2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA 2022)

Virtual Symposium
2-6 April 2022

Pages 1-618

IEEE Catalog Number: CFP22013-POD
ISBN: 978-1-6654-2028-0

1/2

**Copyright © 2022 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:	CFP22013-POD
ISBN (Print-On-Demand):	978-1-6654-2028-0
ISBN (Online):	978-1-6654-2027-3
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

2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)

HPCA 2022

Table of Contents

Message from the General Chairs .... **xviii**
Message from the Program Chair .... **xx**
Message from the Industry Track Program Chair .... **xxv**
Sponsors .... **xxvi**

Session 1A: Accelerators I

Direct Spatial Implementation of Sparse Matrix Multipliers for Reservoir Computing 1
Matthew Denton (Google Research) and Herman Schmit (Google Research)

uSystolic: Byte-Crawling Unary Systolic Array 12
Di Wu (University of Wisconsin-Madison, USA) and Joshua San Miguel (University of Wisconsin-Madison, USA)

CAMA: Energy and Memory Efficient Automata Processing in Content-Addressable Memories 25
Yi Huang (Rice University), Zhiyu Chen (Rice University), Dai Li (Rice University), and Kaiyuan Yang (Rice University)

CoopMC: Algorithm-Architecture Co-Optimization for Markov Chain Monte Carlo Accelerators 38
Yuji Chai (Harvard University, USA; Stochastic Inc., USA), Glenn G. Ko (Harvard University, USA; Stochastic Inc., USA), Wei-Te Mark Ting (Harvard University, USA; Stochastic Inc., USA), Luke Bailey (Harvard University, USA; Stochastic Inc., USA), David Brooks (Harvard University, USA), and Gu-Yeon Wei (Harvard University, USA)

Session 1B: Security I

Leaky Frontends: Security Vulnerabilities in Processor Frontends 53
Shuwen Deng (Yale University, USA), Bowen Huang (Yale University, USA), and Jakub Szefer (Yale University, USA)

DPrime+DAbort: A High-Precision and Timer-Free Directory-Based Side-Channel Attack in Non-Inclusive Cache Hierarchies using Intel TSX 67
Sowoong Kim (UNIST, Republic of Korea), Myeonggyun Han (UNIST, Republic of Korea), and Woongki Baek (UNIST, Republic of Korea)

Abusing Cache Line Dirty States to Leak Information in Commercial Processors 82
Yujie Cui (Peking University, China), Chun Yang (Peking University, China), and Xu Cheng (Peking University, China)

unXpec: Breaking Undo-Based Safe Speculation 98
Mengming Li (Zhejiang University), Chenlu Miao (Zhejiang University), Yilong Yang (Zhejiang University), and Kai Bu (Zhejiang University)

Session 1C: At Scale

Cottage: Coordinated Time Budget Assignment for Latency, Quality and Power Optimization in Web Search 113
Liang Zhou (University of California Riverside, USA), Laxmi N. Bhuyan (University of California Riverside, USA), and K. K. Ramakrishnan (University of California Riverside, USA)

Enabling Efficient Large-Scale Deep Learning Training with Cache Coherent Disaggregated Memory Systems 126
Zixuan Wang (University of California, San Diego), Joonseop Sim (System Architecture Division, SK hynix), Euicheol Lim (System Architecture Division, SK hynix), and Jishen Zhao (University of California, San Diego)

Hercules: Heterogeneity-Aware Inference Serving for At-Scale Personalized Recommendation 141
Liu Ke (Meta; Washington University in St Louis, USA), Udit Gupta (Meta; Harvard University, USA), Mark Hempstead (Tufts University, USA), Carole-Jean Wu (Meta, USA), Hsien-Hsin S. Lee (Meta, USA), and Xuan Zhang (Washington University in St. Louis, USA)

ReTail: Opting for Learning Simplicity to Enable QoS-Aware Power Management in the Cloud 155
Shuang Chen (Cornell University, USA), Angela Jin (Cornell University, USA), Christina Delimitrou (Cornell University, USA), and José F. Martínez (Cornell University, USA)

Session 2A: Accelerators II

ANNA: Specialized Architecture for Approximate Nearest Neighbor Search 169
Yejin Lee (Seoul National University, South Korea), Hyunji Choi (Seoul National University, South Korea), Sunhong Min (Seoul National University, South Korea), Hyunseung Lee (Seoul National University, South Korea), Sangwon Beak (Seoul National University, South Korea), Dawoon Jeong (Seoul National University, South Korea), Jae W. Lee (Seoul National University, South Korea), and Tae Jun Ham (Seoul National University, South Korea)

Hardware-Accelerated Hypergraph Processing with Chain-Driven Scheduling 184
Qinggang Wang (Huazhong University of Science and Technology, China), Long Zheng (Huazhong University of Science and Technology, China), Jingrui Yuan (Huazhong University of Science and Technology, China), Yu Huang (Huazhong University of Science and Technology, China), Pengcheng Yao (Huazhong University of Science and Technology, China), Chuangyi Gui (Huazhong University of Science and Technology, China), Ao Hu (Huazhong University of Science and Technology, China), Xiaofei Liao (Huazhong University of Science and Technology, China), and Hai Jin (Huazhong University of Science and Technology, China)

ScalaGraph: A Scalable Accelerator for Massively Parallel Graph Processing 199
Pengcheng Yao (Huazhong University of Science and Technology, China), Long Zheng (Huazhong University of Science and Technology, China), Yu Huang (Huazhong University of Science and Technology, China), Qinggang Wang (Huazhong University of Science and Technology, China), Chuangyi Gui (Huazhong University of Science and Technology, China), Zhen Zeng (Huazhong University of Science and Technology, China), Xiaofei Liao (Huazhong University of Science and Technology, China), Hai Jin (Huazhong University of Science and Technology, China), and Jingling Xue (University of New South Wales, Australia)

Session 2B: Security II

Adaptive Security Support for Heterogeneous Memory on GPUs 213
Shougang Yuan (North Carolina State University, USA), Amro Awad (North Carolina State University, USA), Ardhi Wiratama Baskara Yudha (University of Central Florida, USA), Yan Solihin (University of Central Florida, USA), and Huiyang Zhou (North Carolina State University, USA)

TNPU: Supporting Trusted Execution with Tree-Less Integrity Protection for Neural Processing Unit 229
Sunho Lee (KAIST), Jungwoo Kim (KAIST), Seonjin Na (KAIST), Jongse Park (KAIST), and Jaehyuk Huh (KAIST)

SecNDP: Secure Near-Data Processing with Untrusted Memory 244
Wenjie Xiong (Meta, USA), Liu Ke (Meta; Washington University in St. Louis, USA), Dimitrije Jankov (Rice University, USA), Michael Kounavis (Meta, USA), Xiaochen Wang (Meta, USA), Eric Northup (Meta, USA), Jie Amy Yang (Meta, USA), Bilge Acun (Meta, USA), Carole-Jean Wu (Meta, USA), Ping Tak Peter Tang (Meta, USA), G. Edward Suh (Meta; Cornell University, USA), Xuan Zhang (Washington University in St. Louis, USA), and Hsien-Hsin S. Lee (Meta, USA)

Session 2C: Quantum I

AFS: Accurate, Fast, and Scalable Error-Decoding for Fault-Tolerant Quantum Computers 259
Poulami Das (Georgia Institute of Technology, Georgia), Christopher A. Pattison (California Institute of Technology, California), Srilatha Manne (Microsoft Quantum and Microsoft Research, Washington), Douglas M. Carmean (Microsoft Quantum and Microsoft Research, Washington), Krysta M. Svore (Microsoft Quantum and Microsoft Research, Washington), Moinuddin Qureshi (Georgia Institute of Technology, Georgia), and Nicolas Delfosse (Microsoft Quantum and Microsoft Research, Washington)

QULATIS: A Quantum Error Correction Methodology Toward Lattice Surgery 274
Yosuke Ueno (The University of Tokyo, Japan; Keio University, Japan), Masaaki Kondo (Keio University, Japan; RIKEN Center for Computational Science), Masamitsu Tanaka (Nagoya University, Japan), Yasunari Suzuki (NTT, Japan; JST PRESTO), and Yutaka Tabuchi (RIKEN, Japan)

VAQEM: A Variational Approach to Quantum Error Mitigation 288
Gokul Subramanian Ravi (University of Chicago), Kaitlin N. Smith (University of Chicago), Pranav Gokhale (Super.tech), Andrea Mari (Unitary Fund), Nathan Earnest (IBM), Ali Javadi-Abhari (IBM), and Frederic T. Chong (University of Chicago)

Session 3A: Accelerators III

DRIPS: Dynamic Rebalancing of Pipelined Streaming Applications on CGRAs 304
Cheng Tan (Pacific Northwest National Laboratory), Nicolas Bohm Agostini (Pacific Northwest National Laboratory), Tong Geng (Pacific Northwest National Laboratory), Chenhao Xie (Pacific Northwest National Laboratory), Jiajia Li (William & Mary), Ang Li (Pacific Northwest National Laboratory), Kevin J. Barker (Pacific Northwest National Laboratory), and Antonino Tumeo (Pacific Northwest National Laboratory)

Parallel Time Batching: Systolic-Array Acceleration of Sparse Spiking Neural Computation 317
Jeong-Jun Lee (University of California, Santa Barbara), Wenrui Zhang (University of California, Santa Barbara), and Peng Li (University of California, Santa Barbara)

Near-Stream Computing: General and Transparent Near-Cache Acceleration 331
Zhengrong Wang (University of California, Los Angeles), Jian Weng (University of California, Los Angeles), Sihao Liu (University of California, Los Angeles), and Tony Nowatzki (University of California, Los Angeles)

Session 3B: Security III

HyBP: Hybrid Isolation-Randomization Secure Branch Predictor 346
Lutan Zhao (University of Chinese Academy of Sciences), Peinan Li (University of Chinese Academy of Sciences), Rui Hou (University of Chinese Academy of Sciences), Michael C. Huang (University of Rochester), Xuehai Qian (University of Southern California), Lixin Zhang (Freelance), and Dan Meng (University of Chinese Academy of Sciences)

IR-ORAM: Path Access Type Based Memory Intensity Reduction for Path-ORAM 360
Mehrnoosh Raoufi (University of Pittsburgh, USA), Youtao Zhang (University of Pittsburgh, USA), and Jun Yang (University of Pittsburgh, USA)

SafeGuard: Reducing the Security Risk from Row-Hammer via Low-Cost Integrity Protection 373
Ali Fakhrzadehgan (University of Texas at Austin), Yale N. Patt (University of Texas at Austin), Prashant J. Nair (University of British Columbia), and Moinuddin K. Qureshi (Georgia Institute of Technology)

Session 3C: Quantum II

Detecting Qubit-Coupling Faults in Ion-Trap Quantum Computers 387
Andrii Maksymov (IonQ), Jason Nguyen (IonQ), Vandiver Chaplin (IonQ), Yunseong Nam (IonQ), and Igor L. Markov (IonQ)

DigiQ: A Scalable Digital Controller for Quantum Computers using SFQ Logic 400
Mohammad Reza Jokar (University of Chicago, USA), Richard Rines (University of Chicago, USA; Super.tech, USA), Ghasem Pasandi (University of Southern California, USA; NVIDIA), Haolin Cong (University of Southern California, USA), Adam Holmes (University of Chicago, USA), Yunong Shi (Amazon Braket, USA), Massoud Pedram (University of Southern California, USA), and Frederic T. Chong (University of Chicago, USA; Super.tech, USA)

HiPerRF: A Dual-Bit Dense Storage SFQ Register File 415
Haipeng Zha (University of Southern California, USA), Naveen Kumar Katam (SeeQC Inc., USA), Massoud Pedram (University of Southern California, USA), and Murali Annavaram (University of Southern California, USA)

Session 4A: Accelerators IV

ReGNN: A Redundancy-Eliminated Graph Neural Networks Accelerator 429
Cen Chen (Shenzhen Institute, Hunan University, China), Kenli Li (Hunan University, China), Yangfan Li (Hunan University, China), and Xiaofeng Zou (Hunan University, China)

LISA: Graph Neural Network Based Portable Mapping on Spatial Accelerators 444
Zhaoying Li (National University of Singapore), Dan Wu (National University of Singapore), Dhananjaya Wijerathne (National University of Singapore), and Tulika Mitra (National University of Singapore)

GCoD: Graph Convolutional Network Acceleration via Dedicated Algorithm and Accelerator Co-Design 460
Haoran You (Rice University), Tong Geng (Pacific Northwest National Laboratory), Yongan Zhang (Rice University), Ang Li (Pacific Northwest National Laboratory), and Yingyan Lin (Rice University)

Atomic Dataflow Based Graph-Level Workload Orchestration for Scalable DNN Accelerators 475
Shixuan Zheng (Tsinghua University, China), Xianjue Zhang (Tsinghua University, China), Leibo Liu (Tsinghua University, China), Shaojun Wei (Tsinghua University, China), and Shouyi Yin (Tsinghua University, China)

Session 4B: Storage, Scheduling, Interfaces

Filesystem Encryption or Direct-Access for NVM Filesystems? Let's Have Both! 490
Kazi Abu Zubair (North Carolina State University, North Carolina), David Mohaisen (University of Central Florida, Florida), and Amro Awad (North Carolina State University, North Carolina)

Efficient Bad Block Management with Cluster Similarity 503
Jui-Nan Yen (National Taiwan University), Yao-Ching Hsieh (National Taiwan University), Cheng-Yu Chen (National Taiwan University), Tseng-Yi Chen (National Central University), Chia-Lin Yang (National Taiwan University), Hsiang-Yun Cheng (Academia Sinica), and Yixin Luo (Carnegie Mellon University)

Using Psychophysics to Guide Power Adaptation for Input Methods on Mobile Architectures 514
Xueliang Li (Shenzhen University, China), Shicong Hong (Shenzhen University, China), Junyang Chen (Shenzhen University, China), Guihai Yan (Chinese Academy of Sciences, China), and Kaishun Wu (Shenzhen University, China)

HD-CPS: Hardware-Assisted Drift-Aware Concurrent Priority Scheduler for Shared Memory Multicores 528
Mohsin Shan (University of Connecticut, Storrs, CT USA) and Omer Khan (University of Connecticut, Storrs, CT USA)

Session 4C: Best Paper Candidates

Improving Locality of Irregular Updates with Hardware Assisted Propagation Blocking 543
Vignesh Balaji (NVIDIA, USA) and Brandon Lucia (Carnegie Mellon University, USA)

Effective Mimicry of Belady's MIN Policy 558
Ishan Shah (The University of Texas at Austin, USA), Akanksha Jain (Google, USA), and Calvin Lin (The University of Texas at Austin, USA)

S2TA: Exploiting Structured Sparsity for Energy-Efficient Mobile CNN Acceleration 573
Zhi-Gang Liu (Arm ML Research), Paul N. Whatmough (Arm ML Research), Yuhao Zhu (University of Rochester), and Matthew Mattina (Arm ML Research)

SupermarQ: A Scalable Quantum Benchmark Suite 587
Teague Tomesh (Princeton University, USA; Super.tech, USA), Pranav Gokhale (Super.tech, USA), Victory Omole (Super.tech, USA), Gokul Subramanian Ravi (University of Chicago, USA), Kaitlin N. Smith (University of Chicago, USA), Joshua Viszlai (University of Chicago, USA), Xin-Chuan Wu (University of Chicago, USA), Nikos Hardavellas (Northwestern University, USA), Margaret R. Martonosi (Princeton University, USA), and Frederic T. Chong (University of Chicago, USA; Super.tech, USA)

Session 5A: Simulation

LoopPoint: Checkpoint-Driven Sampled Simulation for Multi-Threaded Applications 604
Alen Sabu (National University of Singapore), Harish Patil (Intel Corporation), Wim Heirman (Intel Corporation), and Trevor E. Carlson (National University of Singapore)

Compiler-Driven Simulation of Reconfigurable Hardware Accelerators 619
Zhijing Li (Cornell University), Yuwei Ye (Cornell University), Stephen Neuendorffer (Xilinx Inc.), and Adrian Sampson (Cornell University)

NeuroSync: A Scalable and Accurate Brain Simulator Using Safe and Efficient Speculation 633
Hunjun Lee (Seoul National University, South Korea), Chanmyeong Kim (Seoul National University, South Korea), Minseop Kim (Seoul National University, South Korea), Yujin Chung (Seoul National University, South Korea), and Jangwoo Kim (Seoul National University, South Korea)

Session 5B: Cache Hierarchy

Reducing Load Latency with Cache Level Prediction 648
Majid Jalili (The University of Texas at Austin) and Mattan Erez (The University of Texas at Austin)

TCOR: A Tile Cache with Optimal Replacement 662
Diya Joseph (Universitat Politècnica de Catalunya, Spain), Juan L. Aragón (Universidad de Murcia, Spain), Joan-Manuel Parcerisa (Universitat Politècnica de Catalunya, Spain), and Antonio González (Universitat Politècnica de Catalunya, Spain)

Only Buffer When You Need To: Reducing On-Chip GPU Traffic with Reconfigurable Local Atomic Buffers 676
Preyesh Dalmia (University of Wisconsin-Madison), Rohan Mahapatra (University of California, San Diego), and Matthew D. Sinclair (University of Wisconsin Madison, AMD Research)

Session 5C: Quantum III

QuantumNAS: Noise-Adaptive Search for Robust Quantum Circuits 692
Hanrui Wang (Massachusetts Institute of Technology, USA), Yongshan Ding (Yale University, USA), Jiaqi Gu (The University of Texas at Austin, USA), Yujun Lin (Massachusetts Institute of Technology, USA), David Z. Pan (The University of Texas at Austin, USA), Frederic T. Chong (University of Chicago, USA), and Song Han (Massachusetts Institute of Technology, USA)

Not All SWAPs Have the Same Cost: A Case for Optimization-Aware Qubit Routing 709
Ji Liu (North Carolina State University, USA), Peiyi Li (North Carolina State University, USA), and Huiyang Zhou (North Carolina State University, USA)

Q-GPU: A Recipe of Optimizations for Quantum Circuit Simulation using GPUs 726
Yilun Zhao (University of Pittsburgh, USA), Yanan Guo (University of Pittsburgh, USA), Yuan Yao (University of Pittsburgh, USA), Amanda Dumi (University of Pittsburgh, USA), Devin M Mulvey (University of Pittsburgh, USA), Shiv Upadhyay (University of Pittsburgh, USA), Youtao Zhang (University of Pittsburgh, USA), Kenneth D Jordan (University of Pittsburgh, USA), Jun Yang (University of Pittsburgh, USA), and Xulong Tang (University of Pittsburgh, USA)

Session 6A: Synthesis

ScaleHLS: A New Scalable High-Level Synthesis Framework on Multi-Level Intermediate Representation 741
Hanchen Ye (University of Illinois at Urbana-Champaign, USA), Cong Hao (Georgia Institute of Technology, USA), Jianyi Cheng (Imperial College London, UK), Hyunmin Jeong (University of Illinois at Urbana-Champaign, USA), Jack Huang (University of Illinois at Urbana-Champaign, USA), Stephen Neuendorffer (Xilinx Inc., USA), and Deming Chen (University of Illinois at Urbana-Champaign, USA)

HeteroGen: Automatic Synthesis of Heterogeneous Cache Coherence Protocols 756
Nicolai Oswald (The University of Edinburgh), Vijay Nagarajan (The University of Edinburgh), Daniel J. Sorin (Duke University), Vasilis Gavrielatos (The University of Edinburgh), Theo Olausson (The University of Edinburgh), and Reece Carr (The University of Edinburgh)

Session 6B: Traditional Architecture

Reliability-Aware Runahead 772
Ajeya Naithani (Ghent University, Belgium) and Lieven Eeckhout (Ghent University, Belgium)

Adaptable Register File Organization for Vector Processors 786
Cristóbal Ramírez Lazo (Polytechnic University of Catalonia; Barcelona Supercomputing Center, Spain), Enrico Reggiani (Polytechnic University of Catalonia; Barcelona Supercomputing Center, Spain), Carlos Rojas Morales (Barcelona Supercomputing Center, Spain), Roger Figueras Bagué (Barcelona Supercomputing Center, Spain), Luis A. Villa Vargas (Instituto Politécnico Nacional, Mexico), Marco A. Ramírez Salinas (Instituto Politécnico Nacional, Mexico), Mateo Valero Cortés (Polytechnic University of Catalonia; Barcelona Supercomputing Center, Spain), Osman Sabri Ünsal (Barcelona Supercomputing Center, Spain), and Adrián Cristal (Polytechnic University of Catalonia; Barcelona Supercomputing Center, Spain)

Tacker: Tensor-CUDA Core Kernel Fusion for Improving the GPU Utilization While Ensuring QoS 800
Han Zhao (Shanghai Jiao Tong University), Weihao Cui (Shanghai Jiao Tong University), Quan Chen (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Youtao Zhang (University of Pittsburgh), Yanchao Lu (NVIDIA), Chao Li (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Jingwen Leng (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), and Minyi Guo (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute)

Session 7A: Accelerators V

MAGMA: An Optimization Framework for Mapping Multiple DNNs on Multiple Accelerator Cores .. 814
Sheng-Chun Kao (Georgia Institute of Technology) and Tushar Krishna (Georgia Institute of Technology)

SPACX: Silicon Photonics-Based Scalable Chiplet Accelerator for DNN Inference 831
Yuan Li (George Washington University), Ahmed Louri (George Washington University), and Avinash Karanth (Ohio University)

FAST: DNN Training Under Variable Precision Block Floating Point with Stochastic Rounding 846
Sai Qian Zhang (Harvard University), Bradley McDanel (Franklin and Marshall College), and H.T. Kung (Harvard University)

Griffin: Rethinking Sparse Optimization for Deep Learning Architectures 861
Jong Hoon Shin (Samsung Semiconductor Inc., USA), Ali Shafiee (Samsung Semiconductor Inc., USA), Ardavan Pedram (Samsung Semiconductor Inc., USA), Hamzah Abdel-Aziz (Samsung Semiconductor Inc., USA), Ling Li (Samsung Semiconductor Inc., USA), and Joseph Hassoun (Samsung Semiconductor Inc., USA)

CANDLES: Channel-Aware Novel Dataflow-Microarchitecture Co-Design for Low Energy Sparse Neural Network Acceleration 876
Sumanth Gudaparthi (University of Utah, USA), Sarabjeet Singh (University of Utah, USA), Surya Narayanan (University of Utah, USA), Rajeev Balasubramonian (University of Utah, USA), and Visvesh Sathe (University of Washington, USA)

Session 7B: Non-Volatile Memory

ASAP: A Speculative Approach to Persistence 892
Sujay Yadalam (University of Wisconsin-Madison, USA), Nisarg Shah (University of Wisconsin-Madison, USA), Xiangyao Yu (University of Wisconsin-Madison, USA), and Michael Swift (University of Wisconsin-Madison, USA)

Temporal Exposure Reduction Protection for Persistent Memory 908
Yuanchao Xu (North Carolina State University), Chencheng Ye (Huazhong University of Science and Technology), Xipeng Shen (North Carolina State University), and Yan Solihin (University of Central Florida)

MULTI-CLOCK: Dynamic Tiering for Hybrid Memory Systems 925
Adnan Maruf (Florida International University), Ashikee Ghosh (Florida International University), Janki Bhimani (Florida International University), Daniel Campello (Google), Andy Rudoff (Intel Corporation), and Raju Rangaswami (Florida International University)

NVMExplorer: A Framework for Cross-Stack Comparisons of Embedded Non-Volatile Memories . 938
Lillian Pentecost (Harvard University, USA), Alexander Hankin (Tufts University, USA), Marco Donato (Tufts University, USA), Mark Hempstead (Tufts University, USA), Gu-Yeon Wei (Harvard University, USA), and David Brooks (Harvard University, USA)

Session 7C: Network On Chip

Stay in your Lane: A NoC with Low-Overhead Multi-Packet Bypassing 957
Hossein Farrokhbakht (University of Toronto), Paul V. Gratz (Texas A & M), Tushar Krishna (Georgia Institute of Technology), Joshua San Miguel (University of Wisconsin-Madison), and Natalie Enright Jerger (University of Toronto)

FastTrackNoC: A NoC with FastTrack Router Datapaths 971
Ahsen Ejaz (Chalmers University of Technology, Sweden) and Ioannis Sourdis (Chalmers University of Technology, Sweden)

Upward Packet Popup for Deadlock Freedom in Modular Chiplet-Based Systems 986
Yibo Wu (Tsinghua University, China), Liang Wang (Beihang University, China), Xiaohang Wang (South China University of Technology, China), Jie Han (University of Alberta, Canada), Jianfeng Zhu (Tsinghua University, China), Honglan Jiang (Shanghai Jiao Tong University, China), Shouyi Yin (Tsinghua University, China), Shaojun Wei (Tsinghua University, China), and Leibo Liu (Tsinghua University, China)

Saving PAM4 Bus Energy with SMOREs: Sparse Multi-Level Opportunistic Restricted Encodings 1001
Mike O'Connor (NVIDIA, USA), Donghyuk Lee (NVIDIA, USA), Niladrish Chatterjee (NVIDIA, USA), Michael B. Sullivan (NVIDIA, USA), and Stephen W. Keckler (NVIDIA, USA)

Delegated Replies: Alleviating Network Clogging in Heterogeneous Architectures 1014
Xia Zhao (Artificial Intelligence Research Center, Academy of Military Science, China), Lieven Eeckhout (Ghent University, Belgium), and Magnus Jahre (Norwegian University of Science and Technology, Norway)

Session 8A: Accelerators VI

Accelerating Graph Convolutional Networks using Crossbar-Based Processing-In-Memory Architectures 1029
Yu Huang (Huazhong University of Science and Technology, China), Long Zheng (Huazhong University of Science and Technology, China), Pengcheng Yao (Huazhong University of Science and Technology, China), Qinggang Wang (Huazhong University of Science and Technology, China), Xiaofei Liao (Huazhong University of Science and Technology, China), Hai Jin (Huazhong University of Science and Technology, China), and Jingling Xue (UNSW Sydney, Australia)

Enabling High-Quality Uncertainty Quantification in a PIM Designed for Bayesian Neural Network 1043
Xingchen Li (Peking University, China; Advanced Institute of Information Technology), Bingzhe Wu (Peking University, China; Advanced Institute of Information Technology), Guangyu Sun (Peking University, China), Zhe Zhang (Peking University, China), Zhihang Yuan (Peking University, China), Runsheng Wang (Peking University, China), Ru Huang (Peking University, China), Dimin Niu (Alibaba Group Inc., China), Hongzhong Zheng (Alibaba Group Inc., China), Zhichao Lu (Hefei Reliance Memory Ltd., China), Liang Zhao (Hefei Reliance Memory Ltd., China), Meng-Fan Chang (National Tsing Hua University), Tianchan Guan (Alibaba Group Inc., China), and Xin Si (National Tsing Hua University)

RM-SSD: In-Storage Computing for Large-Scale Recommendation Inference 1056
Xuan Sun (City University of Hong Kong), Hu Wan (City University of Hong Kong), Qiao Li (City University of Hong Kong), Chia-Lin Yang (National Taiwan University), Tei-Wei Kuo (City University of Hong Kong; NTU High Performance and Scientific Computing Center, National Taiwan University), and Chun Jason Xue (City University of Hong Kong)

TransPIM: A Memory-Based Acceleration via Software-Hardware Co-Design for Transformer 1071
Minxuan Zhou (University of California, San Diego, USA), Weihong Xu (University of California, San Diego, USA), Jaeyoung Kang (University of California, San Diego, USA), and Tajana Rosing (University of California, San Diego, USA)

PIMCloud: QoS-Aware Resource Management of Latency-Critical Applications in Clouds with Processing-in-Memory 1086
Shuang Chen (Cornell University, USA), Yi Jiang (Cornell University, USA), Christina Delimitrou (Cornell University, USA), and José F. Martínez (Cornell University, USA)

Session 8B: Memory

Exploiting Inter-Block Entropy to Enhance the Compressibility of Blocks with Diverse Data 1100
Jinkwon Kim (KAIST), Mincheol Kang (KAIST), Jeongkyu Hong (Yeungnam University), and Soontae Kim (KAIST)

GBDI: Going Beyond Base-Delta-Immediate Compression with Global Bases 1115
Alexandra Angerd (ZeroPoint Technologies), Angelos Arelakis (ZeroPoint Technologies), Vasilis Spiliopoulos (ZeroPoint Technologies), Erik Sintorn (Chalmers University of Technology), and Per Stenstrom (Chalmers University of Technology)

Virtual Coset Coding for Encrypted Non-Volatile Memories with Multi-Level Cells 1128
Stephen Longofono (University of Pittsburgh, USA), Seyed Mohammad Seyedzadeh (AMD Research, USA), and Alex K. Jones (University of Pittsburgh, USA)

DR-STRaNGe: End-to-End System Design for DRAM-Based True Random Number Generators .. 1141
F. Nisa Bostancı (TOBB University of Economics and Technology; ETH Zürich), Ataberk Olgun (TOBB University of Economics and Technology; ETH Zürich), Lois Orosa (ETH Zürich), A. Giray Yağlıkçı (ETH Zürich), Jeremie S. Kim (ETH Zürich), Hasan Hassan (ETH Zürich), Oğuz Ergin (TOBB University of Economics and Technology), and Onur Mutlu (ETH Zürich)

Mithril: Cooperative Row Hammer Protection on Commodity DRAM Leveraging Managed Refresh ... 1156
Michael Jaemin Kim (Seoul National University, South Korea), Jaehyun Park (Seoul National University, South Korea), Yeonhong Park (Seoul National University, South Korea), Wanju Doh (Seoul National University, South Korea), Namhoon Kim (Seoul National University, South Korea), Tae Jun Ham (Seoul National University, South Korea), Jae W. Lee (Seoul National University, South Korea), and Jung Ho Ahn (Seoul National University, South Korea)

Session 8C: Industrial Session

DarkGates: A Hybrid Power-Gating Architecture to Mitigate the Performance Impact of Dark-Silicon in High Performance Processors 1170
Jawad Haj Yahya (Huawei Technologies - Zurich Research Center; ETH Zurich), Jeremie S. Kim (ETH Zurich), A. Giray Yağlıkçı (ETH Zurich), Jisung Park (ETH Zurich), Efraim Rotem (Intel Corporation), Yanos Sazeides (University of Cyprus), and Onur Mutlu (ETH Zurich)

GPU Subwarp Interleaving 1184
Sana Damani (Georgia Institute of Technology, USA), Mark Stephenson (NVIDIA, USA), Ram Rangan (NVIDIA, India), Daniel Johnson (NVIDIA, USA), Rishkul Kulkarni (NVIDIA, USA), and Stephen W. Keckler (NVIDIA, USA)

Application Defined On-Chip Networks for Heterogeneous Chiplets: An Implementation Perspective 1198
Tianqi Wang (Huawei), Fan Feng (Huawei), Shaolin Xiang (Huawei), Qi Li (Huawei), and Jing Xia (Huawei)

The Specialized High-Performance Network on Anton 3 1211
Keun Sup Shim (D. E. Shaw Research, USA), Brian Greskamp (D. E. Shaw Research, USA), Brian Towles (D. E. Shaw Research, USA), Bruce Edwards (D. E. Shaw Research, USA), J.P. Grossman (D. E. Shaw Research, USA), and David E. Shaw (D. E. Shaw Research, USA)

AI-Enabling Workloads on Large-Scale GPU-Accelerated System: Characterization, Opportunities, and Implications 1224
Baolin Li (Northeastern University), Rohin Arora (Northeastern University), Siddharth Samsi (MIT Lincoln Laboratory), Tirthak Patel (Northeastern University), William Arcand (MIT Lincoln Laboratory), David Bestor (MIT Lincoln Laboratory), Chansup Byun (MIT Lincoln Laboratory), Rohan Basu Roy (Northeastern University), Bill Bergeron (MIT Lincoln Laboratory), John Holodnak (MIT Lincoln Laboratory), Michael Houle (MIT Lincoln Laboratory), Matthew Hubbell (MIT Lincoln Laboratory), Michael Jones (MIT Lincoln Laboratory), Jeremy Kepner (MIT Lincoln Laboratory), Anna Klein (MIT Lincoln Laboratory), Peter Michaleas (MIT Lincoln Laboratory), Joseph McDonald (MIT Lincoln Laboratory), Lauren Milechin (MIT Lincoln Laboratory), Julie Mullen (MIT Lincoln Laboratory), Andrew Prout (MIT Lincoln Laboratory), Benjamin Price (MIT Lincoln Laboratory), Albert Reuther (MIT Lincoln Laboratory), Antonio Rosa (MIT Lincoln Laboratory), Matthew Weiss (MIT Lincoln Laboratory), Charles Yee (MIT Lincoln Laboratory), Daniel Edelman (MIT Lincoln Laboratory), Allan Vanterpool (US Air Force), Anson Cheng (US Air Force), Vijay Gadepally (MIT Lincoln Laboratory), and Devesh Tiwari (Northeastern University)

Author Index