2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA 2024)

Buenos Aires, Argentina
29 June - 3 July 2024

Pages 1-659

IEEE

IEEE Catalog Number:	**CFP24030-POD**
ISBN:	**979-8-3503-2659-8**

1/2

Copyright © 2024 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:	CFP24030-POD
ISBN (Print-On-Demand):	979-8-3503-2659-8
ISBN (Online):	979-8-3503-2658-1

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

2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)

ISCA 2024

Table of Contents

Message from the General Chairs xx
Message from the Program Chairs xxii
Organizing Committee xxvii
Program Committee xxix
Message from the Industry Track Program Chair xxxiv
Sponsors xxxvi

Session 1A: Microarchitectire

GhOST: A GPU Out-of-Order Scheduling Technique for Stall Reduction 1
Ishita Chaturvedi (Princeton University, USA), Bhargav Reddy Godala (Princeton University, USA), Yucan Wu (Princeton University, USA), Ziyang Xu (Princeton University, USA), Konstantinos Iliakis (National Technical University of Athens, Greece), Panagiotis-Eleftherios Eleftherakis (National Technical University of Athens, Greece), Sotirios Xydis (National Technical University of Athens, Greece), Dimitrios Soudris (National Technical University of Athens, Greece), Tyler Sorensen (UC Santa Cruz, USA), Simone Campanoni (Northwestern University, USA), Tor M. Aamodt (University of British Columbia, Canada), and David I. August (Princeton University, USA)

AVM-BTB: Adaptive and Virtualized Multi-Level Branch Target Buffer 17
Yunzhe Liu (State Key Lab of Processors, Insititute of Computing Technology, CAS; University of Chinese Academy of Sciences), Xinyu Li (State Key Lab of Processors, Insititute of Computing Technology, CAS; University of Chinese Academy of Sciences), Tingting Zhang (Loongson Technology Co. Ltd.; Institute of Computing Technology, CAS), Tianyi Liu (The University of Texas at San Antonio), Qi Guo (State Key Lab of Processors, Insititute of Computing Technology, CAS), Fuxin Zhang (State Key Lab of Processors, Insititute of Computing Technology, CAS; University of Chinese Academy of Sciences), and Jian Wang (State Key Lab of Processors, Insititute of Computing Technology, CAS; University of Chinese Academy of Sciences)

The Maya Cache: A Storage-Efficient and Secure Fully-Associative Last-Level Cache 32
Anubhav Bhatla (Indian Institute of Technology Bombay, India), Navneet Navneet (Indian Institute of Technology Bombay, India), and Biswabandan Panda (Indian Institute of Technology Bombay, India)

Session 1B: Emerging Technologies

DS-GL: Advancing Graph Learning via Harnessing Nature's Power Within Scalable Dynamical Systems 45
Ruibing Song (University of Rochester, USA), Chunshu Wu (University of Rochester, USA), Chuan Liu (University of Rochester, USA), Ang Li (Pacific Northwest National Laboratory, USA), Michael Huang (University of Rochester, USA), and Tony (Tong) Geng (University of Rochester, USA)

ReAIM: A ReRAM-Based Adaptive Ising Machine for Solving Combinatorial Optimization Problems 58
Hao-Wei Chiang (National Taiwan University, Taiwan), Chin-Fu Nien (Chang Gung University, Taiwan), Hsiang-Yun Cheng (Center for Information Technology Innovation, Taiwan), and Kuei-Po Huang (Chang Gung University, Taiwan)

Mirage: An RNS-Based Photonic Accelerator for DNN Training 73
Cansu Demirkiran (Boston University, USA), Guowei Yang (Boston University, USA), Darius Bunandar (Lightmatter, USA), and Ajay Joshi (Boston University, USA)

Session 2: Best Paper Session

Constable: Improving Performance and Power Efficiency by Safely Eliminating Load Instruction Execution 88
Rahul Bera (ETH Zurich), Adithya Ranganathan (Processor Architecture Research Lab, Intel), Joydeep Rakshit (Processor Architecture Research Lab, Intel), Sujit Mahto (Processor Architecture Research Lab, Intel), Anant V. Nori (Processor Architecture Research Lab, Intel), Jayesh Gaur (Processor Architecture Research Lab, Intel), Ataberk Olgun (ETH Zurich), Konstantinos Kanellopoulos (ETH Zurich), Mohammad Sadrosadati (ETH Zurich), Sreenivas Subramoney (Processor Architecture Research Lab, Intel), and Onur Mutlu (ETH Zurich)

QuTracer: Mitigating Quantum Gate and Measurement Errors by Tracing Subsets of Qubits 103
Peiyi Li (North Carolina State University, USA), Ji Liu (Argonne National Laboratory, USA), Alvin Gonzales (Argonne National Laboratory, USA), Zain Hamid Saleem (Argonne National Laboratory, USA), Huiyang Zhou (North Carolina State University, USA), and Paul Hovland (Argonne National Laboratory, USA)

Splitwise: Efficient Generative LLM Inference Using Phase Splitting 118
Pratyush Patel (University of Washington, USA), Esha Choukse (Microsoft, USA), Chaojie Zhang (Microsoft, USA), Aashaka Shah (Microsoft, USA), Íñigo Goiri (Microsoft, USA), Saeed Maleki (Microsoft, USA), and Ricardo Bianchini (Microsoft, USA)

HiFi-DRAM: Enabling High-Fidelity DRAM Research by Uncovering Sense Amplifiers with IC Imaging 133
Michele Marazzi (ETH Zurich, Switzerland), Tristan Sachsenweger (ETH Zurich, Switzerland), Flavien Solt (ETH Zurich, Switzerland), Peng Zeng (ETH Zurich, Switzerland), Kubo Takashi (Zentel Japan, Japan), Maksym Yarema (ETH Zurich, Switzerland), and Kaveh Razavi (ETH Zurich, Switzerland)

Mind the Gap: Attainable Data Movement and Operational Intensity Bounds for Tensor Algorithms 150
Qijing Huang (NVIDIA), Po-An Tsai (NVIDIA), Joel S. Emer (MIT CSAIL / NVIDIA), and Angshuman Parashar (NVIDIA)

A Tale of Two Domains: Exploring Efficient Architecture Design for Truly Autonomous Things.... 167
Xiaofeng Hou (Shanghai Jiao Tong University, China), Tongqiao Xu (Shanghai Jiao Tong University, China), Chao Li (Shanghai Jiao Tong University, China), Cheng Xu (Shanghai Jiao Tong University, China), Jiacheng Liu (Chinese University of Hong Kong, China), Yang Hu (Tsinghua University, China), Jieru Zhao (Shanghai Jiao Tong University, China), Jingwen Leng (Shanghai Jiao Tong University, China), Kwang-Ting Cheng (Hong Kong University of Science and Technology, China), and Minyi Guo (Shanghai Jiao Tong University, China)

Session 3A: Networking

Determining the Minimum Number of Virtual Networks for Different Coherence Protocols 182
Weihang Li (Duke University), Andres Goens (University of Amsterdam), Nicolai Oswald (NVIDIA), Vijay Nagarajan (University of Utah), and Daniel J. Sorin (Duke University)

FEATHER: A Reconfigurable Accelerator with Data Reordering Support for Low-Cost On-Chip Dataflow Switching 198
Jianming Tong (Georgia Institute of Technology, USA), Anirudh Itagi (Georgia Institute of Technology, USA), Prasanth Chatarasi (IBM Research, USA), and Tushar Krishna (Georgia Institute of Technology, USA)

Waferscale Network Switches 215
Shuangliang Chen (UIUC), Saptadeep Pal (Etched AI), and Rakesh Kumar (UIUC)

The Case For Data Centre Hyperloops 230
Guillem López-Paradís (Barcelona Supercomputing Center, Universitat Politècnica de Catalunya, Spain), Isaac M. Hair (UC Santa Barbara, USA), Sid Kannan (UC Santa Barbara, USA), Roman Rabbat (UC Santa Barbara, USA), Parker Murray (UC Santa Barbara, USA), Alex Lopes (UC Santa Barbara, USA), Rory Zahedi (UC Santa Barbara, USA), Winston Zuo (UC Santa Barbara, USA), and Jonathan Balkind (UC Santa Barbara, USA)

PID-Comm: A Fast and Flexible Collective Communication Framework for Commodity Processing-in-DIMM Devices 245
Si Ung Noh (Seoul National University, Republic of Korea), Junguk Hong (Seoul National University, Republic of Korea), Chaemin Lim (Yonsei University, Republic of Korea), Seongyeon Park (Seoul National University, Republic of Korea), Jeehyun Kim (Yonsei University, Republic of Korea), Hanjun Kim (Yonsei University, Republic of Korea), Youngsok Kim (Yonsei University, Republic of Korea), and Jinho Lee (Seoul National University, Republic of Korea)

Session 3B: Quantum Computing

Bosehedral: Compiler Optimization for Bosonic Quantum Computing 261
Junyu Zhou (University of Pennsylvania, USA), Yuhao Liu (University of Pennsylvania, USA), Yunong Shi (AWS Quantum Technology, USA), Ali Javadi-Abhari (IBM Quantum, USA), and Gushu Li (University of Pennsylvania, USA)

Tetris: A Compilation Framework for VQA Applications in Quantum Computing 277
Yuwei Jin (Rutgers University, USA), Zirui Li (Rutgers University, USA), Fei Hua (Rutgers University, USA), Tianyi Hao (University of Wisconsin-Madison, USA), Huiyang Zhou (North Carolina State University, USA), Yipeng Huang (Rutgers University, USA), and Eddy Z. Zhang (Rutgers University, USA)

Atomique: A Quantum Compiler for Reconfigurable Neutral Atom Arrays 293
Hanrui Wang (Massachusetts Institute of Technology, US), Pengyu Liu (Carnegie Mellon University, US), Daniel Bochen Tan (University of California, Los Angeles, US), Yilian Liu (Cornell University, US), Jiaqi Gu (Arizona State University, US), David Z. Pan (The University of Texas at Austin, US), Jason Cong (University of California, Los Angeles, US), Umut A. Acar (Carnegie Mellon University, US), and Song Han (Massachusetts Institute of Technology, US)

Suppressing Correlated Noise in Quantum Computers via Context-Aware Compiling 310
Alireza Seif (IBM Quantum, USA), Haoran Liao (University of California, Berkeley, USA), Vinay Tripathi (University of Southern California, USA), Kevin Krsulich (IBM Quantum, USA), Moein Malekakhlagh (IBM Quantum, USA), Mirko Amico (IBM Quantum, USA), Petar Jurcevic (IBM Quantum, USA), and Ali Javadi-Abhari (IBM Quantum, USA)

A SAT Scalpel for Lattice Surgery: Representation and Synthesis of Subroutines for Surface-Code Fault-Tolerant Quantum Computing 325
Daniel Bochen Tan (Google Quantum AI; University of California, Los Angeles), Murphy Yuezhen Niu (Google Quantum AI; University of California, Santa Barbara), and Craig Gidney (Google Quantum AI)

Session 4A: PIM Accelerators

PreSto: An In-Storage Data Preprocessing System for Training Recommendation Models 340
Yunjae Lee (KAIST, South Korea), Hyeseong Kim (KAIST, South Korea), and Minsoo Rhu (KAIST, South Korea)

pSyncPIM: Partially Synchronous Execution of Sparse Matrix Operations for All-Bank PIM Architectures 354
Daehyeon Baek (Samsung SDS, Republic of Korea), Soojin Hwang (KAIST, Republic of Korea), and Jaehyuk Huh (KAIST, Republic of Korea)

NDSEARCH: Accelerating Graph-Traversal-Based Approximate Nearest Neighbor Search through Near Data Processing 368
Yitu Wang (Duke University, USA), Shiyu Li (Duke University, USA), Qilin Zheng (Duke University, USA), Linghao Song (University of California, Los Angeles, USA), Zongwang Li (Samsung Semiconductor, Inc., USA), Andrew Chang (Samsung Semiconductor, Inc., USA), Hai Li (Duke University, USA), and Yiran Chen (Duke University, USA)

Enabling Efficient Large Recommendation Model Training with Near CXL Memory Processing 382
Haifeng Liu (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; Zhejiang Lab, China), Jingyi Zhou (Huazhong University of Science and Technology, China), Chaoqiang Liu (Huazhong University of Science and Technology, China), Runze 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 (University of New South Wales, Australia)

Exploiting Similarity Opportunities of Emerging Vision AI Models on Hybrid Bonding Architecture 396
Zhiheng Yue (Tsinghua University, China), Huizheng Wang (Tsinghua University, China), Jiahao Fang (Tsinghua University, China), Jinyi Deng (Tsinghua University, China), Guangyang Lu (Tsinghua university, China), Fengbin Tu (The Hong Kong University of Science and Technology, China), Ruiqi Guo (Tsinghua university, China), Yuxuan Li (Tsinghua University, China), Yubin Qin (Tsinghua University, China), Yang Wang (Tsinghua University, China), Chao Li (Shanghai Jiao Tong University), Huiming Han (Tsinghua University, China), Shaojun Wei (Tsinghua University, China), Yang Hu (Tsinghua University, China), and Shouyi Yin (Tsinghua University, China)

Session 4B: Cloud Technologies

ElasticRec: A Microservice-Based Model Serving Architecture Enabling Elastic Resource Scaling for Recommendation Models 410
Yujeong Choi (KAIST), Jiin Kim (KAIST), and Minsoo Rhu (KAIST)

Derm: SLA-Aware Resource Management for Highly Dynamic Microservices 424
Liao Chen (University of Macau), Shutian Luo (Yale University), Chenyu Lin (University of Macau, Macau SAR, China), Zizhao Mo (University of Macau, Macau SAR, China), Huanle Xu (University of Macau, Macau SAR, China), Kejiang Ye (Shenzhen Institute of Advanced Technology), and Chengzhong Xu (University of Macau, Macau SAR, China)

SmartOClock: Workload- and Risk-Aware Overclocking in the Cloud 437
Jovan Stojkovic (University of Illinois at Urbana-Champaign), Pulkit A. Misra (Microsoft), Íñigo Goiri (Microsoft), Sam Whitlock (Microsoft), Esha Choukse (Microsoft), Mayukh Das (Microsoft), Chetan Bansal (Microsoft), Jason Lee (Microsoft), Zoey Sun (Microsoft), Haoran Qiu (University of Illinois at Urbana-Champaign), Reed Zimmermann (University of Texas at Austin), Savyasachi Samal (Microsoft), Brijesh Warrier (Microsoft), Ashish Raniwala (Microsoft), and Ricardo Bianchini (Microsoft)

Designing Cloud Servers for Lower Carbon 452
Jaylen Wang (Carnegie Mellon University), Daniel S. Berger (Microsoft and University of Washington), Fiodar Kazhamiaka (Microsoft), Celine Irvene (Microsoft), Chaojie Zhang (Microsoft), Esha Choukse (Microsoft), Kali Frost (Microsoft), Rodrigo Fonseca (Microsoft), Brijesh Warrier (Microsoft), Chetan Bansal (Microsoft), Jonathan Stern (Microsoft), Ricardo Bianchini (Microsoft), and Akshitha Sriraman (Carnegie Mellon University)

EcoFaaS: Rethinking the Design of Serverless Environments for Energy Efficiency 471
Jovan Stojkovic (University of Illinois at Urbana-Champaign), Nikoleta Iliakopoulou (University of Illinois at Urbana-Champaign), Tianyin Xu (University of Illinois at Urbana-Champaign), Hubertus Franke (IBM Research), and Josep Torrellas (University of Illinois at Urbana-Champaign)

Session 5A: Tools and Analysis

AIO: An Abstraction for Performance Analysis Across Diverse Accelerator Architectures 487
Joseph Rogers (Norwegian University of Science and Technology (NTNU), Norway), Taha Soliman (Robert Bosch GmbH Corporate Research, Germany), and Magnus Jahre (Norwegian University of Science and Technology (NTNU), Norway)

FireAxe: Partitioned FPGA-Accelerated Simulation of Large-Scale RTL Designs 501
Joonho Whangbo (UC Berkeley), Edwin Lim (CMU), Chengyi Lux Zhang (UC Berkeley), Kevin Anderson (UC Berkeley), Abraham Gonzalez (UC Berkeley), Raghav Gupta (UC Berkeley), Nivedha Krishnakumar (UC Berkeley), Sagar Karandikar (UC Berkeley), Borivoje Nikolic (UC Berkeley), Yakun Sophia Shao (UC Berkeley), and Krste Asanovic (UC Berkeley)

Harpocrates: Breaking the Silence of CPU Faults through Hardware-in-the-Loop Program Generation 516
Nikos Karystinos (University of Athens, Greece), Odysseas Chatzopoulos (University of Athens, Greece), George-Marios Fragkoulis (University of Athens, Greece), George Papadimitriou (University of Athens, Greece), Dimitris Gizopoulos (University of Athens, Greece), and Sudhanva Gurumurthi (Advanced Micro Devices, Inc, USA)

The Dataflow Abstract Machine Simulator Framework 532
Nathan Zhang (Stanford University, USA), Rubens Lacouture (Stanford University, USA), Gina Sohn (Stanford University, USA), Paul Mure (Massachusetts Institute of Technology, USA), Qizheng Zhang (Stanford University, USA), Fredrik Kjolstad (Stanford University, USA), and Kunle Olukotun (Stanford University, USA)

Session 5B: Accelerators for Emerging Workloads I

Tartan: Microarchitecting a Robotic Processor 548
Mohammad Bakhshalipour (Carnegie Mellon University, USA) and Phillip B. Gibbons (Carnegie Mellon University, USA)

Collision Prediction for Robotics Accelerators 566
Deval Shah (University of British Columbia, Canada) and Tor M. Aamodt (University of British Columbia, Canada)

BLESS: Bandwidth and Locality Enhanced SMEM Seeding Acceleration for DNA Sequencing 582
Seunghee Han (KAIST, South Korea), Seungjae Moon (KAIST, South Korea), Teokkyu Suh (KAIST, South Korea), Jaehoon Heo (KAIST, South Korea), and Joo-Young Kim (KAIST, South Korea)

QUETZAL: Vector Acceleration Framework for Modern Genome Sequence Analysis Algorithms .. 597
Julian Pavon (Barcelona Supercomputing Center, Spain), Ivan Vargas Valdivieso (Barcelona Supercomputing Center), Carlos Rojas (Barcelona Supercomputing Center, Spain), Cesar Hernandez (Barcelona Supercomputing Center, Spain), Mehmet Aslan (TOBB ETU University of Economics & Technology, Turkey), Roger Figueras (Barcelona Supercomputing Center, Spain), Yichao Yuan (University of Michigan, USA), Joel Lindegger (ETH Zurich, Switzerland), Mohammed Alser (ETH Zurich, Switzerland), Francesc Moll (Barcelona Supercomputing Center, Spain), Santiago Marco-Sola (Barcelona Supercomputing Center, Spain), Oguz Ergin (TOBB ETU University of Economics & Technology, Turkey), Nishil Talati (University of Michigan, USA), Onur Mutlu (ETH Zurich, Switzerland), Osman Unsal (Barcelona Supercomputing Center, Spain), Mateo Valero (Barcelona Supercomputing Center, Spain), and Adrian Cristal (Barcelona Supercomputing Center, Spain)

Session 6A: NDP Technologies

HAL: Hardware-Assisted Load Balancing for Energy-Efficient SNIC-Host Cooperative Computing..... 613
Jinghan Huang (University of Illinois Urbana-Champaign, USA), Jiaqi Lou (University of Illinois Urbana-Champaign, USA), Srikar Vanavasam (University of Illinois Urbana-Champaign, USA), Xinhao Kong (Duke University, USA), Houxiang Ji (University of Illinois Urbana-Champaign, USA), Ipoom Jeong (University of Illinois Urbana-Champaign, USA), Danyang Zhuo (Duke University, USA), Eun Kyung Lee (IBM Research, USA), and Nam Sung Kim (University of Illinois Urbana-Champaign, USA)

NDPBridge: Enabling Cross-Bank Coordination in Near-DRAM-Bank Processing Architectures 628
Boyu Tian (Tsinghua University, China), Yiwei Li (Tsinghua University, China), Li Jiang (Shanghai Jiao Tong University, China; Shanghai Qi Zhi Institute, China; Huawei Technologies Co., Ltd., China), Shuangyu Cai (Tsinghua University, China), and Mingyu Gao (Tsinghua University, China; Shanghai Qi Zhi Institute, China)

UM-PIM: DRAM-Based PIM with Uniform & Shared Memory Space 644

Yilong Zhao (Shanghai Jiao Tong University, Shanghai Qi Zhi Institute), Mingyu Gao (Tsinghua University, Shanghai Qi Zhi Institute), Fangxin Liu (Shanghai Jiao Tong University, Shanghai Qi Zhi Institute), Yiwei Hu (Shanghai Jiao Tong University), Zongwu Wang (Shanghai Jiao Tong University, Shanghai Qi Zhi Institute), Han Lin (Huawei Technologies Co. Ltd.), Ji Li (Huawei Technologies Co. Ltd.), He Xian (Shanghai Qi Zhi Institute), Hanlin Dong (Shanghai Qi Zhi Institute), Tao Yang (Shanghai Jiao Tong University), Naifeng Jing (Shanghai Jiao Tong University), Xiaoyao Liang (Shanghai Jiao Tong University), and Li Jiang (Shanghai Jiao Tong University, Shanghai Qi Zhi Institute, Huawei Technologies Co. Ltd.)

MegIS: High-Performance, Energy-Efficient, and Low-Cost Metagenomic Analysis with In-Storage Processing 660

Nika Mansouri Ghiasi (ETH Zurich, Switzerland), Mohammad Sadrosadati (ETH Zurich, Switzerland), Harun Mustafa (ETH Zurich, Switzerland), Arvid Gollwitzer (ETH Zurich, Switzerland), Can Firtina (ETH Zurich, Switzerland), Julien Eudine (ETH Zurich, Switzerland), Haiyu Mao (ETH Zurich, Switzerland), Joël Lindegger (ETH Zurich, Switzerland), Meryem Banu Cavlak (ETH Zurich, Switzerland), Mohammed Alser (ETH Zurich, Switzerland), Jisung Park (POSTECH, Republic of Korea), and Onur Mutlu (ETH Zurich, Switzerland)

On Error Correction for Nonvolatile Processing-In-Memory 678

Husrev Cilasun (University of Minnesota, Twin Cities), Salonik Resch (University of Minnesota, Twin Cities), Zamshed I. Chowdhury (University of Minnesota, Twin Cities), Masoud Zabihi (University of Minnesota, Twin Cities), Yang Lv (University of Minnesota, Twin Cities), Brandon Zink (University of Minnesota, Twin Cities), Jian-Ping Wang (University of Minnesota, Twin Cities), Sachin S. Sapatnekar (University of Minnesota, Twin Cities), and Ulya R. Karpuzcu (University of Minnesota, Twin Cities)

Session 6B: Security

MetaLeak: Uncovering Side Channels in Secure Processor Architectures Exploiting Metadata 693

Hafizul Islam Chowdhuryy (University of Central Florida, USA), Hao Zheng (University of Central Florida, USA), and Fan Yao (University of Central Florida, USA)

sNPU: Trusted Execution Environments on Integrated NPUs 708

Erhu Feng (Shanghai Jiao Tong University, China; Ministry of Education, China), Dahu Feng (Tsinghua university), Dong Du (Shanghai Jiao Tong University, China; Ministry of Education, China), Yubin Xia (Shanghai Jiao Tong University, China; Ministry of Education, China), and Haibo Chen (Shanghai Jiao Tong University, China; Chinese Academy of Sciences, China)

Counter-Light Memory Encryption 724

Xin Wang (Virginia Tech), Jagadish Kotra (AMD), Alex Jones (University of Pittsburgh), Wenjie Xiong (Virginia Tech), and Xun Jian (Virginia Tech)

Perspective: A Principled Framework for Pliable and Secure Speculation in Operating Systems 739
Tae Hoon Kim (Carnegie Mellon University, USA), David Rudo (Carnegie Mellon University, USA), Kaiyang Zhao (Carnegie Mellon University, USA), Zirui Neil Zhao (University of Illinois Urbana-Champaign, USA), and Dimitrios Skarlatos (Carnegie Mellon University, USA)

HEAP: A Fully Homomorphic Encryption Accelerator with Parallelized Bootstrapping 756
Rashmi Agrawal (Boston University, United States), Anantha Chandrakasan (MIT, United States), and Ajay Joshi (Boston University, United States)

Session 6C: Parallel Architectures

Scalable, Programmable and Dense: The HammerBlade Open-Source RISC-V Manycore 770
Dai Cheol Jung (University of Washington), Max Ruttenberg (University of Washington), Paul Gao (University of Washington), Scott Davidson (University of Washington), Daniel Petrisko (University of Washington), Kangli Li (University of Washington), Aditya K Kamath (University of Washington), Lin Cheng (Cornell University), Shaolin Xie (University of Washington), Peitian Pan (Cornell University), Zhongyuan Zhao (Cornell University), Zichao Yue (Cornell University), Bandhav Veluri (University of Washington), Sripathi Muralitharan (University of Washington), Adrian Sampson (Cornell University), Andrew Lumsdaine (University of Washington; PNNL), Zhiru Zhang (Cornell University), Christopher Batten (Cornell University), Mark Oskin (University of Washington), Dustin Richmond (University of California, Santa Cruz), and Michael Taylor (University of Washington)

HADES: Hardware-Assisted Distributed Transactions in the Age of Fast Networks and SmartNICs 785
Apostolos Kokolis (University of Illinois Urbana-Champaign, USA), Antonis Psistakis (University of Illinois Urbana-Champaign, USA), Benjamin Reidys (University of Illinois Urbana-Champaign, USA), Jian Huang (University of Illinois Urbana-Champaign, USA), and Josep Torrellas (University of Illinois Urbana-Champaign, USA)

BlitzCoin: Fully Decentralized Hardware Power Management for Accelerator-Rich SoCs 801
Martin Cochet (IBM Research, USA), Karthik Swaminathan (IBM Research, USA), Erik Jens Loscalzo (Columbia University, USA), Joseph Zuckerman (Columbia University, USA), Maico Cassel dos Santos (Columbia University, USA), Davide Giri (Columbia University, USA), Alper Buyuktosunoglu (IBM Research, USA), Tianyu Jia (Harvard University, USA), David Brooks (Harvard University, USA), Gu-Yeon Wei (Harvard University, USA), Kenneth Shepard (Columbia University, USA), Luca P. Carloni (Columbia University, USA), and Pradip Bose (IBM Research, USA)

MAD-Max Beyond Single-Node: Enabling Large Machine Learning Model Acceleration on Distributed Systems 818
Samuel Hsia (FAIR At Meta; Harvard University, United States of America), Alicia Golden (Harvard University, United States of America), Bilge Acun (FAIR at Meta), Newsha Ardalani (FAIR at Meta), Zachary DeVito (FAIR at Meta), Gu-Yeon Wei (Harvard University, United States of America), David Brooks (Harvard University, United States of America), and Carole-Jean Wu (FAIR at Meta)

Barre Chord: Efficient Virtual Memory Translation for Multi-Chip-Module GPUs 834
Yuan Feng (University of California, Merced), Seonjin Na (Georgia Institute of Technology), Hyesoon Kim (Georgia Institute of Technology), and Hyeran Jeon (University of California, Merced)

Session 7: Industry Session

Intel Accelerators Ecosystem: An SoC-Oriented Perspective 848
Yifan Yuan (Intel Labs), Ren Wang (Intel Labs), Narayan Ranganathan (Intel Labs), Nikhil Rao (Intel Labs), Sanjay Kumar (Intel Labs), Philip Lantz (Intel Labs), Vivekananthan Sanjeepan (Intel Labs), Jorge Cabrera (Intel Labs), Atul Kwatra (Intel Labs), Rajesh Sankaran (Intel Labs), Ipoom Jeong (University of Illinois at Urbana-Champaign), and Nam Sung Kim (University of Illinois at Urbana-Champaign)

Circular Reconfigurable Parallel Processor for Edge Computing 863
Yuan Li (AzurEngine Technologies), Jianbin Zhu (AzurEngine Technologies), Yao Fu (AzurEngine Technologies), Yu Lei (AzurEngine Technologies), Toshio Nagata (AzurEngine Technologies), Ryan Braidwood (AzurEngine Technologies), Haohuan Fu (Tsinghua University), Juepeng Zheng (Sun Yat-Sen University), Wayne Luk (Imperial College London), and Hongxiang Fan (Imperial College London and University of Cambridge)

Realizing the AMD Exascale Heterogeneous Processor Vision 876
Alan Smith (Advanced Micro Devices, Inc.), Gabriel H. Loh (Advanced Micro Devices, Inc.), Michael J. Schulte (Advanced Micro Devices, Inc.), Mike Ignatowski (Advanced Micro Devices, Inc.), Samuel Naffziger (Advanced Micro Devices, Inc.), Mike Mantor (Advanced Micro Devices, Inc.), Mark Fowler (Advanced Micro Devices, Inc.), Nathan Kalyanasundharam (Advanced Micro Devices, Inc.), Vamsi Alla (Advanced Micro Devices, Inc.), Nicholas Malaya (Advanced Micro Devices, Inc.), Joseph L. Greathouse (Advanced Micro Devices, Inc.), Eric Chapman (Advanced Micro Devices, Inc.), and Raja Swaminathan (Advanced Micro Devices, Inc.)

TCP: A Tensor Contraction Processor for AI Workloads 890
Hanjoon Kim (Furiosa AI, Inc.), Younggeun Choi (Furiosa AI, Inc.), Junyoung Park (Furiosa AI, Inc.), Byeongwook Bae (Furiosa AI, Inc.), Hyunmin Jeong (Furiosa AI, Inc.), Sang Min Lee (Furiosa AI, Inc.), Jeseung Yeon (Furiosa AI, Inc.), Minho Kim (Furiosa AI, Inc.), Changjae Park (Furiosa AI, Inc.), Boncheol Gu (Furiosa AI, Inc.), Changman Lee (Furiosa AI, Inc.), Jaeick Bae (Furiosa AI, Inc.), SungGyeong Bae (Furiosa AI, Inc.), Yojung Cha (Furiosa AI, Inc.), Wooyoung Choe (Furiosa AI, Inc.), Jonguk Choi (Furiosa AI, Inc.), Juho Ha (Furiosa AI, Inc.), Hyuck Han (Furiosa AI, Inc.), Namoh Hwang (Furiosa AI, Inc.), Seokha Hwang (Furiosa AI, Inc.), Kiseok Jang (Furiosa AI, Inc.), Haechan Je (Furiosa AI, Inc.), Hojin Jeon (Furiosa AI, Inc.), Jaewoo Jeon (Furiosa AI, Inc.), Hyunjun Jeong (Furiosa AI, Inc.), Yeonsu Jung (Furiosa AI, Inc.), Dongok Kang (Furiosa AI, Inc.), Hyewon Kim (Furiosa AI, Inc.), Minjae Kim (Furiosa AI, Inc.), Muhwan Kim (Furiosa AI, Inc.), Sewon Kim (Furiosa AI, Inc.), Suhyung Kim (Furiosa AI, Inc.), Won Kim (Furiosa AI, Inc.), Yong Kim (Furiosa AI, Inc.), Youngsik Kim (Furiosa AI, Inc.), Younki Ku (Furiosa AI, Inc.), Jeong Ki Lee (Furiosa AI, Inc.), Juyun Lee (Furiosa AI, Inc.), Kyungjae Lee (Furiosa AI, Inc.), Seokho Lee (Furiosa AI, Inc.), Minwoo Noh (Furiosa AI, Inc.), Hyuntaek Oh (Furiosa AI, Inc.), Gyunghee Park (Furiosa AI, Inc.), Sanguk Park (Furiosa AI, Inc.), Jimin Seo (Furiosa AI, Inc.), Jungyoung Seong (Furiosa AI, Inc.), June Paik (Furiosa AI, Inc.), Nuno P. Lopes (INESC-ID / Instituto Superior Técnico, University of Lisbon), and Sungjoo Yoo (Seoul National University)

Session 8A: Machine Learning Accelerators I

Cambricon-D: Full-Network Differential Acceleration for Diffusion Models 903
Weihao Kong (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Yifan Hao (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Qi Guo (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Yongwei Zhao (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Xinkai Song (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Xiaqing Li (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Mo Zou (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Zidong Du (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Rui Zhang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Chang Liu (Cambricon Technologies, China), Yuanbo Wen (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Pengwei Jin (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Xing Hu (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Wei Li (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), Zhiwei Xu (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China), and Tianshi Chen (Cambricon Technologies)

Flagger: Cooperative Acceleration for Large-Scale Cross-Silo Federated Learning Aggregation 915
Xiurui Pan (Peking University, China), Yuda An (Peking University, China), Shengwen Liang (Institute of Computing Technology, Chinese Academy of Sciences, China), Bo Mao (Xiamen University, China), Mingzhe Zhang (Institute of Information Engineering, Chinese Academy of Sciences, China), Qiao Li (Xiamen University, China), Myoungsoo Jung (KAIST and Panmnesia, South Korea), and Jie Zhang (Peking University, China)

Trapezoid: A Versatile Accelerator for Dense and Sparse Matrix Multiplications 931
Yifan Yang (MIT CSAIL), Joel Emer (MIT CSAIL / NVIDIA), and Daniel Sanchez (MIT CSAIL)

NeuraChip: Accelerating GNN Computations with a Hash-Based Decoupled Spatial Accelerator .. 946
Kaustubh Shivdikar (Northeastern University, USA), Nicolas Bohm Agostini (Northeastern University, USA), Malith Jayaweera (Northeastern University, USA), Gilbert Jonatan (KAIST, South Korea), José L. Abellán (University of Murcia, Spain), Ajay Joshi (Boston University, USA), John Kim (KAIST, South Korea), and David Kaeli (Northeastern University, USA)

Session 8B: Compliers and Programming Models

Compiler-Directed Whole-System Persistence 961
Jianping Zeng (Purdue University), Tong Zhang (Samsung Electronics), and Changhee Jung (Purdue University)

Memento: An Adaptive, Compiler-Assisted Register File Cache for GPUs 978
Mojtaba Abaie Shoushtary (Polytechnic University of Catalonia, Spain), Jose Maria Arnau (Polytechnic University of Catalonia, Spain), Jordi Tubella Murgadas (Polytechnic University of Catalonia, Spain), and Antonio Gonzalez (Polytechnic University of Catalonia, Spain)

Soter: Analytical Tensor-Architecture Modeling and Automatic Tensor Program Tuning for Spatial Accelerators 991
Fuyu Wang (Sun Yat-sen University, China), Minghua Shen (Sun Yat-sen University, China), Yufei Ding (University of California, USA), and Nong Xiao (Sun Yat-sen University, China)

ALISA: Accelerating Large Language Model Inference via Sparsity-Aware KV Caching 1005
Youpeng Zhao (University of Central Florida), Di Wu (University of Central Florida), and Jun Wang (University of Central Florida)

Session 9A: Machine Learning Accelerators II

Pre-Gated MoE: An Algorithm-System Co-Design for Fast and Scalable Mixture-of-Expert Inference 1018
Ranggi Hwang (Korea Advanced Institute of Science & Technology, South Korea), Jianyu Wei (University of Science and Technology of China, China, and Microsoft Research, China), Shijie Cao (Microsoft Research, China), Changho Hwang (Microsoft Research, China), Xiaohu Tang (Microsoft Research, China), Ting Cao (Microsoft Research, China), and Mao Yang (Microsoft Research, China)

MECLA: Memory-Compute-Efficient LLM Accelerator with Scaling Sub-Matrix Partition 1032
Yubin Qin (Tsinghua University, China), Yang Wang (Tsinghua University, China), Zhiren Zhao (Tsinghua University, China), Xiaolong Yang (Tsinghua University, China), Yang Zhou (Tsinghua University, China), Shaojun Wei (Tsinghua University, China), Yang Hu (Tsinghua University, China), and Shouyi Yin (Tsinghua University, China)

Tender: Accelerating Large Language Models via Tensor Decomposition and Runtime Requantization 1048
Jungi Lee (Seoul National University, South Korea), Wonbeom Lee (Seoul National University, South Korea), and Jaewoong Sim (Seoul National University, South Korea)

Heterogeneous Acceleration Pipeline for Recommendation System Training 1063
Muhammad Adnan (University of British Columbia), Yassaman Ebrahimzadeh Maboud (University of British Columbia), Divya Mahajan (Georgia Institute of Technology), and Prashant J. Nair (University of British Columbia)

LLMCompass: Enabling Efficient Hardware Design for Large Language Model Inference 1080
Hengrui Zhang (Princeton University, USA), August Ning (Princeton University, USA), Rohan Baskar Prabhakar (Princeton University, USA), and David Wentzlaff (Princeton University, USA)

Session 9B: Memory Systems

DRAMScope: Uncovering DRAM Microarchitecture and Characteristics by Issuing Memory Commands 1097
Hwayong Nam (Seoul National University), Seungmin Baek (Seoul National University), Minbok Wi (Seoul National University), Michael Jaemin Kim (Seoul National University), Jaehyun Park (Seoul National University), Chihun Song (University of Illinois at Urbana-Champaign), Nam Sung Kim (University of Illinois at Urbana-Champaign), and Jung Ho Ahn (Seoul National University)

(MC)^2: Lazy MemCopy at the Memory Controller 1112
Aditya K. Kamath (University of Washington, USA) and Simon Peter (University of Washington, USA)

DyLeCT: Achieving Huge-Page-Like Translation Performance for Hardware-Compressed Memory 1129
Gagandeep Panwar (Virginia Tech), Muhammad Laghari (Virginia Tech), Esha Choukse (Microsoft Research), and Xun Jian (Virginia Tech)

Native DRAM Cache: Re-Architecting DRAM as a Large-Scale Cache for Data Centers 1144
Yesin Ryu (Sungkyunkwan University, Republic of Korea; Samsung Electronics, Republic of Korea), Yoojin Kim (Sungkyunkwan University, Republic of Korea), Giyong Jung (Sungkyunkwan University, Republic of Korea), Jung Ho Ahn (Seoul National University, Republic of Korea), and Jungrae Kim (Sungkyunkwan University, Republic of Korea)

PrIDE: Achieving Secure Rowhammer Mitigation with Low-Cost In-DRAM Trackers 1157
Aamer Jaleel (NVIDIA), Gururaj Saileshwar (University of Toronto), Stephen W. Keckler (NVIDIA), and Moinuddin Qureshi (Georgia Tech)

Session 10A: Prefetching

A New Formulation of Neural Data Prefetching 1173
Quang Duong (The University of Texas at Austin), Akanksha Jain (Google), and Calvin Lin (The University of Texas at Austin)

UDP: Utility-Driven Fetch Directed Instruction Prefetching 1188
Surim Oh (University of California, USA), Mingsheng Xu (University of California, USA), Tanvir Ahmed Khan (Columbia University, USA), Baris Kasikci (University of Washington, USA), and Heiner Litz (University of California, USA)

Triangel: A High-Performance, Accurate, Timely, On-Chip Temporal Prefetcher 1202
Sam Ainsworth (University of Edinburgh) and Lev Mukhanov (Queen Mary University of London)

Alternate Path Fetch 1217
Aniket Deshmukh (University of Texas at Austin, United States), Lingzhe (Chester) Cai (University of Texas at Austin, United States), and Yale N. Patt (University of Texas at Austin, United States)

Alternate Path μ-op Cache Prefetching 1230
Sawan Singh (University of Murcia, Spain), Arthur Perais (CNRS, France), Alexandra Jimborean (University of Murcia, Spain), and Alberto Ros (University of Murcia, Spain)

Session 10B: Accelerators for Emerging Workloads II

DaCapo: Accelerating Continuous Learning in Autonomous Systems for Video Analytics 1246
Yoonsung Kim (KAIST), Changhun Oh (KAIST), Jinwoo Hwang (KAIST), Wonung Kim (KAIST), Seongryong Oh (KAIST), Yubin Lee (KAIST), Hardik Sharma (Meta), Amir Yazdanbakhsh (Google DeepMind), and Jongse Park (KAIST)

BlissCam: Boosting Eye Tracking Efficiency with Learned In-Sensor Sparse Sampling 1262
Yu Feng (Shanghai Jiao Tong University), Tianrui Ma (Washington University in St. Louis), Yuhao Zhu (University of Rochester), and Xuan Zhang (Northeastern University)

BitNN: A Bit-Serial Accelerator for K-Nearest Neighbor Search in Point Clouds 1278
Meng Han (Beihang University, China), Liang Wang (Beihang University, China), Limin Xiao (Beihang University, China), Hao Zhang (Beihang University, China), Tianhao Cai (Beihang University, China), Jiale Xu (Shanghai Qi Zhi Institute, China), Yibo Wu (Tsinghua University, China), Chenhao Zhang (Beihang University, China), and Xiangrong Xu (Beihang University, China)

Cicero: Addressing Algorithmic and Architectural Bottlenecks in Neural Rendering by Radiance Warping and Memory Optimizations 1293
Yu Feng (Shanghai Jiao Tong University, China), Zihan Liu (Shanghai Jiao Tong University, China), Jingwen Leng (Shanghai Jiao Tong University, China), Minyi Guo (Shanghai Jiao Tong University, China), and Yuhao Zhu (University of Rochester, US)

GameStreamSR: Enabling Neural-Augmented Game Streaming on Commodity Mobile Platforms 1309
Sandeepa Bhuyan (The Pennsylvania State University, USA), Ziyu Ying (The Pennsylvania State University, USA), Mahmut T. Kandemir (The Pennsylvania State University, USA), Mahanth Gowda (The Pennsylvania State University, USA), and Chita R. Das (The Pennsylvania State University, USA)

Author Index