2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA 2019)

Phoenix, Arizona, USA
22 – 26 June 2019

IEEE

IEEE Catalog Number: CFP19030-POD
ISBN: 978-1-7281-4838-0

****** 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:	CFP19030-POD
ISBN (Print-On-Demand):	978-1-7281-4838-0
ISBN (Online):	978-1-4503-6669-4
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

Contents

ISCA-46 Sponsors . . . . . **vii**

General Chair Foreword . . . . . **x**

Program Chair Foreword . . . . . **xii**

ISCA-46 Organizing Committee . . . . . **xxiii**

ISCA-46 Program Committee . . . . . **xxiv**

ISCA-46 External Review Committee . . . . . **xxvi**

Perceptron-Based Prefetch Filtering **1**
Eshan Bhatia (*Texas A&M University*); Daniel A. Jiménez (*Barcelona Supercomputing Center / Texas A&M University*); Paul Gratz (*Texas A&M University*); Elvira Teran (*Texas A&M International University*); Seth Pugsley (*Intel Labs*); Gino Chacon (*Texas A&M University*)

Post-Silicon CPU Adaptations Made Practical Using Machine Learning **14**
Stephen J Tarsa (*Intel Labs*); Rangeen Basu Roy Chowdhury, Julien Sebot (*Intel Corporation*); Gautham Chinya (*Intel Labs*); Jayesh Gaur (*Intel Labs India*); Karthik Sankaranarayanan, Chit-Kwan Lin (*Intel Labs*); Robert Chappell, Ronak Singhal (*Intel Corporation*); Hong Wang (*Intel Labs*)

Bit-Level Perceptron Prediction for Indirect Branch Prediction **27**
Elba Garza, Samira Mirbagher Ajorpaz, Tahsin Ahmad Khan (*Texas A&M University*); Daniel A. Jiménez (*Barcelona Supercomputing Center and Texas A&M University*)

Generative and Multi-phase Learning for Computer Systems Optimization **39**
Yi Ding, Nikita Mishra, Henry Hoffmann (*University of Chicago*)

OO-VR: NUMA Friendly Object-Oriented VR Rendering Framework For Future NUMA-Based Multi-GPU Systems **53**
Chenhao Xie (*University of Houston; Pacific Northwest National Laboratory*); Xin Fu (*University of Houston*); Mingsong Chen (*East China Normal University*); Shuaiwen Leon Song (*Pacific Northwest National Laboratory; The University of Sydney*)

PES: Proactive Event Scheduling for Energy-Efficient Mobile Web Computing **66**
Yu Feng, Yuhao Zhu (*University of Rochester*)

3D-based Video Understanding Acceleration by Leveraging Temporal Locality and Activation Sparsity **79**
Huixiang Chen, Mingcong Song, Jiechen Zhao (*University of Florida*); Yuting Dai (*Guizhou University*); Tao Li (*University of Florida*)

Energy-Efficient Video Processing for Virtual Reality **91**
Yue Leng (*UIUC*); Chi-chun Chen, Qiuyue Sun (*University of Rochester*); Jian Huang (*UIUC*); Yuhao Zhu (*University of Rochester*)

Triad-NVM: Persistency for Integrity-Protected and Encrypted Non-Volatile Memories **104**
Amro Awad, Yan Solihin (*University of Central Florida*); Laurent Njilla (*Air Force Research Lab*); Mao Ye, Kazi Zubair (*University of Central Florida*)

GraphSSD: Graph Semantics Aware SSD 116
Kiran kumar matam (*USC*); Gunjae Koo (*Hongik University*); Haipeng Zha (*USC*); Hung-Wei Tseng (*NCSU*); Murali Anavarum (*USC*)

CROW: A Low-Cost Substrate for Improving DRAM Performance, Energy Efficiency, and Reliability 129
Hasan Hassan, Minesh Patel (*ETH Zurich*); Jeremie S. Kim (*Carnegie Mellon University; ETH Zurich*); A. Giray Yaglikci (*ETH Zurich*); Nandita Vijaykumar (*Carnegie Mellon University; ETH Zurich*); Nika Mansouri Ghiasi (*ETH Zurich*); Saugata Ghose (*Carnegie Mellon University*); Onur Mutlu (*ETH Zurich; Carnegie Mellon University*)

Janus: Optimizing Memory and Storage Support for Non-Volatile Memory Systems 143
Sihang Liu, Korakit Seemakhupt (*University of Virginia*); Gennady Pekhimenko (*University of Toronto*); Aasheesh Kolli (*Pennsylvania State University and VMware Research*); Samira Khan (*University of Virginia*)

Anubis: Low-Overhead and Practical Recovery Time for Secure Non-Volatile Memories 157
Kazi Abu Zubair, Amro Awad (*University of Central Florida*)

Emerald: Graphics Modeling for SoC Systems 169
Ayub A. Gubran, Tor M. Aamodt (*University of British Columbia*)

Linebacker: Preserving Victim Cache Lines in Idle Register Files of GPUs 183
Yunho Oh (*EPFL*); Gunjae Koo (*Hongik University*); Murali Annavaram (*University of Southern California*); Won Woo Ro (*Yonsei University*)

MGPUSim: Enabling Multi-GPU Performance Modeling and Optimization 197
Yifan Sun, Trinayan Baruah (*Northeastern University*); Saiful A. Mojumder (*Boston University*); Shi Dong, Xiang Gong, Shane Treadway, Yuhui Bao, Spencer Hance, Carter McCardwell, Vincent Zhao, Harrison Barclay (*Northeastern University*); Amir Kavyan Ziabari, Zhongliang Chen (*AMD*); Rafael Ubal (*Northeastern University*); José L. Abellán (*Universidad Católica San Antonio de Murcia*); John Kim (*KAIST*); Ajay Joshi (*Boston University*); David Kaeli (*Northeastern University*)

Opportunistic Computing in GPU Architectures 210
Ashutosh Pattnaik, Xulong Tang (*Penn State*); Onur Kayiran (*AMD Research*); Adwait Jog (*College of William & Mary*); Asit Mishra (*NVIDIA*); Mahmut T. Kandemir, Anand Sivasubramaniam, Chita Das (*Penn State*)

Interplay between Hardware Prefetcher and Page Eviction Policy in CPU-GPU Unified Virtual Memory 224
Debashis Ganguly, Rami Melhem (*Department of Computer Science, University of Pittsburgh*); Jun Yang (*Electrical and Computer Engineering, University of Pittsburgh*); Ziyu Zhang (*Department of Computer Science, University of Pittsburgh*)

Sparse ReRAM Engine: Joint Exploration of Activation and Weight Sparsity in Compressed Neural Networks 236
Tzu-Hsien Yang (*National Taiwan Univervsity*); Hsiang-Yun Cheng (*Academia Sinica*); Chia-Lin Yang, I-Ching Tseng (*National Taiwan Univervsity*); Han-Wen Hu, Hung-Sheng Chang, Hsiang-Pang Li (*Macronix International Co., Ltd.*)

MnnFast: A Fast and Scalable System Architecture for Memory-Augmented Neural Networks . . 250
Hanhwi Jang (*POSTECH*); Joonsung Kim (*Seoul National University*); Jae-Eon Jo (*POSTECH*); Jaewon Lee, Jangwoo Kim (*Seoul National University*)

TIE: Energy-efficient tensor train-based inference engine for deep neural network **264**
Chunhua Deng (*Rutgers University*); Fangxuan Sun (*Nanjing University*); Xuehai Qian (*University of Southern California*); Jun Lin, Zhongfeng Wang (*Nanjing University*); Bo Yuan (*Rutgers University*)

Accelerating Distributed Reinforcement Learning with In-Switch Computing **279**
Youjie Li, Iou-Jen Liu, Yifan Yuan, Deming Chen, Alexander Schwing, Jian Huang (*UIUC*)

Eager Pruning: Algorithm and Architecture Support for Fast Training of Deep Neural Networks . **292**
Jiaqi Zhang, Xiangru Chen, Mingcong Song, Tao Li (*University of Florida*)

Laconic Deep Learning Inference Acceleration **304**
Sayeh Sharify, Alberto Delmas Lascorz, Mostafa Mahmoud, Milos Nikolic, Kevin Siu, Dylan Malone Stuart, Zissis Poulos, Andreas Moshovos (*Toronto*)

MicroScope: Enabling Microarchitectural Replay Attacks **318**
Dimitrios Skarlatos, Mengjia Yan, Bhargava Gopireddy, Read Sprabery, Josep Torrellas, Christopher Fletcher (*University of Illinois at Urbana Champaign*)

SecDir: A Secure Directory to Defeat Directory Side-Channel Attacks **332**
Mengjia Yan, Jen-Yang Wen, Christopher Fletcher, Josep Torrellas (*University of Illinois at Urbana Champaign*)

Secure TLBs **346**
Shuwen Deng, Wenjie Xiong, Jakub Szefer (*Yale University*)

New Attacks and Defense for Encrypted-Address Cache **360**
Moinuddin K Qureshi (*Georgia Tech*)

InvisiPage: Oblivious Demand Paging for Secure Enclaves **372**
Shaizeen Aga, Satish Narayanasamy (*University of Michigan, Ann Arbor*)

TWiCe: Preventing Row-hammering by Exploiting Time Window Counters **385**
Eojin Lee, Ingab Kang (*Seoul National University*); Sukhan Lee (*Seoul National University / Samsung Electronics*); G. Edward Suh (*Cornell University*); Jung Ho Ahn (*Seoul National University*)

Duality Cache for Data Parallel Acceleration **397**
Daichi Fujiki, Scott Mahlke, Reetuparna Das (*University of Michigan*)

Adaptive Memory-Side Last-Level GPU Caching **411**
Xia Zhao, Almutaz Adileh (*Ghent University*); Zhibin Yu (*Shenzhen Institutes of Advanced Technology,CAS*); Zhiying Wang (*National University of Defense Technology*); Aamer Jaleel (*NVIDIA*); Lieven Eeckhout (*Ghent University*)

SCU: A GPU Stream Compaction Unit for Graph Processing **424**
Albert Segura, Jose-Maria Arnau, Antonio Gonzalez (*Universitat Politecnica de Catalunya*)

Filter Caching for Free: The Untapped Potential of the Store Buffer **436**
Ricardo Alves (*Uppsala University*); Alberto Ros (*Universidad de Murcia*); David Black-Schaffer, Stefanos Kaxiras (*Uppsala University*)

Efficient Meta-Data Management for Irregular Data Prefetching **449**
Hao Wu (*ARM, UT Austin*); Krishnendra Nathella, Dam Sunwoo (*ARM*); Akanksha Jain, Calvin Lin (*UT Austin*)

AsmDB: Understanding and Mitigating Front-end Stalls in Warehouse-Scale Computers **462**
Grant Ayers (*Stanford University*); Nayana Prasad Nagendra, David I. August (*Princeton University*); Hyoun Kyu Cho, Svilen Kanev (*Google*); Christos Kozyrakis (*Stanford University*); Trivikram Krishnamurthy (*Nvidia*); Heiner Litz (*UC Santa Cruz*); Tipp Moseley, Parthasarathy Ranganathan (*Google*)

Fine-grained Warm Water Cooling for Improving Datacenter Economy **474**
Weixiang Jiang, Ziyang Jia, Sirui Feng, Fangming Liu, Hai Jin (*Huazhong University of Science and Technology*)

DeepAttest: An End-to-End Attestation Framework for Deep Neural Networks **487**
Huili Chen, Cheng Fu, Bita Darvish Rouhani, Jishen Zhao, Farinaz Koushanfar (*UC San Diego*)

A Time-Space Sharing Scheduling Abstraction for Next Generation of Shared Cloud via Vertical Labels **499**
Yuzhao Wang (*Huazhong University of Science and Technology, Shenzhen Institute of Advanced Technology*); Lele Li (*Shenzhen Institute of Advanced Technology*); You Wu (*University of Southern California*); Junqing Yu (*Huazhong University of Science and Technology*); Zhibin Yu (*Shenzhen Institute of Advancec Technology*); Xuehai Qian (*University of Southern California*)

SoftSKU: Optimizing Server Architectures for Microservice Diversity @Scale **513**
Akshitha Sriraman (*University of Michigan*); Abhishek Dhanotia (*Facebook*); Thomas F. Wenisch (*University of Michigan*)

Full-Stack, Real-System Quantum Computer Studies: Architectural Comparisons and Design Insights **527**
Prakash Murali (*Princeton University*); Norbert Matthias Linke (*University of Maryland*); Margaret Martonosi (*Princeton University*); Ali Javadi Abhari (*IBM T. J. Watson Research Center*); Nhung Hong Nguyen, Cinthia Huerta Alderete (*University of Maryland*)

Statistical Assertions for Validating Patterns and Finding Bugs in Quantum Programs **541**
Yipeng Huang, Margaret Martonosi (*Princeton University*)

Asymptotic Improvements to Quantum Circuits via Qutrits **554**
Pranav Gokhale, Jonathan M. Baker, Casey Duckering (*University of Chicago*); Natalie C. Brown (*Georgia Institute of Technology*); Kenneth R. Brown (*Duke University*); Frederic T. Chong (*University of Chicago*)

A Stochastic-Computing based Deep Learning Framework using Adiabatic Quantum-Flux-Parametron Superconducting Technology **567**
Ruizhe Cai, Ao Ren (*Northeastern University*); Olivia Chen (*Yokohama National University*); Ning Liu, Caiwen Ding (*Northeastern University*); Xuehai Qian (*University of Southern California*); Jie Han (*University of Alberta*); Wenhui Luo, Nobuyuki Yoshikawa (*Yokohama National University*); Yanzhi Wang (*Northeastern University*)

A Quantum Computational Compiler and Design Tool for Technology-Specific Targets **579**
Kaitlin Smith, Mitch Thornton (*Southern Methodist University*)

IntelliNoC: A Holistic Design Framework for Energy-Efficient and Reliable On-Chip Communication for Manycores **589**
Ke Wang, Ahmed Louri (*George Washington University*); Avinash Karanth, Razvan Bunescu (*Ohio University*)

HALO: Accelerating Flow Classification for Scalable Packet Processing in NFV **601**
Yifan Yuan (*UIUC*); Yipeng Wang, Ren Wang (*Intel Labs*); Jian Huang (*UIUC*)

Scalable Interconnects for Reconfigurable Spatial Architectures **615**
Yaqi Zhang, Alexander Rucker, Matthew Vilim, Raghu Prabhakar, William Hwang, Kunle Olukotun (*Stanford University*)

CoNDA: Efficient Cache Coherence Support for Near-Data Accelerators **629**
Amirali Boroumand, Saugata Ghose (*Carnegie Mellon University*); Minesh Patel (*Carnegie Mellon University, ETH*); Rachata Ausavarungnirun (*Carnegie Mellon University, King Mongkut's University of Technology North Bangkok*); Hasan Hassan (*Carnegie Mellon University, ETH*); Brandon Lucia, Kevin Hsieh (*Carnegie Mellon University*); Nastaran Hajinazar (*Carnegie Mellon University, Simon Fraser University*); Krishna T. Malladi, Hongzhong Zheng (*Samsung Semiconductor, Inc.*); Onur Mutlu (*ETH, Carnegie Mellon University*)

Designing Vertical Processors in Monolithic 3D **643**
Bhargava Gopireddy, Josep Torrellas (*University of Illinois at Urbana Champaign*)

Time Squeezing for Tiny Devices **657**
Yuanbo Fan, Simone Campanoni, Russ Joseph (*Northwestern University*)

XPC: Architectural Support for Secure and Efficient Cross Process Call **671**
Dong Du, Zhichao Hua, Yubin Xia, Binyu Zang, Haibo Chen (*Shanghai Jiao Tong University*)

AxMemo: Hardware-Compiler Co-Design for Approximate Code Memoization **685**
Zhenhong Liu (*UIUC*); Amir Yazdanbakhsh (*Google Brain*); Dong Kai Wang (*UIUC*); Hadi Esmaeilzadeh (*UCSD*); Nam Sung Kim (*UIUC*)

Translation Ranger: Operating System Support for Contiguity-Aware TLBs **698**
Zi Yan (*Rutgers University/NVIDIA*); Daniel Lustig, David Nellans (*NVIDIA*); Abhishek Bhattacharjee (*Yale University*)

Bouncer: Static Program Analysis in Hardware **711**
Joseph McMahan, Michael Christensen (*University of California, Santa Barbara*); Kyle Dewey (*California State University Northridge*); Ben Hardekopf, Timothy Sherwood (*University of California, Santa Barbara*)

Efficient Invisible Speculative Execution through Selective Delay and Value Prediction **723**
Christos Sakalis, Stefanos Kaxiras (*Uppsala University*); Alberto Ros (*University of Murcia*); Alexandra Jimborean (*Uppsala University*); Magnus Själander (*Norwegian University of Science and Technology*)

Stream-based Memory Access Specialization for General Purpose Processors **736**
Zhengrong Wang, Tony Nowatzki (*UCLA*)

Using SMT to accelerate nested virtualization **750**
Lluís Vilanova (*Technion*); Nadav Amit (*VMWare Research*); Yoav Etsion (*Technion*)

Master of None Acceleration: A Comparison of Accelerator Architectures for Analytical Query Processing **762**
Andrea Lottarini (*Google*); João P. Cerqueira, Thomas J. Repetti, Stephen A. Edwards, Kenneth A. Ross, Mingoo Seok, Martha A. Kim (*Columbia University*)

Cryogenic Computer Architecture Modeling with Memory-Side Case Studies **774**
Gyu-Hyeon Lee, Dongmoon Min, Il-Kwon Byun, Jangwoo Kim (*Department of Electrical and Computer Engineering, Seoul National University*)

Cambricon-F: Machine Learning Computers with Fractal von Neumann Architecture **788**
Yongwei Zhao, Zidong Du, Qi Guo, Shaoli Liu (*Institute of Computing Technology, Chinese Academy of Sciences*); Ling Li (*Institute of Software, Chinese Academy of Sciences*); Zhiwei Xu, Tianshi Chen, Yunji Chen (*Institute of Computing Technology, Chinese Academy of Sciences*)

FloatPIM: In-Memory Acceleration of Deep Neural Network Training with High Precision **802**
Mohsen Imani, Saransh Gupta, Yeseong Kim, Tajana Rosing (*UC San Diego*)

Author index **816**

ISCA-46 Other External Reviewers **819**