok Jagannathan ; Ajaya Durg ; Dheemanth Nagaraj ; Bharat Kaul ; Pradeep Dubey ; Anand Raghunathan</i>	
SCNN: AN ACCELERATOR FOR COMPRESSED-SPARSE CONVOLUTIONAL NEURAL NETWORKS	27
<i>Angshuman Parashar ; Minsoo Rhu ; Anurag Mukkara ; Antonio Puglielli ; Rangharajan Venkatesan ; Brucek Khailany ; Joel Emer ; Stephen W. Keckler ; William J. Dally</i>	
BESPOKE PROCESSORS FOR APPLICATIONS WITH ULTRA-LOW AREA AND POWER CONSTRAINTS	41
<i>Hari Cherupalli ; Henry Duwe ; Weidong Ye ; Rakesh Kumar ; John Sartori</i>	
A PROGRAMMABLE GALOIS FIELD PROCESSOR FOR THE INTERNET OF THINGS	55
<i>Yajing Chen ; Shengshuo Lu ; Cheng Fu ; David Blaauw ; Ronald Dreslinski ; Trevor Mudge ; Hun-Seok Kim</i>	
XPRO: A CROSS-END PROCESSING ARCHITECTURE FOR DATA ANALYTICS IN WEARABLES	69
<i>Aosen Wang ; Lizhong Chen ; Wenyao Xu</i>	
REGAINING LOST CYCLES WITH HOTCALLS: A FAST INTERFACE FOR SGX SECURE ENCLAVES	81
<i>Ofir Weisse ; Valeria Bertacco ; Todd Austin</i>	
INVISIMEM: SMART MEMORY DEFENSES FOR MEMORY BUS SIDE CHANNEL	94
<i>Shaizeen Aga ; Satish Narayanasamy</i>	
OBFUSMEM: A LOW-OVERHEAD ACCESS OBFUSCATION FOR TRUSTED MEMORIES	107
<i>Amro Awad ; Yipeng Wang ; Deborah Shands ; Yan Solihin</i>	
THERMOGATER: THERMALLY-AWARE ON-CHIP VOLTAGE REGULATION	120
<i>S. Karen Khatamifard ; Longfei Wang ; Weize Yu ; Selçuk Köse ; Ulya R. Karpuzcu</i>	
POWERCHIEF: INTELLIGENT POWER ALLOCATION FOR MULTI-STAGE APPLICATIONS TO IMPROVE RESPONSIVENESS ON POWER CONSTRAINED CMP	133
<i>Hailong Yang ; Quan Chen ; Moeiz Riaz ; Zhongzhi Luan ; Lingjia Tang ; Jason Mars</i>	
CHARSTAR: CLOCK HIERARCHY AWARE RESOURCE SCALING IN TILED ARCHITECTURES	147
<i>Gokul Subramanian Ravi ; Mikko H. Lipasti</i>	
CHASING AWAY RATS: SEMANTICS AND EVALUATION FOR RELAXED ATOMICS ON HETEROGENEOUS SYSTEMS	161
<i>Matthew D. Sinclair ; Johnathan Alsop ; Sarita V. Adve</i>	
HIDING THE LONG LATENCY OF PERSIST BARRIERS USING SPECULATIVE EXECUTION	175
<i>Seunghee Shin ; James Tuck ; Yan Solihin</i>	
NON-SPECULATIVE LOAD-LOAD REORDERING IN TSO	187
<i>Alberto Ros ; Trevor E. Carlson ; Mehdi Alipour ; Stefanos Kaxiras</i>	
MTRACECHECK: VALIDATING NON-DETERMINISTIC BEHAVIOR OF MEMORY CONSISTENCY MODELS IN POST-SILICON VALIDATION	201
<i>Doowon Lee ; Valeria Bertacco</i>	
REDUNDANT MEMORY ARRAY ARCHITECTURE FOR EFFICIENT SELECTIVE PROTECTION	214
<i>Ruohuang Zheng ; Michael C. Huang</i>	
CLANK: ARCHITECTURAL SUPPORT FOR INTERMITTENT COMPUTATION	228
<i>Matthew Hicks</i>	

MERLIN: EXPLOITING DYNAMIC INSTRUCTION BEHAVIOR FOR FAST AND ACCURATE MICROARCHITECTURE LEVEL RELIABILITY ASSESSMENT	241
<i>Manolis Kalliorakis ; Dimitris Gizopoulos ; Ramon Canal ; Antonio Gonzalez</i>	
THE REACH PROFILER (REAPER): ENABLING THE MITIGATION OF DRAM RETENTION FAILURES VIA PROFILING AT AGGRESSIVE CONDITIONS.....	255
<i>Minesh Patel ; Jeremie S. Kim ; Onur Mutlu</i>	
QUALITY OF SERVICE SUPPORT FOR FINE-GRAINED SHARING ON GPUS	269
<i>Zhenning Wang ; Jun Yang ; Rami Melhem ; Bruce Childers ; Youtao Zhang ; Minyi Guo</i>	
ACCELERATING GPU HARDWARE TRANSACTIONAL MEMORY WITH SNAPSHOT ISOLATION	282
<i>Sui Chen ; Lu Peng ; Samuel Irving</i>	
DECOUPLED AFFINE COMPUTATION FOR SIMT GPUS	295
<i>Kai Wang ; Calvin Lin</i>	
ACCESS PATTERN-AWARE CACHE MANAGEMENT FOR IMPROVING DATA UTILIZATION IN GPU	307
<i>Gunjae Koo ; Yunho Oh ; Won Woo Ro ; Murali Annavaram</i>	
MCM-GPU: MULTI-CHIP-MODULE GPUS FOR CONTINUED PERFORMANCE SCALABILITY.....	320
<i>Akhil Arunkumar ; Evgeny Bolotin ; Benjamin Cho ; Ugljesa Milic ; Eiman Ebrahimi ; Oreste Villa ; Aamer Jaleel ; Carole-Jean Wu ; David Nellans</i>	
EDDIE: EM-BASED DETECTION OF DEVIATIONS IN PROGRAM EXECUTION.....	333
<i>Alireza Nazari ; Nader Sehatbakhsh ; Monjur Alam ; Alenka Zajic ; Milos Prvulovic</i>	
SECURE HIERARCHY-AWARE CACHE REPLACEMENT POLICY (SHARP): DEFENDING AGAINST CACHE-BASED SIDE CHANNEL ATTACKS	347
<i>Mengjia Yan ; Bhargava Gopireddy ; Thomas Shull ; Josep Torrellas</i>	
LEMONADE FROM LEMONS: HARNESSING DEVICE WEAROUT TO CREATE LIMITED-USE SECURITY ARCHITECTURES	361
<i>Zhaoxia Deng ; Ariel Feldman ; Stuart A. Kurtz ; Frederic T. Chong</i>	
LOGCA: A HIGH-LEVEL PERFORMANCE MODEL FOR HARDWARE ACCELERATORS.....	375
<i>Muhammad Shoaib Bin Altaf ; David A. Wood</i>	
PLASTICINE: A RECONFIGURABLE ARCHITECTURE FOR PARALLEL PATTERNS.....	389
<i>Raghu Prabhakar ; Yaqi Zhang ; David Koeplinger ; Matt Feldman ; Tian Zhao ; Stefan Hadjis ; Ardavan Pedram ; Christos Kozyrakis ; Kunle Olukotun</i>	
A PROGRAMMABLE HARDWARE ACCELERATOR FOR SIMULATING DYNAMICAL SYSTEMS	403
<i>Jaeha Kung ; Yun Long ; Duckhwan Kim ; Saibal Mukhopadhyay</i>	
STREAM-DATAFLOW ACCELERATION	416
<i>Tony Nowatzki ; Vinay Gangadhar ; Newsha Ardalani ; Karthikeyan Sankaralingam</i>	
HARDWARE TRANSLATION COHERENCE FOR VIRTUALIZED SYSTEMS.....	430
<i>Zi Yan ; Ján Veselý ; Guilherme Cox ; Abhishek Bhattacharjee</i>	
HYBRID TLB COALESCING: IMPROVING TLB TRANSLATION COVERAGE UNDER DIVERSE FRAGMENTED MEMORY ALLOCATIONS.....	444
<i>Chang Hyun Park ; Taekyung Heo ; Jungi Jeong ; Jaehyuk Huh</i>	
DO-IT-YOURSELF VIRTUAL MEMORY TRANSLATION.....	457
<i>Hanna Alam ; Tianhao Zhang ; Mattan Erez ; Yoav Etsion</i>	
RETHINKING TLB DESIGNS IN VIRTUALIZED ENVIRONMENTS: A VERY LARGE PART-OF-MEMORY TLB	469
<i>Jee Ho Ryoo ; Nagendra Gulur ; Shuang Song ; Lizy K. John</i>	
LANGUAGE-LEVEL PERSISTENCY.....	481
<i>Aasheesh Kolli ; Vaibhav Gogte ; Ali Saidi ; Stephan Diestelhorst ; Peter M. Chen ; Satish Narayanasamy ; Thomas F. Wenisch</i>	
SHORTCUT: ARCHITECTURAL SUPPORT FOR FAST OBJECT ACCESS IN SCRIPTING LANGUAGES.....	494
<i>Jiho Choi ; Thomas Shull ; Maria J. Garzaran ; Josep Torrellas</i>	
ARCHITECTURAL SUPPORT FOR SERVER-SIDE PHP PROCESSING	507
<i>Dibakar Gope ; David J. Schlais ; Mikko H. Lipasti</i>	
HETEROOS — OS DESIGN FOR HETEROGENEOUS MEMORY MANAGEMENT IN DATACENTER	521
<i>Sudarsun Kannan ; Ada Gavrilovska ; Vishal Gupta ; Karsten Schwan</i>	
MAXIMIZING CNN ACCELERATOR EFFICIENCY THROUGH RESOURCE PARTITIONING	535
<i>Yongming Shen ; Michael Ferdman ; Peter Milder</i>	

SCALPEL: CUSTOMIZING DNN PRUNING TO THE UNDERLYING HARDWARE	
PARALLELISM	548
<i>Jiecao Yu ; Andrew Lukefahr ; David Palframan ; Ganesh Dasika ; Reetuparna Das ; Scott Mahlke</i>	
UNDERSTANDING AND OPTIMIZING ASYNCHRONOUS LOW-PRECISION STOCHASTIC GRADIENT DESCENT	561
<i>Christopher De Sa ; Matthew Feldman ; Christopher Ré ; Kunle Olukotun</i>	
AGGRESSIVE PIPELINING OF IRREGULAR APPLICATIONS ON RECONFIGURABLE HARDWARE	575
<i>Zhaoshi Li ; Leibo Liu ; Yangdong Deng ; Shouyi Yin ; Yao Wang ; Shaojun Wei</i>	
FRACTAL: AN EXECUTION MODEL FOR FINE-GRAIN NESTED SPECULATIVE PARALLELISM	587
<i>Suvinay Subramanian ; Mark C. Jeffrey ; Maleen Abeydeera ; Hyun Ryong Lee ; Victor A. Ying ; Joel Emer ; Daniel Sanchez</i>	
PARALLEL AUTOMATA PROCESSOR	600
<i>Arun Subramaniyan ; Reetuparna Das</i>	
VIYOJIT: DECOUPLING BATTERY AND DRAM CAPACITIES FOR BATTERY-BACKED DRAM	613
<i>Rajat Kateja ; Anirudh Badam ; Sriram Govindan ; Bikash Sharma ; Greg Ganger</i>	
DICE: COMPRESSING DRAM CACHES FOR BANDWIDTH AND CAPACITY	627
<i>Vinson Young ; Prashant J. Nair ; Moinuddin K. Qureshi</i>	
THE MONDRIAN DATA ENGINE	639
<i>Mario Drumond ; Alexandros Daglis ; Nooshin Mirzadeh ; Dmitrii Ustiugov ; Javier Picorel ; Babak Falsafi ; Boris Grot ; Dionisios Pnevmatikatos</i>	
JENGA: SOFTWARE-DEFINED CACHE HIERARCHIES	652
<i>Po-An Tsai ; Nathan Beckmann ; Daniel Sanchez</i>	
APPROX-NOC: A DATA APPROXIMATION FRAMEWORK FOR NETWORK-ON-CHIP ARCHITECTURES	666
<i>Rahul Boyapati ; Jiayi Huang ; Pritam Majumder ; Ki Hwan Yum ; Eun Jung Kim</i>	
THERE AND BACK AGAIN: OPTIMIZING THE INTERCONNECT IN NETWORKS OF MEMORY CUBES	678
<i>Matthew Poremba ; Itir Akgun ; Jieming Yin ; Onur Kayiran ; Yuan Xie ; Gabriel H. Loh</i>	
FOOTPRINT: REGULATING ROUTING ADAPTIVENESS IN NETWORKS-ON-CHIP	691
<i>Binzhang Fu ; John Kim</i>	
EBDA: A NEW THEORY ON DESIGN AND VERIFICATION OF DEADLOCK-FREE INTERCONNECTION NETWORKS	703
<i>Masoumeh Ebrahimi ; Masoud Daneshtalab</i>	
Author Index	