

2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO 2024)

**Austin, Texas, USA
2-6 November 2024**

Pages 1-839



**IEEE Catalog Number: CFP24071-POD
ISBN: 979-8-3503-5058-6**

**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:	CFP24071-POD
ISBN (Print-On-Demand):	979-8-3503-5058-6
ISBN (Online):	979-8-3503-5057-9
ISSN:	1072-4451

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 57th IEEE/ACM International Symposium on Microarchitecture (MICRO) **MICRO 2024**

Table of Contents

Message from the MICRO 2024 General Chairs	xxiii
Message from the MICRO 2024 Program Chairs	xxvi
MICRO 2024 Organizing Committee	xxviii
MICRO 2024 Steering Committee	xxx
MICRO 2024 Program Committee	xxxi
Keynotes	xxxv

Session 1A: Virtual Memory/Virtualization

Hardware-Assisted Virtualization of Neural Processing Units for Cloud Platforms	1
<i>Yuqi Xue (University of Illinois Urbana-Champaign), Yiqi Liu (University of Illinois Urbana-Champaign), Lifeng Nai (Google), and Jian Huang (University of Illinois Urbana-Champaign)</i>	
Elastic Translations: Fast Virtual Memory with Multiple Translation Sizes	17
<i>Stratos Psomadakis (National Technical University of Athens, Greece), Chloe Alverti (University of Urbana-Champaign, USA), Vasileios Karakostas (University of Athens, Greece), Christos Katsakioris (National Technical University of Athens, Greece), Dimitrios Siakavaras (National Technical University of Athens, Greece), Konstantinos Nikas (National Technical University of Athens, Greece), Georgios Goumas (National Technical University of Athens, Greece), and Nectarios Koziris (National Technical University of Athens, Greece)</i>	
Distributed Page Table: Harnessing Physical Memory as an Unbounded Hashed Page Table	36
<i>Osang Kwon (Sungkyunkwan University), Yongho Lee (Sungkyunkwan University), Junhyeok Park (Sungkyunkwan University), Sungbin Jang (Sungkyunkwan University), Byungchul Tak (Kyungpook National University), and Seokin Hong (Sungkyunkwan University)</i>	

Session 1B: Accelerators for Computer Vision

CamPU: A Multi-Camera Processing Unit for Deep Learning-Based 3D Spatial Computing Systems. 50	
<i>Dongseok Im (KAIST, South Korea) and Hoi-Jun Yoo (KAIST, South Korea)</i>	

AdapTiV: Sign-Similarity Based Image-Adaptive Token Merging for Vision Transformer Acceleration	64
<i>Seungjae Yoo (KAIST, South Korea), Hangeol Kim (KAIST, South Korea), and Joo-Young Kim (KAIST, South Korea)</i>	
Fusion-3D: Integrated Acceleration for Instant 3D Reconstruction and Real-Time Rendering	78
<i>Sixu Li (Georgia Institute of Technology, USA), Yang Zhao (Georgia Institute of Technology, USA), Chaojian Li (Georgia Institute of Technology, USA), Bowei Guo (Georgia Institute of Technology, USA), Jingqun Zhang (Georgia Institute of Technology, USA), Wenbo Zhu (Georgia Institute of Technology, USA), Zhifan Ye (Georgia Institute of Technology, USA), Cheng Wan (Georgia Institute of Technology, USA), and Yingyan Lin (Georgia Institute of Technology, USA)</i>	

Session 1C: Security Architecture

Secure Prefetching for Secure Cache Systems	92
<i>Sumon Nath (Indian Institute of Technology Bombay, India), Agustín Navarro-Torres (University of Murcia, Spain), Alberto Ros (University of Murcia, Spain), and Biswabandan Panda (Indian Institute of Technology Bombay, India)</i>	
HyperTEE: A Decoupled TEE Architecture with Secure Enclave Management	105
<i>Yunkai Bai (University of Chinese Academy of Sciences), Peinan Li (University of Chinese Academy of Sciences), Yubiao Huang (University of Chinese Academy of Sciences), Michael C. Huang (University of Rochester), Shijun Zhao (University of Chinese Academy of Sciences), Lutan Zhao (University of Chinese Academy of Sciences), Fengwei Zhang (Southern University of Science and Technology), Dan Meng (University of Chinese Academy of Sciences), and Rui Hou (University of Chinese Academy of Sciences)</i>	
Defending Against EMI Attacks on Just-In-Time Checkpoint for Resilient Intermittent Systems	121
<i>Choi Jaeseok (University of Central Florida, USA), Joe Hyunwoo (ETRI, South Korea), Jung Changhee (Purdue University, USA), and Choi Jongouk (University of Central Florida, USA)</i>	

Session 2A: Simulation

A Mess of Memory System Benchmarking, Simulation and Application Profiling	136
<i>Pouya Esmaili-Dokht (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Francesco Sgherzi (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Valèria Soldera Girelli (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Isaac Boixaderas (Barcelona Supercomputing Center), Mariana Carmin (Barcelona Supercomputing Center), Alireza Monemi (Barcelona Supercomputing Center), Adrià Armejach (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Estanislao Mercadal (Barcelona Supercomputing Center), Germán Lloret (Barcelona Supercomputing Center), Petar Radojković (Barcelona Supercomputing Center), Miquel Moreto (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Judit Giménez (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Xavier Martorell (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Eduard Ayguadé (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Jesus Labarta (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Emanuele Confalonieri (Micron technology), Rishabh Dubey (Micron technology), and Jason Adlard (Micron technology)</i>	
vTrain: A Simulation Framework for Evaluating Cost-Effective and Compute-Optimal Large Language Model Training	153
<i>Jehyeon Bang (KAIST), Yujeong Choi (KAIST), Myeongwoo Kim (Samsung Advanced Institute of Technology), Yongdeok Kim (Samsung Advanced Institute of Technology), and Minsoo Rhu (KAIST)</i>	
HyFiSS: A Hybrid Fidelity Stall-Aware Simulator for GPGPUs	168
<i>Jianchao Yang (National University of Defense Technology, China), Mei Wen (National University of Defense Technology, China), Dong Chen (Huawei Technologies Co., Ltd, China), Zhaoyun Chen (National University of Defense Technology, China), Zeyu Xue (National University of Defense Technology, China), Yuhang Li (National University of Defense Technology, China), Junzhong Shen (National University of Defense Technology, China), and Yang Shi (National University of Defense Technology, China)</i>	

Session 2B: Compiler Techniques/Optimizations

Unleashing CPU Potential for Executing GPU Programs Through Compiler/Runtime Optimizations....	186
<i>Ruobing Han (Georgia Institute of Technology, USA), Jisheng Zhao (Georgia Institute of Technology, USA), and Hyesoon Kim (Georgia Institute of Technology, USA)</i>	
A Framework for Fine-Grained Program Versioning	201
<i>Yishen Chen (MIT CSAIL) and Saman Amarasinghe (MIT CSAIL)</i>	
LightWSP: Whole-System Persistence on the Cheap	215
<i>Yuchen Zhou (Purdue University), Jianping Zeng (Purdue University), and Changhee Jung (Purdue University)</i>	

Session 2C: Reliability and Fault Tolerance

DelayAVF: Calculating Architectural Vulnerability Factors for Delay Faults	231
<i>Peter W. Deutsch (MIT, USA), Vincent Quentin Ulitzsch (MIT/TU, Germany), Sudhanva Gurumurthi (Advanced Micro Devices, Inc., USA), Vilas Sridharan (Advanced Micro Devices, Inc., USA), Joel S. Emer (MIT, USA), and Mengjia Yan (MIT, USA)</i>	
Polymorphic Error Correction	246
<i>Evgeny Manzhosov (Columbia University, USA) and Simha Sethumadhavan (Columbia University, USA)</i>	
DRCTL: A Disorder-Resistant Computation Translation Layer Enhancing the Lifetime and Performance of Memristive CIM Architecture	263
<i>Heng Zhou (Huazhong University of Science and Technology, China), Bing Wu (Huazhong University of Science and Technology, China), Huan Cheng (Huazhong University of Science and Technology, China), Jinpeng Liu (Huazhong University of Science and Technology, China), Taoming Lei (Huazhong University of Science and Technology, China), Dan Feng (Huazhong University of Science and Technology, China), and Wei Tong (Huazhong University of Science and Technology, China)</i>	

Session 3A: GPU Microarchitecture I

A Case for Speculative Address Translation with Rapid Validation for GPUs	278
<i>Junhyeok Park (Sungkyunkwan University), Osang Kwon (Sungkyunkwan University), Yongho Lee (Sungkyunkwan University), Seongwook Kim (Sungkyunkwan University), Gwangeun Byeon (Sungkyunkwan University), Jihun Yoon (Sungkyunkwan University), Prashant J. Nair (University of British Columbia), and Seokin Hong (Sungkyunkwan University)</i>	
SUV: Static Analysis Guided Unified Virtual Memory	293
<i>Pratheek B (Indian Institute of Science, India), Guilherme Cox (NVIDIA, USA), Jan Vesely (NVIDIA, USA), and Arkaprava Basu (Indian Institute of Science, India)</i>	
STAR: Sub-Entry Sharing-Aware TLB for Multi-Instance GPU	309
<i>Bingyao Li (University of Pittsburgh), Yueqi Wang (University of Pittsburgh), Tianyu Wang (University of Pittsburgh), Lieven Eeckhout (Ghent University), Jun Yang (University of Pittsburgh), Aamer Jaleel (NVIDIA), and Xulong Tang (University of Pittsburgh)</i>	
CacheCraft: Enhancing GPU Performance Under Memory Protection Through Reconstructed Caching	324
<i>Soyoung Park (Sungkyunkwan University, South Korea), Hojung Namkoong (Sungkyunkwan University, South Korea), Boyeol Choi (Sungkyunkwan University, South Korea), Michael B. Sullivan (NVIDIA, USA), and Jungrae Kim (Sungkyunkwan University, South Korea)</i>	

Session 3B: Security: Accelerators and Cryptography

Trinity: A General Purpose FHE Accelerator	338
<i>Xianglong Deng (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, China; University of Chinese Academy of Sciences, China), Shengyu Fan (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, China; University of Chinese Academy of Sciences, China), Zhicheng Hu (University of Electronic Science and Technology of China, China), Zhuoyu Tian (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, China; University of Chinese Academy of Sciences, China), Zihao Yang (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, China; University of Chinese Academy of Sciences, China), Jiangrui Yu (Peking University, China), Dingyuan Cao (University of Illinois, USA), Dan Meng (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, China), Rui Hou (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, China), Meng Li (Peking University, China), Qian Lou (University of Central Florida, USA), and Mingzhe Zhang (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, China)</i>	
UFC: A Unified Accelerator for Fully Homomorphic Encryption	352
<i>Minxuan Zhou (Illinois Institute of Technology), Yujin Nam (University of California San Diego), Xuan Wang (University of California San Diego), Youhak Lee (University of California San Diego), Chris Wilkerson (Intel Labs), Raghavan Kumar (Intel Labs), Sachin Taneja (Intel Labs), Sanu Mathew (Intel Labs), Rosario Cammarota (Intel Labs), and Tajana Rosing (University of California San Diego)</i>	
Accelerating Zero-Knowledge Proofs Through Hardware-Algorithm Co-Design	366
<i>Nikola Samardzic (MIT CSAIL), Simon Langowski (MIT CSAIL), Srinivas Devadas (MIT CSAIL), and Daniel Sanchez (MIT CSAIL)</i>	
A Compiler-Like Framework for Optimizing Cryptographic Big Integer Multiplication on GPUs ...	380
<i>Zhuoran Ji (Shandong University, China; Quan Cheng Laboratory, China), Jianyu Zhao (Shandong University, China), Zhaorui Zhang (The Hong Kong Polytechnic University, China), Jiming Xu (Ant Group, China), Shoumeng Yan (Ant Group, China), and Lei Ju (Quan Cheng Laboratory, China; Shandong University, China)</i>	

Session 3C: Open-Source Hardware/Hardware Design Tools

Beehive: A Flexible Network Stack for Direct-Attached Accelerators	393
<i>Katie Lim (University of Washington, USA), Matthew Giordano (University of Washington, USA), Theano Stavrinos (University of Washington, USA), Irene Zhang (Microsoft Research, USA), Jacob Nelson (Microsoft Research, USA), Baris Kasikci (University of Washington, USA), and Thomas Anderson (University of Washington, USA)</i>	
Stellar: An Automated Design Framework for Dense and Sparse Spatial Accelerators	409
<i>Hasan Nazim Genc (University of California, Berkeley), Hansung Kim (University of California, Berkeley), Prashanth Ganesh (University of California, Berkeley), and Yakun Sophia Shao (University of California, Berkeley)</i>	

LUCIE: A Universal Chiplet-Interposer Design Framework for Plug-and-Play Integration	423
<i>Zixi Li (Princeton University, USA) and David Wentzlaff (Princeton University, USA)</i>	
A Scalable, Efficient, and Robust Dynamic Memory Management Library for HLS-Based FPGAs ...	437
<i>Qinggong Wang (Huazhong University of Science and Technology, China; Zhejiang Lab, China), Long Zheng (Huazhong University of Science and Technology, China), Zhaozeng An (Huazhong University of Science and Technology, China), Shuyi Xiong (Huazhong University of Science and Technology, China), Runze Wang (Huazhong University of Science and Technology, China), Yu Huang (Huazhong University of Science and Technology, China; Zhejiang Lab, China), Pengcheng Yao (Huazhong University of Science and Technology, China; Zhejiang Lab, 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)</i>	

Session 4A: CPU Microarchitecture I

Customizing Cache Indexing Through Entropy Estimation	451
<i>Kevin Weston (Texas A&M University, USA), Avery Johnson (Texas A&M University, USA), Vahid Janfaza (Texas A&M University, USA), Farabi Mahmud (Texas A&M University, USA), and Abdullah Muzahid (Texas A&M University, USA)</i>	
The Last-Level Branch Predictor	464
<i>David Schall (University of Edinburgh, UK), Andreas Sandberg (Arm Limited, UK), and Boris Grot (University of Edinburgh, UK)</i>	
Timely, Efficient, and Accurate Branch Precomputation	480
<i>Aniket Deshmukh (University of Texas, USA), Lingzhe Chester Cai (University of Texas, USA), and Yale N. Patt (University of Texas, USA)</i>	
Localizing the Tag Comparisons in the Wakeup Logic to Reduce Energy Consumption of the Issue Queue	493
<i>Kenichiro Mori (Nagoya University, Japan), Sota Kosugi (Nagoya University, Japan), Hiroto Yoshida (Nagoya University, Japan), Hajime Shimada (Nagoya University, Japan), and Hideki Ando (Nagoya University, Japan)</i>	
RTL2M μ PATH: Multi- μ PATH Synthesis with Applications to Hardware Security Verification	507
<i>Yao Hsiao (Stanford University), Nikos Nikoleris (Arm), Artem Khyzha (Arm), Dominic P. Mulligan (Amazon Web Services), Gustavo Petri (Amazon Web Services), Christopher W. Fletcher (University of California, Berkeley), and Caroline Trippel (Stanford University)</i>	

Session 4B: Accelerators for ML

SRender: Boosting Neural Radiance Field Efficiency via Sensitivity-Aware Dynamic Precision Rendering	525
<i>Zhuoran Song (Shanghai Jiao Tong University, China), Houshu He (Shanghai Jiao Tong University, China), Fangxin Liu (Shanghai Jiao Tong University, China), Yifan Hao (Chinese Academy of Sciences, China), Xinkai Song (Chinese Academy of Sciences, China), Li Jiang (Shanghai Jiao Tong University, China), and Xiaoyao Liang (Shanghai Jiao Tong University, China)</i>	
Cambricon-C: Efficient 4-bit Matrix Unit via Primitivization	538
<i>Yi Chen (USTC SKLP, ICT, CAS), Yongwei Zhao (SKLP, ICT, CAS), Yifan Hao (SKLP, ICT, CAS), Yuanbo Wen (SKLP, ICT, CAS), Yuntao Dai (USTC), Xiaqing Li (SKLP, ICT, CAS), Yang Liu (SKLP, ICT, CAS UCAS), Rui Zhang (SKLP, ICT, CAS), Mo Zou (SKLP, ICT, CAS), Xinkai Song (SKLP, ICT, CAS), Xing Hu (SKLP, ICT, CAS SHIC), Zidong Du (SKLP, ICT, CAS SHIC), Huaping Chen (USTC), Qi Guo (SKLP, ICT, CAS), and Tianshi Chen (Cambrion Technologies)</i>	
BBS: Bi-Directional Bit-Level Sparsity for Deep Learning Acceleration	551
<i>Yuzong Chen (Cornell University, USA), Jian Meng (Cornell University, USA), Jae-sun Seo (Cornell University, USA), and Mohamed Abdelfattah (Cornell University, USA)</i>	
SCAR: Scheduling Multi-Model AI Workloads on Heterogeneous Multi-Chiplet Module Accelerators	565
<i>Mohanad Odema (University of California, Irvine, USA), Luke Chen (University of California, Irvine, USA), Hyoukjun Kwon (University of California, Irvine, USA), and Mohammad Abdullah Al Faruque (University of California, Irvine, USA)</i>	
SCALE: A Structure-Centric Accelerator for Message Passing Graph Neural Networks	580
<i>Lingxiang Yin (University of Central Florida, USA), Sanjay Gandham (University of Central Florida, USA), Mingjie Lin (University of Central Florida, USA), and Hao Zheng (University of Central Florida, USA)</i>	

Session 4C: Processing in/near Memory I

Low-Overhead General-Purpose Near-Data Processing in CXL Memory Expanders	594
<i>Hyungkyu Ham (POSTECH), Jeongmin Hong (POSTECH), Geonwoo Park (POSTECH), Yunseon Shin (POSTECH), Okkyun Woo (POSTECH), Wonhyuk Yang (POSTECH), Jinhoon Bae (POSTECH), Eunhyeok Park (POSTECH), Hyojin Sung (Seoul National University), Euicheol Lim (SK hynix), and Gwangsun Kim (POSTECH)</i>	
PIFS-Rec: Process-In-Fabric-Switch for Large-Scale Recommendation System Inferences	612
<i>Pingyi Huo (The Pennsylvania State University), Anusha Devoulapally (The Pennsylvania State University), Hasan Al Maruf (AMD, Inc), Minseo Park (AMD, Inc), Krishnakumar Nair (AMD, Inc), Meena Arunachalam (AMD, Inc), Gulsum Gudukbay Akbulut (The Pennsylvania State University), Mahmut Taylan Kandemir (The Pennsylvania State University), and Vijaykrishnan Narayanan (The Pennsylvania State University)</i>	

PIM-MMU: A Memory Management Unit for Accelerating Data Transfers in Commercial PIM Systems	627
<i>Dongjae Lee (KAIST), Bongjoon Hyun (KAIST), Taehun Kim (KAIST), and Minsoo Rhu (KAIST)</i>	
Azul: An Accelerator for Sparse Iterative Solvers Leveraging Distributed On-Chip Memory	643
<i>Axel Feldmann (MIT CSAIL), Courtney Golden (MIT CSAIL), Yifan Yang (MIT CSAIL), Joel Emer (MIT CSAIL/NVIDIA), and Daniel Sanchez (MIT CSAIL)</i>	
FloatAP: Supporting High-Performance Floating-Point Arithmetic in Associative Processors	657
<i>Kailin Yang (Cornell University, USA) and José Martínez (Cornell University, USA)</i>	

Session 5A: GPU Synchronization/Concurrency

Atomic Cache: Enabling Efficient Fine-Grained Synchronization with Relaxed Memory Consistency on GPGPUs Through In-Cache Atomic Operations	671
<i>Yicong Zhang (Sun Yat-sen University, China), Mingyu Wang (Sun Yat-sen University, China), Wangguang Wang (Sun Yat-sen University, China), Yangzhan Mai (Sun Yat-sen University, China), Haiqiu Huang (Sun Yat-sen University, China), and Zhiyi Yu (Sun Yat-sen University, China)</i>	
Concurrency-Aware Register Stacks for Efficient GPU Function Calls	686
<i>Ni Kang (Purdue University, USA), Ahmad Alawneh (Purdue University, USA), Mengchi Zhang (Purdue University, USA), and Timothy G. Rogers (Purdue University, USA)</i>	
CPElide: Efficient Multi-Chiplet GPU Implicit Synchronization	700
<i>Preyesh Dalmia (University of Wisconsin-Madison, USA), Rajesh Shashi Kumar (University of Wisconsin-Madison, USA), and Matthew D. Sinclair (University of Wisconsin-Madison, USA)</i>	

Session 5B: Quantum

Flag-Proxy Networks: Overcoming the Architectural, Scheduling and Decoding Obstacles of Quantum LDPC Codes	718
<i>Suhas Vittal (Georgia Institute of Technology), Ali Javadi-Abhari (IBM T.J. Watson Research Center), Andrew Cross (IBM T.J. Watson Research Center), Lev Bishop (IBM T.J. Watson Research Center), and Moinuddin Qureshi (Georgia Institute of Technology)</i>	
Qoncord: A Multi-Device Job Scheduling Framework for Variational Quantum Algorithms	735
<i>Meng Wang (The University of British Columbia, Canada), Poulami Das (The University of Texas at Austin, USA), and Prashant J. Nair (The University of British Columbia, Canada)</i>	
Surf-Deformer: Mitigating Dynamic Defects on Surface Code via Adaptive Deformation	750
<i>Keyi Yin (University of California San Diego, USA), Xiang Fang (University of California San Diego, USA), Travis S. Humble (Oak Ridge National Laboratory, USA), Ang Li (Pacific Northwest National Laboratory, USA), Yunong Shi (AWS Quantum Technologies, USA), and Yufei Ding (University of California San Diego, USA)</i>	

Session 5C: Debugging Correctness/Performance

Hestia: An Efficient Cross-Level Debugger for High-Level Synthesis	765
<i>Ruifan Xu (Peking University), Jin Luo (Peking University), Yawen Zhang (Peking University), Yibo Lin (Peking University), Runsheng Wang (Peking University), Ru Huang (Peking University), and Yun Liang (Peking University)</i>	
Looking into the Black Box: Monitoring Computer Architecture Simulations in Real-Time with AkitaRTM	780
<i>Ali Mosallaei (University of Michigan, USA), Katherine Isaacs (University of Utah, USA), and Yifan Sun (William & Mary, USA)</i>	
Over-Synchronization in GPU Programs	795
<i>Ajay Nayak (Indian Institute of Science, India) and Arkaprava Basu (Indian Institute of Science, India)</i>	

Session 6A: Cache Coherence

Temporarily Unauthorized Stores: Write First, Ask for Permission Later	810
<i>Juan M. Cebrian (University of Murcia, Spain), Magnus Jahre (Norwegian University of Science and Technology (NTNU), Norway), and Alberto Ros (University of Murcia, Spain)</i>	
Leveraging Cache Coherence to Detect and Repair False Sharing On-the-Fly	823
<i>Vipin Patel (Indian Institute of Technology Kanpur, India), Swarnendu Biswas (Indian Institute of Technology Kanpur, India), and Mainak Chaudhuri (Indian Institute of Technology Kanpur, India)</i>	
Chaining Transactions for Effective Concurrency Management in Hardware Transactional Memory	840
<i>Víctor Nicolás-Conesa (University of Murcia, Spain), Rubén Títos-Gil (University of Murcia, Spain), Ricardo Fernández-Pascual (University of Murcia, Spain), Manuel E. Acacio (University of Murcia, Spain), and Alberto Ros (University of Murcia, Spain)</i>	

Session 6B: Networks-on-Chip

TACOS: Topology-Aware Collective Algorithm Synthesizer for Distributed Machine Learning	856
<i>William Won (Georgia Institute of Technology), Midhilesh Elavazhagan (Intel), Sudarshan Srinivasan (Intel), Swati Gupta (Massachusetts Institute of Technology), and Tushar Krishna (Georgia Institute of Technology)</i>	
Ring Road: A Scalable Polar-Coordinate-Based 2D Network-on-Chip Architecture	871
<i>Yinxiao Feng (Tsinghua University, China), Wei Li (Tsinghua University, China), and Kaisheng Ma (Tsinghua University, China)</i>	

Uncovering Real GPU NoC Characteristics: Implications on Interconnect Architecture	885
<i>Zhixian Jin (KAIST, Republic of Korea), Christopher Rocca (KAIST, Republic of Korea), Jiho Kim (KAIST, Republic of Korea), Hans Kasan (KAIST, Republic of Korea), Minsoo Rhu (KAIST, Republic of Korea), Ali Bakhoda (Microsoft, USA), Tor Aamodt (University of British Columbia, Canada), and John Kim (KAIST, Republic of Korea)</i>	

Session 6C: Rowhammer

MINT: Securely Mitigating Rowhammer with a Minimalist In-DRAM Tracker	899
<i>Moinuddin Qureshi (Georgia Tech), Salman Qazi (Google), and Amer Jaleel (Nvidia)</i>	
BreakHammer: Enhancing RowHammer Mitigations by Carefully Throttling Suspect Threads	915
<i>Oğuzhan Canpolat (TOBB ETÜ, Turkey & ETH Zürich, Switzerland), Abdullah Giray Yağlıkçı (ETH Zürich, Switzerland), Ataberk Olgun (ETH Zürich, Switzerland), Ismail Emir Yuksel (ETH Zürich, Switzerland), Yahya Can Tuğrul (TOBB ETÜ, Turkey & ETH Zürich, Switzerland), Konstantinos Kanellopoulos (ETH Zürich, Switzerland), Oğuz Ergin (University of Sharjah, United Arab Emirates & ETH Zürich, Switzerland & TOBB ETÜ, Turkey), and Onur Mutlu (ETH Zürich, Switzerland & Stanford University, United States)</i>	
ImPress: Securing DRAM Against Data-Disturbance Errors via Implicit Row-Press Mitigation	935
<i>Anish Saxena (Georgia Institute of Technology), Amer Jaleel (NVIDIA), and Moinuddin Qureshi (Georgia Institute of Technology)</i>	

Session 7A: Memory

Self-Managing DRAM: A Low-Cost Framework for Enabling Autonomous and Efficient DRAM Maintenance Operations	949
<i>Hasan Hassan (ETH Zurich), Ataberk Olgun (ETH Zurich), Abdullah Giray Yaglikci (ETH Zurich), Haocong Luo (ETH Zurich), and Onur Mutlu (ETH Zurich)</i>	
Memory Allocation Under Hardware Compression	966
<i>Muhammad Laghari (Virginia Tech), Yuqing Liu (Virginia Tech), Gagandeep Panwar (Virginia Tech), David Bears (Virginia Tech), Chandler Jearls (Virginia Tech), Raghavendra Srinivas (Virginia Tech), Esha Choukse (Microsoft Research), Kirk W. Cameron (Virginia Tech), Ali R. Butt (Virginia Tech), and Xun Jian (Virginia Tech)</i>	
Genie Cache: Non-Blocking Miss Handling and Replacement in Page-Table-Based DRAM Cache ..	983
<i>Youngin Kim (Yonsei University, South Korea) and William Song (Yonsei University, South Korea)</i>	
StarNUMA: Mitigating NUMA Challenges with Memory Pooling	997
<i>Albert Cho (Georgia Institute of Technology, USA) and Alexandros Daglis (Georgia Institute of Technology, USA)</i>	

Session 7B: GPU Microarchitecture II

ThreadFuser: A SIMT Analysis Framework for MIMD Programs	1013
<i>Ahmad Alawneh (Purdue University, USA), Ni Kang (Purdue University, USA), Mahmoud Khairy (Purdue University, USA), and Timothy G. Rogers (Purdue University, USA)</i>	
Extending GPU Ray-Tracing Units for Hierarchical Search Acceleration	1027
<i>Aaron Barnes (Purdue University, USA), Fangjia Shen (Purdue University, USA), and Timothy G. Rogers (Purdue University, USA)</i>	
Generalizing Ray Tracing Accelerators for Tree Traversals on GPUs	1041
<i>Dongho Ha (Yonsei University, South Korea), Lufei Liu (University of British Columbia, Canada), Yuan Hsi Chou (University of British Columbia, Canada), Seokjin Go (Yonsei University, South Korea), Won Woo Ro (Yonsei University, South Korea), Hung-Wei Tseng (University of California, Riverside, USA), and Tor M. Aamodt (University of British Columbia, Canada)</i>	
LIBRA: Memory Bandwidth- and Locality-Aware Parallel Tile Rendering	1058
<i>Aurora Tomás (Universitat Politècnica de Catalunya), Juan L. Aragón (Universidad de Murcia), Joan-Manuel Parcerisa (Universitat Politècnica de Catalunya), and Antonio González (Universitat Politècnica de Catalunya)</i>	

Session 7C: Neuromorphic Processors & SNN

Rearchitecting a Neuromorphic Processor for Spike-Driven Brain-Computer Interfacing	1073
<i>Hunjun Lee (Hanyang University, South Korea), Yeongwoo Jang (Seoul National University, South Korea), Daye Jung (Seoul National University, South Korea), Seunghyun Song (Seoul National University, South Korea), and Jangwoo Kim (Seoul National University, South Korea)</i>	
COMPASS: SRAM-Based Computing-in-Memory SNN Accelerator with Adaptive Spike Speculation .	1090
<i>Zongwu Wang (Shanghai Jiao Tong University, China; Shanghai Qi Zhi Institute, China), Fangxin Liu (Shanghai Jiao Tong University, China; Shanghai Qi Zhi Institute, China), Ning Yang (Shanghai Jiao Tong University, China; Shanghai Qi Zhi Institute, China), Shiyuan Huang (Shanghai Jiao Tong University, China; Shanghai Qi Zhi Institute, China), Haomin Li (Shanghai Jiao Tong University, China; Shanghai Qi Zhi Institute, China), and Li Jiang (Shanghai Jiao Tong University, China; Shanghai Qi Zhi Institute, China)</i>	
LoAS: Fully Temporal-Parallel Dataflow for Dual-Sparse Spiking Neural Networks	1107
<i>Ruokai Yin (Yale University, USA), Youngeun Kim (Yale University, USA), Di Wu (University of Central Florida, USA), and Priyadarshini Panda (Yale University, USA)</i>	
ActiveN: A Scalable and Flexibly-Programmable Event-Driven Neuromorphic Processor	1122
<i>Xiaoyi Liu (Tsinghua University, China), Zhongzhu Pu (Tsinghua University, China), Peng Qu (Tsinghua University, China), Weimin Zheng (Tsinghua University, China), and Youhui Zhang (Tsinghua University, China; Zhongguancun Laboratory, China)</i>	

Session 8A: Side-Channel Attacks/Defenses

Ghost Arbitration: Mitigating Interconnect Side-Channel Timing Attacks in GPU	1138
<i>Zhixian Jin (KAIST), Jaeguk Ahn (KAIST), Jiho Kim (KAIST), Hans Kasan (KAIST), Jina Song (KAIST), Wonjun Song (Kangwon National University), and John Kim (KAIST)</i>	
IvLeague: Side Channel-resistant Secure Architectures Using Isolated Domains of Dynamic Integrity Trees	1153
<i>Md Hafizul Islam Chowdhuryy (University of Central Florida) and Fan Yao (University of Central Florida)</i>	
Veiled Pathways: Investigating Covert and Side Channels Within GPU Uncore	1169
<i>Yuanqing Miao (Penn State University, USA), Yingtian Zhang (Penn State University, USA), Dinghao Wu (Penn State University, USA), Danfeng Zhang (Duke Univeristy, USA), Gang Tan (Penn State University, USA), Rui Zhang (Penn State University, USA), and Mahmut Kandemir (Penn State University, USA)</i>	

Session 8B: Dataflow & Recommendation Systems

The TYR Dataflow Architecture: Improving Locality by Taming Parallelism	1184
<i>Nikhil Agarwal (Carnegie Mellon University), Mitchell Fream (Carnegie Mellon University), Souradip Ghosh (Carnegie Mellon University), Brian C. Schwedock (Samsung), and Nathan Beckmann (Carnegie Mellon University)</i>	
Sparsepipe: Sparse Inter-Operator Dataflow Architecture with Cross-Iteration Reuse	1201
<i>Yunan Zhang (University of California, Riverside, USA), Po-An Tsai (Nvidia, USA), and Hung-Wei Tseng (University of California, Riverside, USA)</i>	
Pushing the Performance Envelope of DNN-based Recommendation Systems Inference on GPUs	1217
<i>Rishabh Jain (The Pennsylvania State University, USA), Vivek M. Bhasi (The Pennsylvania State University, USA), Adwait Jog (University of Virginia, USA), Anand Sivasubramaniam (The Pennsylvania State University, USA), Mahmut T. Kandemir (The Pennsylvania State University, USA), and Chita R. Das (The Pennsylvania State University, USA)</i>	

Session 8C: SRC Competition Presentations

Session 9A: Accelerators for Sparsity & GNN

Terminus: A Programmable Accelerator for Read and Update Operations on Sparse Data Structures	1233
<i>Hyun Ryong Lee (Massachusetts Institute of Technology) and Daniel Sanchez (Massachusetts Institute of Technology)</i>	

- SOFA: A Compute-Memory Optimized Sparsity Accelerator via Cross-Stage Coordinated Tiling 1247
Huizheng Wang (Tsinghua University, China), Jiahao Fang (Tsinghua University, China), Xinru Tang (Tsinghua University, China), Zhiheng Yue (Tsinghua University, China), Jinxi Li (Tsinghua University, China), Yubin Qin (Tsinghua University, China), Sihan Guan (Tsinghua University, China), Qize Yang (Tsinghua University, China), Yang Wang (Tsinghua University, China), Chao Li (Shanghai Jiao Tong University, China), Yang Hu (Tsinghua University, China; Shanghai Artificial Intelligence Laboratory, China), and Shouyi Yin (Tsinghua University, China; Shanghai Artificial Intelligence Laboratory, China)
- RAHP: A Redundancy-Aware Accelerator for High-Performance Hypergraph Neural Network .. 1264
Hui Yu (Huazhong University of Science and Technology, China), Yu Zhang (Huazhong University of Science and Technology, China), Ligang He (University of Warwick, United Kingdom), Yingqi Zhao (Huazhong University of Science and Technology, China), Xintao Li (Huazhong University of Science and Technology, China), Ruida Xin (Huazhong University of Science and Technology, China), Jin Zhao (Huazhong University of Science and Technology, China), Xiaofei Liao (Huazhong University of Science and Technology, China), Haikun Liu (Huazhong University of Science and Technology, China), Bingsheng He (National University of Singapore, Singapore), and Hai Jin (Huazhong University of Science and Technology, China)

Session 9B: Processing in/near Memory II

- Leviathan: A Unified System for General-Purpose Near-Data Computing 1278
Brian C. Schwedock (Samsung) and Nathan Beckmann (Carnegie Mellon University)
- TMiner: A Vertex-Based Task Scheduling Architecture for Graph Pattern Mining 1295
Zerun Li (University of Chinese Academy of Sciences, China), Xiaoming Chen (Institute of Computing Technology, Chinese Academy of Sciences, China), and Yinhe Han (Institute of Computing Technology, Chinese Academy of Sciences, China)
- PointCIM: A Computing-in-Memory Architecture for Accelerating Deep Point Cloud Analytics .. 1309
Xuan-Jun Chen (National Taiwan University, Taiwan), Han-Ping Chen (National Taiwan University, Taiwan), and Chia-Lin Yang (National Taiwan University, Taiwan)

Session 9C: Reconfigurable Architectures

- Blenda: Dynamically-Reconfigurable Stacked DRAM 1323
Mohammad Bakhshalipour (Unaffiliated), Hamidreza Zare (Pennsylvania State University), Farid Samandi (Stony Brook University), Fatemeh Golshan (University of Pittsburgh), Pejman Lotfi-Kamran (IPM), and Hamid Sarbazi-Azad (Sharif University of Technology; IPM)

ICED: An Integrated CGRA Framework Enabling DVFS-Aware Acceleration	1338
<i>Cheng Tan (Google; Arizona State University), Miaomiao Jiang (Shandong University; Arizona State University), Deepak Patil (Arizona State University), Yanghui Ou (Cornell University), Zhaoying Li (National University of Singapore), Lei Ju (Shandong University), Tulika Mitra (National University of Singapore), Hyunchul Park (Google), Antonino Tumeo (Pacific Northwest National Laboratory), and Jeff Zhang (Arizona State University)</i>	
SambaNova SN40L: Scaling the AI Memory Wall with Dataflow and Composition of Experts	1353
<i>Raghu Prabhakar (SambaNova Systems, Inc., USA), Ram Sivaramakrishnan (SambaNova Systems, Inc., USA), Darshan Gandhi (SambaNova Systems, Inc., USA), Yun Du (SambaNova Systems, Inc., USA), Mingran Wang (SambaNova Systems, Inc., USA), Xiangyu Song (SambaNova Systems, Inc., USA), Kejie Zhang (SambaNova Systems, Inc., USA), Tianren Gao (SambaNova Systems, Inc., USA), Angela Wang (SambaNova Systems, Inc., USA), Xiaoyan Li (SambaNova Systems, Inc., USA), Yongning Sheng (SambaNova Systems, Inc., USA), Joshua Brot (SambaNova Systems, Inc., USA), Denis Sokolov (SambaNova Systems, Inc., USA), Apurv Vivek (SambaNova Systems, Inc., USA), Calvin Leung (SambaNova Systems, Inc., USA), Arjun Sabnis (SambaNova Systems, Inc., USA), Jiayu Bai (SambaNova Systems, Inc., USA), Tuowen Zhao (SambaNova Systems, Inc., USA), Mark Gottscho (SambaNova Systems, Inc., USA), David Jackson (SambaNova Systems, Inc., USA), Mark Luttrell (SambaNova Systems, Inc., USA), Manish K. Shah (SambaNova Systems, Inc., USA), Zhengyu Chen (SambaNova Systems, Inc., USA), Kaizhao Liang (SambaNova Systems, Inc., USA), Swayambhoo Jain (SambaNova Systems, Inc., USA), Urmish Thakker (SambaNova Systems, Inc., USA), Dawei Huang (SambaNova Systems, Inc., USA), Sumti Jairath (SambaNova Systems, Inc., USA), Kevin J. Brown (SambaNova Systems, Inc., USA), and Kunle Olukotun (SambaNova Systems, Inc., USA)</i>	

Session 10A: CPU Microarchitecture II

Scalar Vector Runahead	1367
<i>Jaime Roelandts (Ghent University, Belgium), Ajeya Naithani (Ghent University, Belgium), Sam Ainsworth (University of Edinburgh, UK), Timothy M. Jones (University of Cambridge, UK), and Lieven Eeckhout (Ghent University, Belgium)</i>	
Weeding out Front-End Stalls with Uneven Block Size Instruction Cache	1382
<i>Brunner Roman (Norwegian University of Science and Technology (NTNU), Norway) and Kumar Rakesh (Norwegian University of Science and Technology (NTNU), Norway)</i>	
Mosaic: Harnessing the Micro-Architectural Resources of Servers in Serverless Environments.....	1397
<i>Jovan Stojkovic (University of Illinois at Urbana-Champaign), Esha Choukse (Microsoft Azure Research - Systems), Enrique Saurez (Microsoft Azure Research - Systems), Íñigo Goiri (Microsoft Azure Research - Systems), and Josep Torrellas (University of Illinois at Urbana-Champaign)</i>	

SOPHGO BM1684X: A Commercial High Performance Terminal AI Processor with Large Model Support	1413
<i>Peng Gao (Beijing University of Posts and Telecommunications, China), Yang Liu (Beijing University of Posts and Telecommunications, China), Jun Wang (SOPHGO TECHNOLOGIES PTE. LTD., China), Wanlin Cai (SOPHGO TECHNOLOGIES PTE. LTD., China), Guangchong Shen (SOPHGO TECHNOLOGIES PTE. LTD., China), Zonghui Hong (SOPHGO TECHNOLOGIES PTE. LTD., China), Jiali Qu (SOPHGO TECHNOLOGIES PTE. LTD., China), and Ning Wang (Beijing University of Posts and Telecommunications, China)</i>	

Session 10B: Large Language Models

Duplex: A Device for Large Language Models with Mixture of Experts, Grouped Query Attention, and Continuous Batching	1429
<i>Sungmin Yun (Seoul National University, South Korea), Kwanhee Kyung (Seoul National University, South Korea), Juhwan Cho (Seoul National University, South Korea), Jaewan Choi (Seoul National University, South Korea), Jongmin Kim (Seoul National University, South Korea), Byeongho Kim (Samsung Electronics, South Korea), Sukhan Lee (Samsung Electronics, South Korea), Kyomin Sohn (Samsung Electronics, South Korea), and Jung Ho Ahn (Seoul National University, South Korea)</i>	
VGA: Hardware Accelerator for Scalable Long Sequence Model Inference	1444
<i>Seung Yul Lee (Seoul National University, Republic of Korea), Hyunseung Lee (Seoul National University, Republic of Korea), Jihoon Hong (Seoul National University, Republic of Korea), SangLyul Cho (Seoul National University, Republic of Korea), and Jae W. Lee (Seoul National University, Republic of Korea)</i>	
FuseMax: Leveraging Extended Einsums to Optimize Attention Accelerator Design	1458
<i>Nandeeka Nayak (University of California, Berkeley, USA), Xinrui Wu (Tsinghua University, China), Toluwanimi Odemuyiwa (University of California, Davis, USA), Michael Pellauer (NVIDIA, USA), Joel Emer (Massachusetts Institute of Technology/NVIDIA, USA), and Christopher Fletcher (University of California, Berkeley, USA)</i>	

Cambricon-LLM: A Chiplet-Based Hybrid Architecture for On-Device Inference of 70B LLM	1474
<i>Zhongkai Yu (SKL of Processors, Institute of Computing Technology, CAS, China; University of Chinese Academy of Sciences, China), Shengwen Liang (SKL of Processors, Institute of Computing Technology, CAS, China), Tianyun Ma (University of Science and Technology of China, China), Yunke Cai (SKL of Processors, Institute of Computing Technology, CAS, China; University of Chinese Academy of Sciences, China), Ziyuan Nan (SKL of Processors, Institute of Computing Technology, CAS, China; University of Chinese Academy of Sciences, China), Di Huang (SKL of Processors, Institute of Computing Technology, CAS, China;), Xinkai Song (SKL of Processors, Institute of Computing Technology, CAS, China;), Yifan Hao (SKL of Processors, Institute of Computing Technology, CAS, China;), Jie Zhang (Peking University, China), Tian Zhi (SKL of Processors, Institute of Computing Technology, CAS, China;), Yongwei Zhao (SKL of Processors, Institute of Computing Technology, CAS, China;), Zidong Du (SKL of Processors, Institute of Computing Technology, CAS, China;), Xing Hu (SKL of Processors, Institute of Computing Technology, CAS, China;), Qi Guo (SKL of Processors, Institute of Computing Technology, CAS, China;), and Tianshi Chen (Cambricon Technologies Co., Ltd., China)</i>	

Session 10C: Storage Systems and CXL

Ares-Flash: Efficient Parallel Integer Arithmetic Operations Using NAND Flash Memory	1489
<i>Jian Chen (Tsinghua University, China), Congming Gao (Xiamen University, China), Youyou Lu (Tsinghua University, China), Yuhao Zhang (Tsinghua University, China), and Jiwu Shu (Tsinghua University, China)</i>	
Demystifying a CXL Type-2 Device: A Heterogeneous Cooperative Computing Perspective	1504
<i>Houxiang Ji (University of Illinois Urbana-Champaign), Srikar Vanavasam (University of Illinois Urbana-Champaign), Yang Zhou (University of Illinois Urbana-Champaign), Qirong Xia (University of Illinois Urbana-Champaign), Jinghan Huang (University of Illinois Urbana-Champaign), Yifan Yuan (Intel Labs), Ren Wang (Intel Labs), Pekon Gupta (Intel Altera), Bhushan Chitlur (Intel Altera), Ipoom Jeong (Yonsei University), and Nam Sung Kim (University of Illinois Urbana-Champaign)</i>	
NeoMem: Hardware/Software Co-Design for CXL-Native Memory Tiering	1518
<i>Zhe Zhou (School of Integrated Circuits; Peking University), Yiqi Chen (School of Integrated Circuits), Tao Zhang (Microsoft Research), Yang Wang (Microsoft Research), Ran Shu (Microsoft Research), Shuotao Xu (Microsoft Research), Peng Cheng (Microsoft Research), Lei Qu (Microsoft Research), Yongqiang Xiong (Microsoft Research), Jie Zhang (Peking University; Zhongguancun Laboratory), and Guangyu Sun (School of Integrated Circuits; Beijing Advanced Innovation Center for Integrated Circuits)</i>	

Session 11A: Emerging Technologies/Applications

- SuperCore: An Ultra-Fast Superconducting Processor For Cryogenic Applications 1532
Junhyuk Choi (Seoul National University, South Korea), Ilkwon Byun (Kyushu University, Japan), Juwon Hong (Seoul National University, South Korea), Dongmoon Min (Seoul National University, South Korea), Junpyo Kim (Seoul National University, South Korea), Jungmin Cho (Seoul National University, South Korea), Hyeonseong Jeong (Seoul National University, South Korea), Masamitsu Tanaka (Nagoya University, Japan), Koji Inoue (Kyushu University, Japan), and Jangwoo Kim (Seoul National University, South Korea)
- SOPHIE: A Scalable Recurrent Ising Machine Using Optically Addressed Phase Change Memory 1548
Guowei Yang (Boston University), Sina Karimi (Boston University), Carlos A. Ríos Ocampo (University of Maryland), Ayse K. Coskun (Boston University), and Ajay Joshi (Boston University)
- GauSPU: 3D Gaussian Splatting Processor for Real-Time SLAM Systems 1562
Lizhou Wu (Fudan University, China), Haozhe Zhu (Fudan University, China), Siqi He (Fudan University, China), Jiawei Zheng (Fudan University, China), Chixiao Chen (Fudan University, China), and Xiaoyang Zeng (Fudan University, China)

Session 11B: FPGA Architectures and Accelerators

- Multi-Issue Butterfly Architecture for Sparse Convex Quadratic Programming 1574
Maolin Wang (The Hong Kong University of Science and Technology), Ian McNerney (Imperial College London), Bartolomeo Stellato (Princeton University), Fengbin Tu (The Hong Kong University of Science and Technology), Stephen Boyd (Stanford University), Hayden Kwok-Hay So (University of Hong Kong), and Kwang-Ting Cheng (The Hong Kong University of Science and Technology)
- HgPCN: A Heterogeneous Architecture for E2E Embedded Point Cloud Inference 1588
Yiming Gao (University of Florida), Chao Jiang (University of Florida), Wesley Piard (University of Florida), Xiangru Chen (University of Florida), Bhavesh Patel (Dell EMC), and Herman Lam (University of Florida)
- Acamar: A Dynamically Reconfigurable Scientific Computing Accelerator for Robust Convergence and Minimal Resource Underutilization 1601
Ubaid Bakhtiar (University of Maryland, USA), Helya Hosseini (University of Maryland, USA), and Bahar Asgari (University of Maryland, USA)
- Bridging the Gap Between LLMs and LNS with Dynamic Data Format and Architecture Codesign 1617
Pouya Haghi (University of Rochester, USA), Chunshu Wu (University of Rochester, USA), Zahra Azad (University of Rochester, USA), Yanfei Li (Pacific Northwest National Laboratory, USA), Andrew Gui (Pacific Northwest National Laboratory, USA), Yuchen Hao (Meta, USA), Ang Li (Pacific Northwest National Laboratory, USA), and Tong Geng (University of Rochester, USA)

Session 11C: Processing in/near Memory III

PyPIM: Integrating Digital Processing-in-Memory from Microarchitectural Design to Python Tensors	1632
<i>Orian Leitersdorf (Technion - Israel Institute of Technology, Israel), Rommy Ronen (Technion - Israel Institute of Technology, Israel), and Shahar Kvatinsky (Technion - Israel Institute of Technology, Israel)</i>	
Stream-Based Data Placement for Near-Data Processing with Extended Memory	1648
<i>Yiwei Li (Tsinghua University), Boyu Tian (Tsinghua University), Yi Ren (Tsinghua University), and Mingyu Gao (Tsinghua University; Shanghai Qi Zhi Institute)</i>	
Cambricon-M: A Fibonacci-Coded Charge-Domain SRAM-Based CIM Accelerator for DNN Inference 1663	
<i>Hongrui Guo (Institute of Computing Technology, CAS, China; University of Chinese Academy of Sciences, China), Mo Zou (Institute of Computing Technology, CAS, China), Yifan Hao (Institute of Computing Technology, CAS, China), Zidong Du (Institute of Computing Technology, CAS, China; Shanghai Innovation Center for Processor Technologies), Erxiang Ren (Beijing Jiaotong University), Yang Liu (Institute of Computing Technology, CAS, China; University of Chinese Academy of Sciences, China), Yongwei Zhao (Institute of Computing Technology, CAS, China), Tianrui Ma (Institute of Computing Technology, CAS, China), Rui Zhang (Institute of Computing Technology, CAS, China), Xing Hu (Institute of Computing Technology, CAS, China; Shanghai Innovation Center for Processor Technologies), Fei Qiao (Tsinghua University), Zhiwei Xu (Institute of Computing Technology, CAS, China; University of Chinese Academy of Sciences, China), Qi Guo (Institute of Computing Technology, CAS, China), and Tianshi Chen (3Cambricon Technologies Co., Ltd., China)</i>	
MeMCISA: Memristor-Enabled Memory-Centric Instruction-Set Architecture for Database Workloads	1678
<i>Yihang Zhu (Peking University, China), Lei Cai (Peking University, China), Lianfeng Yu (Peking University, China), Anjunyi Fan (Peking University, China), Longhao Yan (Peking University, China), Zhaokun Jing (Peking University, China), Bonan Yan (Peking University, China), Pek Jun Tiw (Peking University, China), Yuqi Li (Peking University, China), Yaoyu Tao (Peking University, China), and Yuchao Yang (Peking University, China)</i>	

Author Index