

# **2020 IEEE International Symposium on High Performance Computer Architecture (HPCA 2020)**

**San Diego, California, USA  
22 – 26 February 2020**



**IEEE Catalog Number: CFP20013-POD  
ISBN: 978-1-7281-6150-1**

**Copyright © 2020 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:	CFP20013-POD
ISBN (Print-On-Demand):	978-1-7281-6150-1
ISBN (Online):	978-1-7281-6149-5
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](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

CURRAN ASSOCIATES INC.  
**proceedings**  
.com

# 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA) **HPCA 2020**

## Table of Contents

Message from the Program Chairs	xiii
Organizing Committee	xv
Program Committee	xvi
Industry Session Committee	xviii
External Review Committee	xix
Sponsors	xxi

### Machine Learning Acceleration

Deep Learning Acceleration with Neuron-to-Memory Transformation	1
<i>Mohsen Imani (University of California San Diego), Mohammad Samragh Razlighi (University of California San Diego), Yeseong Kim (University of California San Diego), Saransh Gupta (University of California San Diego), Farinaz Koushanfar (University of California San Diego), and Tajana Rosing (University of California San Diego)</i>	
HyGCN: A GCN Accelerator with Hybrid Architecture	15
<i>Mingyu Yan (Chinese Academy of Sciences), Lei Deng (University of California, Santa Barbara), Xing Hu (University of California, Santa Barbara), Ling Liang (University of California, Santa Barbara), Yujing Feng (Chinese Academy of Sciences), Xiaochun Ye (Chinese Academy of Sciences), Zhimin Zhang (Chinese Academy of Sciences), Dongrui Fan (Chinese Academy of Sciences), and Yuan Xie (University of California, Santa Barbara)</i>	

### Reliability and Fault Tolerance

ACR: Amnesic Checkpointing and Recovery	30
<i>Ismail Akturk (University of Missouri, Columbia) and Ulya R. Karpuzcu (University of Minnesota, Twin Cities)</i>	

Asymmetric Resilience: Exploiting Task-Level Idempotency for Transient Error Recovery in Accelerator-Based Systems .44.....  
*Jingwen Leng (Shanghai Jiao Tong University), Alper Buyuktosunoglu (IBM T. J. Watson Research Center), Ramon Bertran (IBM T. J. Watson Research Center), Pradip Bose (IBM T. J. Watson Research Center), Quan Chen (Shanghai Jiao Tong University), Minyi Guo (Shanghai Jiao Tong University), and Vijay Janapa Reddi (Harvard University)*

## Best Paper Nominees

SIGMA: A Sparse and Irregular GEMM Accelerator with Flexible Interconnects for DNN Training .58  
*Eric Qin (Georgia Institute of Technology), Ananda Samajdar (Georgia Institute of Technology), Hyoukjun Kwon (Georgia Institute of Technology), Vineet Nadella (Georgia Institute of Technology), Sudarshan Srinivasan (Intel), Dipankar Das (Intel), Bharat Kaul (Intel), and Tushar Krishna (Georgia Institute of Technology)*

EMSim: A Microarchitecture-Level Simulation Tool for Modeling Electromagnetic Side-Channel Signals .71.....  
*Nader Sehatbakhsh (Georgia Institute of Technology), Baki Berkay Yilmaz (Georgia Institute of Technology), Alenka Zajic (Georgia Institute of Technology), and Milos Prvulovic (Georgia Institute of Technology)*

Impala: Algorithm/Architecture Co-Design for In-Memory Multi-Stride Pattern Matching .86.....  
*Elaheh Sadredini (University of Virginia), Reza Rahimi (University of Virginia), Marzieh Lenjani (University of Virginia), Mircea Stan (University of Virginia), and Kevin Skadron (University of Virginia)*

A Deep Reinforcement Learning Framework for Architectural Exploration: A Routerless NoC Case Study .99.....  
*Ting-Ru Lin (University of Southern California), Drew Penney (Oregon State University), Massoud Pedram (University of Southern California), and Lizhong Chen (Oregon State University)*

## Security and NoC

IRONHIDE: A Secure Multicore that Efficiently Mitigates Microarchitecture State Attacks for Interactive Applications .111.....  
*Hamza Omar (University of Connecticut) and Omer Khan (University of Connecticut)*

A New Side-Channel Vulnerability on Modern Computers by Exploiting Electromagnetic Emanations from the Power Management Unit .123.....  
*Nader Sehatbakhsh (Georgia Institute of Technology), Baki Berkay Yilmaz (Georgia Institute of Technology), Alenka Zajic (Georgia Institute of Technology), and Milos Prvulovic (Georgia Institute of Technology)*

Leaking Information Through Cache LRU States .139.....  
*Wenjie Xiong (Yale University) and Jakub Szefer (Yale University)*

Baldur: A Power-Efficient and Scalable Network Using All-Optical Switches .153.....  
*Mohammad Reza Jokar (University of Chicago), Junyi Qiu (University of Illinois at Urbana-Champaign), Frederic T. Chong (University of Chicago), Lynford L. Goddard (University of Illinois at Urbana-Champaign), John M. Dallesasse (University of Illinois at Urbana-Champaign), Milton Feng (University of Illinois at Urbana-Champaign), and Yanjing Li (University of Chicago)*

## Cloud

Twig: Multi-Agent Task Management for Colocated Latency-Critical Cloud Services .167.....  
*Rajiv Nishtala (Norwegian University of Science and Technology), Vinicius Petrucci (Federal University of Bahia, University of Pittsburgh), Paul Carpenter (Barcelona Supercomputing Center), and Magnus Sjalander (Norwegian University of Science and Technology)*

QuickNN: Memory and Performance Optimization of k-d Tree Based Nearest Neighbor Search for 3D Point

Clouds .180.....  
*Reid Pinkham (University of Michigan), Shuqing Zeng (General Motors), and Zhengya Zhang (University of Michigan)*

CLITE: Efficient and QoS-Aware Co-Location of Multiple Latency-Critical Jobs for Warehouse Scale Computers .193.....  
*Tirthak Patel (Northeastern University) and Devesh Tiwari (Northeastern University)*

Q-Zilla: A Scheduling Framework and Core Microarchitecture for Tail-Tolerant Microservices .207.  
*Amirhossein Mirhosseini (University of Michigan), Brendan L. West (University of Michigan), Geoffrey W. Blake (Amazon Web Services), and Thomas F. Wenisch (University of Michigan)*

## Accelerator and DSA

PREMA: A Predictive Multi-Task Scheduling Algorithm For Preemptible Neural Processing Units 220  
*Yujeong Choi (Korea Advanced Institute of Science and Technology) and Minsoo Rhu (Korea Advanced Institute of Science and Technology)*

Domain-Specialized Cache Management for Graph Analytics .234.....  
*Priyank Faldu (The University of Edinburgh), Jeff Diamond (Oracle Labs), and Boris Grot (The University of Edinburgh)*

ALRESCHA: A Lightweight Reconfigurable Sparse-Computation Accelerator .249.....  
*Bahar Asgari (Georgia Institute of Technology), Ramyad Hadidi (Georgia Institute of Technology), Tushar Krishna (Georgia Institute of Technology), Hyesoon Kim (Georgia Institute of Technology), and Sudhakar Yalamanchili (Georgia Institute of Technology)*

SpArch: Efficient Architecture for Sparse Matrix Multiplication .261.....  
*Zhekai Zhang (Massachusetts Institute of Technology), Hanrui Wang (Massachusetts Institute of Technology), Song Han (Massachusetts Institute of Technology), and William J. Dally (Stanford University / NVIDIA)*

## Memory and Memory Hierarchy

Mitigating Voltage Drop in Resistive Memories by Dynamic RESET Voltage Regulation and Partition  
RESET .275.....

*Farzaneh Zokae (Indiana University Bloomington) and Lei Jiang  
(Indiana University Bloomington)*

DRAM-Less: Hardware Acceleration of Data Processing with New Memory .287.....

*Jie Zhang (Korea Advanced Institute of Science and Technology),  
Gyuyoung Park (Korea Advanced Institute of Science and Technology),  
David Donofrio (Lawrence Berkeley National Laboratory), John Shalf  
(Lawrence Berkeley National Laboratory), and Myoungsoo Jung (Korea  
Advanced Institute of Science and Technology)*

ELP2IM: Efficient and Low Power Bitwise Operation Processing in DRAM .303.....

*Xin Xin (University of Pittsburgh), Youtao Zhang (University of  
Pittsburgh), and Jun Yang (University of Pittsburgh)*

ResiRCA: A Resilient Energy Harvesting ReRAM Crossbar-Based Accelerator for Intelligent Embedded  
Processors .315.....

*Keni Qiu (Capital Normal University), Nicholas Jao (Pennsylvania State  
University), Mengying Zhao (Shandong University), Cyan Subhra Mishra  
(Pennsylvania State University), Gulsum Gudukbay (Pennsylvania State  
University), Sethu Jose (Pennsylvania State University), Jack Sampson  
(Pennsylvania State University), Mahmut Taylan Kandemir (Pennsylvania  
State University), and Vijaykrishnan Narayanan (Pennsylvania State  
University)*

## Machine Learning Acceleration

A<sup>3</sup>: Accelerating Attention Mechanisms in Neural Networks with Approximation .328.....

*Tae Jun Ham (Seoul National University), Sung Jun Jung (Seoul National  
University), Seonghak Kim (Seoul National University), Young H. Oh  
(Sungkyunkwan University), Yeonhong Park (Seoul National University),  
Yoonho Song (Seoul National University), Jung-Hun Park (Seoul National  
University), Sanghee Lee (Seoul National University), Kyoung Park (SK  
Hynix), Jae W. Lee (Seoul National University), and Deog-Kyoon Jeong  
(Seoul National University)*

AccPar: Tensor Partitioning for Heterogeneous Deep Learning Accelerators .342.....

*Linghao Song (Duke University), Fan Chen (Duke University), Youwei  
Zhuo (University of Southern California), Xuehai Qian (University of  
Southern California), Hai Li (Duke University), and Yiran Chen (Duke  
University)*

## Fault Tolerance and Security

- FLOWER and FaME: A Low Overhead Bit-Level Fault-map and Fault-Tolerance Approach for Deeply Scaled Memories .356.....  
*Donald Kline Jr. (University of Pittsburgh), Jiangwei Zhang (University of Pittsburgh), Rami Melhem (University of Pittsburgh), and Alex K. Jones (University of Pittsburgh)*
- Multi-Range Supported Oblivious RAM for Efficient Block Data Retrieval .369.....  
*Yuezhi Che (Illinois Institute of Technology) and Rujia Wang (Illinois Institute of Technology)*

## Microarchitecture

- CASINO Core Microarchitecture: Generating Out-of-Order Schedules Using Cascaded In-Order Scheduling Windows .383.....  
*Ipoom Jeong (Yonsei University), Seihoon Park (Yonsei University), Changmin Lee (Samsung Electronics), and Won Woo Ro (Yonsei University)*
- Precise Runahead Execution .397.....  
*Ajeya Naithani (Ghent University), Josué Feliu (Universitat Politècnica de Valencia), Almutaz Adileh (Ghent University), and Lieven Eeckhout (Ghent University)*
- BBS: Micro-Architecture Benchmarking Blockchain Systems through Machine Learning and Fuzzy Set .411.....  
*Liang Zhu (Shenzhen Institutes of Advanced Technology (SIAT), CAS), Chao Chen (Shenzhen Institutes of Advanced Technology (SIAT), CAS), Zihao Su (Shenzhen Institutes of Advanced Technology (SIAT), CAS), Weiguang Chen (Shenzhen Institutes of Advanced Technology (SIAT), CAS), Tao Li (University of Florida), and Zhibin Yu (Shenzhen Institutes of Advanced Technology (SIAT), CAS)*
- Delay and Bypass: Ready and Criticality Aware Instruction Scheduling in Out-of-Order Processors .... 424  
*Mehdi Alipour (Uppsala University), Stefanos Kaxiras (Uppsala University), David Black-Schaffer (Uppsala University), and Rakesh Kumar (Norwegian University of Science and Technology)*

## NoC

- EquiNox: Equivalent NoC Injection Routers for Silicon Interposer-Based Throughput Processors .435  
*Yunfan Li (Oregon State University) and Lizhong Chen (Oregon State University)*
- DRAIN: Deadlock Removal for Arbitrary Irregular Networks .447.....  
*Mayank Parasar (Georgia Institute of Technology), Hossein Farrokhbakht (University of Toronto), Natalie Enright Jerger (University of Toronto), Paul V. Gratz (Texas A & M), Tushar Krishna (Georgia Institute of Technology), and Joshua San Miguel (University of Wisconsin-Madison)*

SnackNoC: Processing in the Communication Layer .461.....  
*Karthik Sangaiah (Drexel University), Michael Lui (Drexel University),  
Ragh Kuttappa (Drexel University), Baris Taskin (Drexel University),  
and Mark Hempstead (Tufts University)*

PIXEL: Photonic Neural Network Accelerator .474.....  
*Kyle Shiflett (Ohio University), Dylan Wright (Ohio University),  
Avinash Karanth (Ohio University), and Ahmed Louri (George Washington  
University)*

## Industry Session 1

The Architectural Implications of Facebook's DNN-Based Personalized Recommendation .488.....  
*Udit Gupta (Facebook, Harvard University), Carole-Jean Wu (Facebook),  
Xiaodong Wang (Facebook), Maxim Naumov (Facebook), Brandon Reagen  
(Facebook), David Brooks (Harvard University), Bradford Cattel  
(Facebook), Kim Hazelwood (Facebook), Mark Hempstead (Facebook), Bill  
Jia (Facebook), Hsien-Hsin S. Lee (Facebook), Andrey Malevich  
(Facebook), Dheevatsa Mudigere (Facebook), Mikhail Smelyanskiy  
(Facebook), Liang Xiong (Facebook), and Xuan Zhang (Facebook)*

NVDIMM-C: A Byte-Addressable Non-Volatile Memory Module for Compatibility with Standard DDR  
Memory  
Interfaces .502.....  
*Changmin Lee (Samsung Electronics), Wonjae Shin (Samsung Electronics),  
Dae Jeong Kim (Samsung Electronics), Yongjun Yu (Samsung Electronics),  
Sung-Joon Kim (Samsung Electronics), Taekyeong Ko (Samsung  
Electronics), Deokho Seo (Samsung Electronics), Jongmin Park (Samsung  
Electronics), Kwanghee Lee (Samsung Electronics), Seongho Choi  
(Samsung Electronics), Namhyung Kim (Samsung Electronics), Vishak G  
(Samsung Electronics), Arun George (Samsung Electronics), Vishwas V  
(Samsung Electronics), Donghun Lee (SAP Labs Korea), Kangwoo Choi (SAP  
Labs Korea), Changbin Song (SAP Labs Korea), Dohan Kim (Samsung  
Electronics), Insu Choi (Samsung Electronics), Ilgyu Jung (Samsung  
Electronics), Yong Ho Song (Samsung Electronics), and Jinman Han  
(Samsung Electronics)*

Missing the Forest for the Trees: End-to-End AI Application Performance in Edge Data Centers .515  
*Daniel Richins (The University of Texas at Austin), Dharmisha Doshi  
(Intel), Matthew Blackmore (Intel), Aswathy Thulaseedharan Nair  
(Intel), Neha Pathapati (Intel), Ankit Patel (Intel), Brainard Daguman  
(Intel), Daniel Dobrijalowski (Intel), Ramesh Illikkal (Intel), Kevin  
Long (Intel), David Zimmerman (Intel), and Vijay Janapa Reddi (The  
University of Texas at Austin, Harvard University)*

## Accelerators and DSA 2

Communication Lower Bound in Convolution Accelerators .529.....  
*Xiaoming Chen (Chinese Academy of Sciences), Yinhe Han (Chinese  
Academy of Sciences), and Yu Wang (Tsinghua University)*



Enabling Highly Efficient Capsule Networks Processing Through A PIM-Based Architecture Design ...  
542

*Xingyao Zhang (University of Houston), Shuaiwen Leon Song (University of Sydney), Chenhao Xie (Pacific Northwest National Lab), Jing Wang (Capital Normal University), Weigong Zhang (Beijing Advanced Innovation Center for Imaging Theory and Technology), and Xin Fu (University of Houston)*

Fulcrum: A Simplified Control and Access Mechanism Toward Flexible and Practical In-Situ Accelerators .556.....

*Marzieh Lenjani (University of Virginia), Patricia Gonzalez (University of Virginia), Elaheh Sadredini (University of Virginia), Shuangchen Li (University of California, Santa Barbara), Yuan Xie (University of California, Santa Barbara), Ameen Akel (Micron Technology, Inc.), Sean Eilert (Micron Technology, Inc.), Mircea R. Stan (University of Virginia), and Kevin Skadron (University of Virginia)*

## GPUs

BCoal: Bucketing-Based Memory Coalescing for Efficient and Secure GPUs .570.....

*Gurunath Kadam (William & Mary), Danfeng Zhang (Pennsylvania State University), and Adwait Jog (William & Mary)*

HMG: Extending Cache Coherence Protocols Across Modern Hierarchical Multi-GPU Systems .582

*Xiaowei Ren (NVIDIA), Daniel Lustig (NVIDIA), Evgeny Bolotin (NVIDIA), Aamer Jaleel (NVIDIA), Oreste Villa (NVIDIA), and David Nellans (NVIDIA)*

Griffin: Hardware-Software Support for Efficient Page Migration in Multi-GPU Systems .596.....

*Trinayan Baruah (Northeastern University), Yifan Sun (Northeastern University), Ali Tolga Dinçer (Istanbul Technical University), Saiful A. Mojumder (Boston University), José L. Abellán (Universidad Católica San Antonio de Murcia), Yash Ukidave (Millennium USA), Ajay Joshi (Boston University), Norman Rubin (Northeastern University), John Kim (KAIST), and David Kaeli (Northeastern University)*

## Industry Session 2

EFLOPS: Algorithm and System Co-Design for a High Performance Distributed Training Platform .....  
610

*Jianbo Dong (Alibaba Group), Zheng Cao (Alibaba Group), Tao Zhang (Alibaba Group), Jianxi Ye (Alibaba Group), Shaochuang Wang (Alibaba Group), Fei Feng (Alibaba Group), Li Zhao (Alibaba Group), Xiaoyong Liu (Alibaba Group), Liuyihan Song (Alibaba Group), Liwei Peng (Alibaba Group), Yiqun Guo (Alibaba Group), Xiaowei Jiang (Alibaba Group), Lingbo Tang (Alibaba Group), Yin Du (Alibaba Group), Yingya Zhang (Alibaba Group), Pan Pan (Alibaba Group), and Yuan Xie (Alibaba Group)*

Techniques for Reducing the Connected-Standby Energy Consumption of Mobile Devices ..... 623  
*Jawad Haj-Yahya (ETH Zurich), Yanos Sazeides (University of Cyprus),  
Mohammed Alser (ETH Zurich), Efraim Rotem (Intel), and Onur Mutlu (ETH  
Zurich)*

Experiences with ML-Driven Design: A NoC Case Study ..... 637  
*Jieming Yin (Advanced Micro Devices, Inc.), Subhash Sethumurugan  
(University of Minnesota, Twin Cities), Yasuko Eckert (Advanced Micro  
Devices, Inc.), Chintan Patel (Advanced Micro Devices, Inc.), Alan  
Smith (Advanced Micro Devices, Inc.), Eric Morton (Advanced Micro  
Devices, Inc.), Mark Oskin (University of Washington), Natalie Enright  
Jerger (University of Toronto), and Gabriel H. Loh (Advanced Micro  
Devices, Inc.)*

## Memory and Memory Hierarchy and Cloud

Hybrid2: Combining Caching and Migration in Hybrid Memory Systems ..... 649  
*Evangelos Vasilakis (Chalmers University of Technology), Vassilis  
Papafstathiou (Foundation for Research and Technology - Hellas),  
Pedro Trancoso (Chalmers University of Technology), and Ioannis  
Sourdis (Chalmers University of Technology)*

Charge-Aware DRAM Refresh Reduction with Value Transformation ..... 663  
*Seikwon Kim (Samsung Electronics), Wonsang Kwak (KAIST), Changdae Kim  
(ETRI), Daehyeon Baek (KAIST), and Jaehyuk Huh (KAIST)*

DWT: Decoupled Workload Tracing for Data Centers ..... 677  
*Jian Chen (Alibaba Group), Ying Zhang (Alibaba Group), Xiaowei Jiang  
(Alibaba Group), Li Zhao (Alibaba Group), Zheng Cao (Alibaba Group),  
and Qiang Liu (Alibaba Group)*

## Accelerators and DSA 3

Tensaurus: A Versatile Accelerator for Mixed Sparse-Dense Tensor Computations ..... 689  
*Nitish Srivastava (Cornell University), Hanchen Jin (Cornell  
University), Shaden Smith (Microsoft AI and Research), Hongbo Rong  
(Intel Parallel Computing Lab), David Albonesi (Cornell University),  
and Zhiru Zhang (Cornell University)*

A Hybrid Systolic-Dataflow Architecture for Inductive Matrix Algorithms ..... 703  
*Jian Weng (University of California, Los Angeles), Sihao Liu  
(University of California, Los Angeles), Zhengrong Wang (University of  
California, Los Angeles), Vidushi Dadu (University of California, Los  
Angeles), and Tony Nowatzki (University of California, Los Angeles)*

Improving Predication Efficiency through Compaction/Restoration of SIMD Instructions ..... 717  
*Adrián Barredo (Barcelona Supercomputing Center), Juan M. Cebrian  
(Barcelona Supercomputing Center), Miquel Moretó (Barcelona  
Supercomputing Center), Marc Casas (Barcelona Supercomputing Center),  
and Mateo Valero (Barcelona Supercomputing Center)*

**Author Index** ..... 729