



48th International Symposium on Computer Architecture



ISCA 2021

Worldwide Event

June 14-19, 2021

Main Program

<https://iscaconf.org>

Main Program

Session 1: Industry track

[Ten Lessons from Three Generations Shaped Google's TPUv4i1](#)

Norman P. Jouppi (Google LLC), Doe Hyun Yoon (Google LLC), Matthew Ashcraft (Google LLC), Mark Gottscho (Google LLC), Thomas B. Jablin (Google LLC), George Kurian (Google LLC), James Laudon (Google LLC), Sheng Li (Google LLC), Peter Ma (Google LLC), Xiaoyu Ma (Google LLC), Thomas Norrie (Google LLC), Nishant Patil (Google LLC), Sushma Prasad (Google LLC), Clifford Young (Google LLC), Zongwei Zhou (Google LLC), David Patterson (Google LLC)

[Sparsity-Aware and Re-Configurable NPU Architecture for Samsung Flagship Mobile SoC: Industrial Product15](#)

Jun-Woo Jang (Samsung Advanced Institute of Technology, Korea), Sehwan Lee (Samsung Advanced Institute of Technology, Korea), Dongyoung Kim (Samsung Advanced Institute of Technology, Korea), Hyunsun Park (Samsung Advanced Institute of Technology, Korea), Ali Shafiee Ardestani (Samsung Semiconductor, Inc., USA), Yeongjae Choi (Samsung Advanced Institute of Technology, Korea), Channoh Kim (Samsung Advanced Institute of Technology, Korea), Yoojin Kim (Samsung Advanced Institute of Technology, Korea), Hyeongseok Yu (Samsung Advanced Institute of Technology, Korea), Hamzah Abdel-Aziz (Samsung Semiconductor, Inc., USA), Jun-Seok Park (Samsung Electronics, Korea), Heonsoo Lee (Samsung Electronics, Korea), Dongwoo Lee (Samsung Electronics, Korea), Myeong Woo Kim (Samsung Advanced Institute of Technology, Korea), Hanwoong Jung (Samsung Advanced Institute of Technology, Korea), Heewoo Nam (Samsung Advanced Institute of Technology, Korea), Dongguen Lim (Samsung Advanced Institute of Technology, Korea), Seungwon Lee (Samsung Advanced Institute of Technology, Korea), Joon-Ho Song (Samsung Advanced Institute of Technology, Korea), Suknam Kwon (Samsung Electronics, Korea), Joseph Hassoun (Samsung Semiconductor, Inc., USA), SukHwan Lim (Samsung Electronics, Korea), Changkyu Choi (Samsung Advanced Institute of Technology, Korea)

[Energy Efficiency Boost in the AI-Infused POWER 10 Processor: Industrial Product29](#)

Brian W. Thompto (International Business Machines Corporation, Armonk, USA), Dung Q. Nguyen (International Business Machines Corporation, Armonk, USA), José E. Moreira (International Business Machines Corporation, Armonk, USA), Ramon Bertran (International Business Machines Corporation, Armonk, USA), Hans Jacobson (International Business Machines Corporation, Armonk, USA), Richard J. Eickemeyer (International Business Machines Corporation, Armonk, USA), Rahul M. Rao (International Business Machines Corporation, Armonk, USA), Michael Goulet (International Business Machines Corporation, Armonk, USA), Marcy Byers (International Business Machines Corporation, Armonk, USA), Christopher J. Gonzalez (International Business Machines Corporation, Armonk, USA), Karthik Swaminathan (International Business Machines Corporation, Armonk, USA), Nagu R. Dhanwada (International Business Machines Corporation, Armonk, USA), Silvia M. Müller (International Business

Machines Corporation, Armonk, USA), Andreas Wagner (International Business Machines Corporation, Armonk, USA), Satish Kumar Sadasivam (International Business Machines Corporation, Armonk, USA), Robert K. Montoye (International Business Machines Corporation, Armonk, USA), William J. Starke (International Business Machines Corporation, Armonk, USA), Christian G. Zoellin (International Business Machines Corporation, Armonk, USA), Michael S. Floyd (International Business Machines Corporation, Armonk, USA), Jeffrey Stuecheli (International Business Machines Corporation, Armonk, USA), Nandhini Chandramoorthy (International Business Machines Corporation, Armonk, USA), John-David Wellman (International Business Machines Corporation, Armonk, USA), Alper Buyuktosunoglu (International Business Machines Corporation, Armonk, USA), Matthias Pflanz (International Business Machines Corporation, Armonk, USA), Balaram Sinharoy (International Business Machines Corporation, Armonk, USA), Pradip Bose (International Business Machines Corporation, Armonk, USA)

[Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology: Industrial Product43](#)

Sukhan Lee (Memory Business Division, Samsung Electronics), Shin-haeng Kang (Memory Business Division, Samsung Electronics), Jaehoon Lee (Memory Business Division, Samsung Electronics), Hyeonsu Kim (Samsung Advanced Institute of Technology, Samsung Electronics), Eojin Lee (Memory Business Division, Samsung Electronics), Seungwoo Seo (Samsung Advanced Institute of Technology, Samsung Electronics), Hosang Yoon (Samsung Advanced Institute of Technology, Samsung Electronics), Seungwon Lee (Samsung Advanced Institute of Technology, Samsung Electronics), Kyoungwhan Lim (Memory Business Division, Samsung Electronics), Hyunsung Shin (Memory Business Division, Samsung Electronics), Jinhyun Kim (Memory Business Division, Samsung Electronics), Seongil O (Memory Business Division, Samsung Electronics), Anand Iyer (Device Solutions America, Samsung Electronics), David Wang (Device Solutions America, Samsung Electronics), Kyomin Sohn (Memory Business Division, Samsung Electronics), Nam Sung Kim (Memory Business Division, Samsung Electronics)

[Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families57](#)

Samuel Naffziger (Advanced Micro Devices, Inc.), Noah Beck (Advanced Micro Devices, Inc.), Thomas Burd (Advanced Micro Devices, Inc.), Kevin Lepak (Advanced Micro Devices, Inc.), Gabriel H. Loh (Advanced Micro Devices, Inc.), Mahesh Subramony (Advanced Micro Devices, Inc.), Sean White (Advanced Micro Devices, Inc.)

Session 2A: Microarchitecture-1

[Zero Inclusion Victim: Isolating Core Caches from Inclusive Last-Level Cache Evictions71](#)

Mainak Chaudhuri (Indian Institute of Technology Kanpur)

[Exploiting Page Table Locality for Agile TLB Prefetching85](#)

Georgios Vavouliotis (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Lluç Alvarez (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya), Vasileios

Karakostas (National Technical University of Athens), Konstantinos Nikas (National Technical University of Athens), Nectarios Koziris (National Technical University of Athens), Daniel A. Jiménez (Texas A&M University), Marc Casas (Barcelona Supercomputing Center; Universitat Politècnica de Catalunya)

[A Cost-Effective Entangling Prefetcher for Instructions99](#)

Alberto Ros (University of Murcia, Spain), Alexandra Jimborean (University of Murcia, Spain)

Session 2B: Memory-1

[Don't Forget the I/O when Allocating Your LLC112](#)

Yifan Yuan (UIUC), Mohammad Alian (University of Kansas), Yipeng Wang (Intel), Ren Wang (Intel), Ilia Kurakin (Intel), Charlie Tai (Intel), Nam Sung Kim (UIUC)

[PF-DRAM: A Precharge-Free DRAM Structure126](#)

Nezam Rohbani (Institute for Research in Fundamental Sciences (IPM), Iran), Sina Darabi (Sharif University of Technology, Iran), Hamid Sarbazi-Azad (Institute for Research in Fundamental Sciences (IPM), Iran; Sharif University of Technology)

[Efficient Multi-GPU Shared Memory via Automatic Optimization of Fine-Grained Transfers139](#)

Harini Muthukrishnan (University of Michigan), David Nellans (NVIDIA), Daniel Lustig (NVIDIA), Jeffrey A. Fessler (University of Michigan), Thomas F. Wenisch (University of Michigan)

Session 3A: Machine Learning-1

[RaPiD: AI Accelerator for Ultra-low Precision Training and Inference153](#)

Swagath Venkataramani (IBM Research, Yorktown Heights), Vijayalakshmi Srinivasan (IBM Research, Yorktown Heights), Wei Wang (IBM Research, Yorktown Heights), Sanchari Sen (IBM Research, Yorktown Heights), Jintao Zhang (IBM Research, Yorktown Heights), Ankur Agrawal (IBM Research, Yorktown Heights), Monodeep Kar (IBM Research, Yorktown Heights), Shubham Jain (IBM Research, Yorktown Heights), Alberto Mannari (IBM Research, Switzerland), Hoang Tran (IBM Research, Yorktown Heights), Yulong Li (IBM Research, Yorktown Heights), Eri Ogawa (IBM Research, Japan), Kazuaki Ishizaki (IBM Research, Japan), Hiroshi Inoue (IBM Research, Japan), Marcel Schaal (IBM Research, Yorktown Heights), Mauricio Serrano (IBM Research, Yorktown Heights), Jungwook Choi (IBM Research, Yorktown Heights), Xiao Sun (IBM Research, Yorktown Heights), Naigang Wang (IBM Research, Yorktown Heights), Chia-Yu Chen (IBM Research, Yorktown Heights), Allison Allain (IBM Research, Yorktown Heights), James Bonanno (IBM, Austin), Nianzheng Cao (IBM Research, Yorktown Heights), Robert Casatuta (IBM, Hopewell Junction), Matthew Cohen (IBM Research, Yorktown Heights), Bruce Fleischer (IBM Research, Yorktown Heights), Michael Guillorn (IBM Research, Yorktown Heights), Howard Haynie (IBM, Poughkeepsie), Jinwook Jung (IBM Research, Yorktown Heights), Mingu Kang (IBM Research, Yorktown Heights), Kyu-hyoun Kim (IBM Research, Yorktown Heights), Siyu Koswatta (IBM Research, Yorktown Heights), Saekyu Lee (IBM Research, Yorktown Heights), Martin Lutz (IBM Research, Yorktown Heights), Silvia Mueller (IBM, Germany), Jinwook Oh (IBM Research,

Yorktown Heights), Ashish Ranjan (IBM Research, Yorktown Heights), Zhibin Ren (IBM Research, Yorktown Heights), Scot Rider (IBM, Poughkeepsie), Kerstin Schelm (IBM, Germany), Micheal Scheuermann (IBM Research, Yorktown Heights), Joel Silberman (IBM Research, Yorktown Heights), Jie Yang (IBM Research, Yorktown Heights), Vidhi Zalani (IBM Research, Yorktown Heights), Xin Zhang (IBM Research, Yorktown Heights), Ching Zhou (IBM Research, Yorktown Heights), Matt Ziegler (IBM Research, Yorktown Heights), Vinay Shah (IBM, United Kingdom), Moriyoshi Ohara (IBM Research, Japan), Pong-Fei Lu (IBM Research, Yorktown Heights), Brian Curran (IBM, Poughkeepsie), Sunil Shukla (IBM Research, Yorktown Heights), Leland Chang (IBM Research, Yorktown Heights), Kailash Gopalakrishnan (IBM Research, Yorktown Heights)

[REDUCT: Keep it Close, Keep it Cool! — Scaling DNN Inference on Multi-core CPUs with Near-Cache Compute167](#)

Anant Nori (Processor Architecture Research Lab, Intel Labs, Intel), Rahul Bera (ETH Zurich), Shankar Balachandran (Processor Architecture Research Labs Intel), Joydeep Rakshit (Processor Architecture Research Labs Intel), Omer Om (Processor Architecture Research Labs Intel), Avishai Abuhatzera (Intel), Belliappa Kuttanna (Intel), Sreenivas Subramoney (Processor Architecture Research Labs Intel)

[Communication Algorithm-Architecture Co-Design for Distributed Deep Learning181](#)

Jiayi Huang (UC Santa Barbara), Pritam Majumder (Texas A&M University), Sungkeun Kim (Texas A&M University), Abdullah Muzahid (Texas A&M University), Ki Hwan Yum (Texas A&M University), Eun Jung Kim (Texas A&M University)

Session 3B: Microarchitecture-2

[Vector Runahead195](#)

Ajeya Naithani (Ghent University), Sam Ainsworth (University of Edinburgh), Timothy M. Jones (University of Cambridge), Lieven Eeckhout (Ghent University)

[Unlimited Vector Extension with Data Streaming Support209](#)

Joao Mario Domingos (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal), Nuno Neves (INESC-ID, Instituto de Telecomunicações, Portugal), Nuno Roma (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal), Pedro Tomás (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal)

[Speculative Vectorisation with Selective Replay223](#)

Peng Sun (University of Cambridge, UK), Giacomo Gabrielli (Arm Research, UK), Timothy M. Jones (University of Cambridge, UK)

Session 4A: Processing in/near Memory

[ABC-DIMM: Alleviating the Bottleneck of Communication in DIMM-Based Near-Memory Processing with Inter-DIMM Broadcast](#)[237](#)

Weiyi Sun (Tsinghua University, China), Zhaoshi Li (Tsinghua University, China), Shouyi Yin (Tsinghua University, China), Shaojun Wei (Tsinghua University, China), Leibo Liu (Tsinghua University, China)

[Sieve: Scalable In-Situ DRAM-Based Accelerator Designs for Massively Parallel k-mer Matching](#)[251](#)

Lingxi Wu (University of Virginia), Rasool Sharifi (University of Virginia), Marzieh Lenjani (University of Virginia), Kevin Skadron (University of Virginia), Ashish Venkat (University of Virginia)

[FORMS: Fine-Grained Polarized ReRAM-Based In-Situ Computation for Mixed-Signal DNN Accelerator](#)[265](#)

Geng Yuan (Northeastern University), Payman Behnam (Georgia Institute of Technology), Zhengang Li (Northeastern University), Ali Shafiee (Samsung), Sheng Lin (Northeastern University), Xiaolong Ma (Northeastern University), Hang Liu (Stevens Institute of Technology), Xuehai Qian (University of Southern California), Mahdi Nazm Bojnordi (University of Utah), Yanzhi Wang (Northeastern University), Caiwen Ding (University of Connecticut)

[BOSS: Bandwidth-Optimized Search Accelerator for Storage-Class Memory](#)[279](#)

Jun Heo (Seoul National University), Seung Yul Lee (Seoul National University), Sunhong Min (Seoul National University), Yeonhong Park (Seoul National University), Sung Jun Jung (Seoul National University), Tae Jun Ham (Seoul National University), Jae W. Lee (Seoul National University)

Session 4B: Data Center

[SATORI: Efficient and Fair Resource Partitioning by Sacrificing Short-Term Benefits for Long-Term Gains](#)[292](#)

Rohan Basu Roy (Northeastern University), Tirthak Patel (Northeastern University), Devesh Tiwari (Northeastern University)

[Confidential Serverless Made Efficient with Plug-In Enclaves](#)[306](#)

Mingyu Li (Shanghai Jiao Tong University, China), Yubin Xia (Shanghai Jiao Tong University, China), Haibo Chen (Shanghai Jiao Tong University, China)

[Flex: High-Availability Datacenters with Zero Reserved Power](#)[319](#)

Chaojie Zhang (Microsoft Research, University of Chicago), Alok Gautam Kumbhare (Microsoft Research), Ioannis Manousakis (Microsoft Azure), Deli Zhang (Microsoft Research, Facebook), Pulkit A. Misra (Microsoft Research), Rod Assis (Microsoft Azure), Kyle Woolcock (Microsoft Azure), Nithish Mahalingam (Microsoft Azure), Brijesh Warriar (Microsoft Azure), David

Gauthier (Microsoft CO+I), Lalu Kunnath (Microsoft CO+I), Steve Solomon (Microsoft CO+I), Osvaldo Morales (Microsoft CO+I), Marcus Fontoura (Microsoft Azure), Ricardo Bianchini (Microsoft Research)

[BlockMaestro: Enabling Programmer-Transparent Task-Based Execution in GPU Systems333](#)

AmirAli Abdolrashidi (University of California, Riverside), Hodjat Asghari Esfeden (University of California, Riverside), Ali Jahanshahi (University of California, Riverside), Kaustubh Singh (University of California, Riverside), Nael Abu-Ghazaleh (University of California, Riverside), Daniel Wong (University of California, Riverside)

Session 5A: Security-1

[Opening Pandora's Box: A Systematic Study of New Ways Microarchitecture Can Leak Private Data347](#)

Jose Rodrigo Sanchez Vicarte (University of Illinois at Urbana-Champaign), Pradyumna Shome (University of Illinois at Urbana-Champaign), Nandeeka Nayak (University of Illinois at Urbana-Champaign), Caroline Trippel (Stanford University), Adam Morrison (Tel Aviv University), David Kohlbrenner (University of Washington), Christopher W. Fletcher (University of Illinois at Urbana-Champaign)

[I See Dead \$\mu\$ ops: Leaking Secrets via Intel/AMD Micro-Op Caches361](#)

Xida Ren (University of Virginia), Logan Moody (University of Virginia), Mohammadkazem Taram (University of California, San Diego), Matthew Jordan (University of Virginia), Dean M. Tullsen (University of California, San Diego), Ashish Venkat (University of Virginia)

[TimeCache: Using Time to Eliminate Cache Side Channels when Sharing Software375](#)

Divya Ojha (University of Rochester, USA), Sandhya Dwarkadas (University of Rochester, USA)

Session 5B: Accelerators-1

[Accelerated Seeding for Genome Sequence Alignment with Enumerated Radix Trees388](#)

Arun Subramaniyan (University of Michigan, USA), Jack Wadden (University of Michigan, USA), Kush Goliya (University of Michigan, USA), Nathan Ozog (University of Michigan, USA), Xiao Wu (University of Michigan, USA), Satish Narayanasamy (University of Michigan, USA), David Blaauw (University of Michigan, USA), Reetuparna Das (University of Michigan, USA)

[Aurochs: An Architecture for Dataflow Threads402](#)

Matthew Vilim (Stanford), Alexander Rucker (Stanford), Kunle Olukotun (Stanford)

[PipeZK: Accelerating Zero-Knowledge Proof with a Pipelined Architecture416](#)

Ye Zhang (Peking University; Shanghai Tree-Graph Blockchain Research Institute), Shuo Wang (Peking University), Xian Zhang (Microsoft Research), Jiangbin Dong (Xi'an Jiaotong University; Institute for Interdisciplinary Information Core Technology, Xi'an), Xingzhong Mao (Institute for Interdisciplinary Information Core Technology, Xi'an), Fan Long (University of Toronto), Cong Wang (International Digital Economy Academy at Guangdong-Hong Kong-Macau Greater Bay

Area), Dong Zhou (Tsinghua University), Mingyu Gao (Tsinghua University; Institute for Interdisciplinary Information Core Technology, Xi'an), Guangyu Sun (Peking University)

Session 6A: Compilers

[Taming the Zoo: The Unified GraphIT Compiler Framework for Novel Architectures429](#)

Ajay Brahmakshatriya (MIT CSAIL), Emily Furst (University of Washington), Victor A. Ying (MIT CSAIL), Claire Hsu (MIT CSAIL), Changwan Hong (MIT CSAIL), Max Ruttenberg (MIT CSAIL), Yunming Zhang (MIT CSAIL), Dai Cheol Jung (University of Washington), Dustin Richmond (University of Washington), Michael B. Taylor (University of Washington), Julian Shun (MIT CSAIL), Mark Oskin (University of Washington), Daniel Sanchez (MIT CSAIL), Saman Amarasinghe (MIT CSAIL)

[Supporting Legacy Libraries on Non-Volatile Memory: A User-Transparent Approach443](#)

Chencheng Ye (Huazhong University of Science and Technology, China), Yuanchao Xu (North Carolina State University, USA), Xipeng Shen (North Carolina State University, USA), Xiaofei Liao (Huazhong University of Science and Technology, China), Hai Jin (Huazhong University of Science and Technology, China), Yan Solihin (University of Central Florida, USA)

[Execution Dependence Extension \(EDE\): ISA Support for Eliminating Fences456](#)

Thomas Shull (Oracle Labs), Ilias Vougioukas (Arm Research), Nikos Nikoleris (Arm Research), Wendy Elsasser (Arm Research), Josep Torrellas (University of Illinois at Urbana-Champaign)

[Hetero-ViTAL: A Virtualization Stack for Heterogeneous FPGA Clusters470](#)

Yue Zha (University of Pennsylvania, USA), Jing Li (University of Pennsylvania, USA)

Session 6B: Memory-2

[CODIC: A Low-Cost Substrate for Enabling Custom In-DRAM Functionalities and Optimizations484](#)

Lois Orosa (ETH Zürich), Yaohua Wang (National University of Defense Technology (NUDT)), Mohammad Sadrosadati (Institute for Research in Fundamental Sciences (IPM); ETH Zürich), Jeremie S. Kim (ETH Zürich), Minesh Patel (ETH Zürich), Ivan Puddu (ETH Zürich), Haocong Luo (ETH Zürich), Kaveh Razavi (ETH Zürich), Juan Gómez-Luna (ETH Zürich), Hasan Hassan (ETH Zürich), Nika Mansouri-Ghiasi (ETH Zürich), Saugata Ghose (University of Illinois at Urbana-Champaign), Onur Mutlu (ETH Zürich)

[NVOverlay: Enabling Efficient and Scalable High-Frequency Snapshotting to NVM498](#)

Ziqi Wang (Carnegie Mellon University), Chul-Hwan Choo (Samsung Electronics), Michael A. Kozuch (Intel Labs), Todd C. Mowry (Carnegie Mellon University), Gennady Pekhimenko (University of Toronto), Vivek Seshadri (Microsoft Research India), Dimitrios Skarlatos (Carnegie Mellon University)

[Rebooting Virtual Memory with Midgard512](#)

Siddharth Gupta (EcoCloud, EPFL), Atri Bhattacharyya (EcoCloud, EPFL), Yunho Oh (Sungkyunkwan University), Abhishek Bhattacharjee (Yale University), Babak Falsafi (EcoCloud, EPFL), Mathias Payer (EcoCloud, EPFL)

[Dvé: Improving DRAM Reliability and Performance On-Demand via Coherent Replication526](#)

Adarsh Patil (University of Edinburgh), Vijay Nagarajan (University of Edinburgh), Rajeev Balasubramonian (University of Utah), Nicolai Oswald (University of Edinburgh)

Session 7A: Accelerators-2

[Enabling Compute-Communication Overlap in Distributed Deep Learning Training Platforms540](#)

Saeed Rashidi (Georgia Institute of Technology, USA), Matthew Denton (Georgia Institute of Technology, USA), Srinivas Sridharan (Facebook, USA), Sudarshan Srinivasan (Intel, India), Amoghavarsha Suresh (Stony Brook University, USA), Jade Nie (Facebook, USA), Tushar Krishna (Georgia Institute of Technology, USA)

[CoSA: Scheduling by Constrained Optimization for Spatial Accelerators554](#)

Qijing Huang (UC Berkeley), Minwoo Kang (UC Berkeley), Grace Dinh (UC Berkeley), Thomas Norell (UC Berkeley), Aravind Kalaiah (Facebook), James Demmel (UC Berkeley), John Wawrzynek (UC Berkeley), Yakun Sophia Shao (UC Berkeley)

[η-LSTM: Co-Designing Highly-Efficient Large LSTM Training via Exploiting Memory-Saving and Architectural Design Opportunities567](#)

Xingyao Zhang (University of Houston and University of Washington, USA), Haojun Xia (University of Sydney, Australia), Donglin Zhuang (University of Sydney, Australia), Hao Sun (University of Sydney, Australia), Xin Fu (University of Houston, USA), Michael B. Taylor (University of Washington, USA), Shuaiwen Leon Song (University of Sydney, Australia)

Session 7B: Graph Processing

[FlexMiner: A Pattern-Aware Accelerator for Graph Pattern Mining581](#)

Xuhao Chen (Massachusetts Institute of Technology), Tianhao Huang (Massachusetts Institute of Technology), Shuotao Xu (Massachusetts Institute of Technology), Thomas Bourgeat (Massachusetts Institute of Technology), Chanwoo Chung (Massachusetts Institute of Technology), Arvind Arvind (Massachusetts Institute of Technology)

[PolyGraph: Exposing the Value of Flexibility for Graph Processing Accelerators595](#)

Vidushi Dadu (University of California, Los Angeles), Sihao Liu (University of California, Los Angeles), Tony Nowatzki (University of California, Los Angeles)

[Large-Scale Graph Processing on FPGAs with Caches for Thousands of Simultaneous Misses609](#)

Mikhail Asiatiki (Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland), Paolo lenne (Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland)

Session 8A: Low Temperature/Low Energy Computing

[Cost-Efficient Overclocking in Immersion-Cooled Datacenters623](#)

Majid Jalili (Microsoft Research), Ioannis Manousakis (Microsoft Azure), Íñigo Goiri (Microsoft Research), Pulkit A. Misra (Microsoft Research), Ashish Raniwala (Microsoft Azure), Husam Alissa (Microsoft CO+I), Bharath Ramakrishnan (Microsoft CO+I), Phillip Tuma (3M), Christian Belady (Microsoft CO+I), Marcus Fontoura (Microsoft Azure), Ricardo Bianchini (Microsoft Research)

[CryoGuard: A Near Refresh-Free Robust DRAM Design for Cryogenic Computing637](#)

Gyu-Hyeon Lee (Seoul National University), Seongmin Na (Seoul National University), Ilkwon Byun (Seoul National University), Dongmoon Min (Seoul National University), Jangwoo Kim (Seoul National University)

[Superconducting Computing with Alternating Logic Elements651](#)

Georgios Tzimpragos (UC Santa Barbara), Jennifer Volk (UC Santa Barbara), Alex Wynn (MIT Lincoln Laboratory), James E. Smith (University of Wisconsin-Madison), Timothy Sherwood (UC Santa Barbara)

[Failure Sentinels: Ubiquitous Just-in-Time Intermittent Computation via Low-Cost Hardware Support for Voltage Monitoring665](#)

Harrison Williams (Virginia Tech, USA), Michael Moukarzel (Virginia Tech, USA), Matthew Hicks (Virginia Tech, USA)

Session 8B: Machine Learning-2

[SPACE: Locality-Aware Processing in Heterogeneous Memory for Personalized Recommendations679](#)

Hongju Kal (Yonsei University), Seokmin Lee (Yonsei University), Gun Ko (Yonsei University), Won Woo Ro (Yonsei University)

[ELSA: Hardware-Software Co-Design for Efficient, Lightweight Self-Attention Mechanism in Neural Networks692](#)

Tae Jun Ham (Seoul National University), Yejin Lee (Seoul National University), Seong Hoon Seo (Seoul National University), Soosung Kim (Seoul National University), Hyunji Choi (Seoul National University), Sung Jun Jung (Seoul National University), Jae W. Lee (Seoul National University)

[Cambricon-Q: A Hybrid Architecture for Efficient Training706](#)

Yongwei Zhao (Institute of Computing Technology, CAS; University of Chinese Academy of Sciences; Cambricon Tech. Ltd), Chang Liu (Institute of Computing Technology, CAS; University of Chinese Academy of Sciences; Cambricon Tech. Ltd), Zidong Du (Institute of Computing Technology, CAS; Cambricon Tech. Ltd), Qi Guo (Institute of Computing Technology, CAS), Xing Hu (Institute of Computing Technology, CAS), Yimin Zhuang (Institute of Computing Technology,

CAS; University of Chinese Academy of Sciences; Cambricon Tech. Ltd), Zhenxing Zhang (Institute of Computing Technology, CAS; University of Chinese Academy of Sciences; Cambricon Tech. Ltd), Xinkai Song (Institute of Computing Technology, CAS; University of Chinese Academy of Sciences; Cambricon Tech. Ltd), Wei Li (Institute of Computing Technology, CAS), Xishan Zhang (Institute of Computing Technology, CAS; Cambricon Tech. Ltd), Ling Li (Institute of Software, CAS), Zhiwei Xu (Institute of Computing Technology, CAS; University of Chinese Academy of Sciences), Tianshi Chen (Cambricon Tech. Ltd)

[TENET: A Framework for Modeling Tensor Dataflow Based on Relation-Centric Notation720](#)

Liqiang Lu (Peking University), Naiqing Guan (Peking University, University of Toronto), Yuyue Wang (Peking University), Liancheng Jia (Peking University), Zizhang Luo (Peking University), Jieming Yin (Lehigh University), Jason Cong (University of California, Los Angeles), Yun Liang (Peking University)

Session 9A: Memory-3

[Ripple: Profile-Guided Instruction Cache Replacement for Data Center Applications734](#)

Tanvir Ahmed Khan (University of Michigan), Dexin Zhang (University of Science and Technology of China), Akshitha Sriraman (University of Michigan), Joseph Devietti (University of Pennsylvania), Gilles Pokam (Intel Corporation), Heiner Litz (University of California, Santa Cruz), Baris Kasikci (University of Michigan)

[Quantifying Server Memory Frequency Margin and Using It to Improve Performance in HPC Systems748](#)

Da Zhang (Virginia Tech), Gagandeep Panwar (Virginia Tech), Jagadish B. Kotra (AMD Research), Nathan DeBardleben (Los Alamos National Laboratory), Sean Blanchard (Los Alamos National Laboratory), Xun Jian (Virginia Tech)

[Revamping Storage Class Memory with Hardware Automated Memory-Over-Storage Solution762](#)

Jie Zhang (Korea Advanced Institute of Science and Technology (KAIST)), Miryeong Kwon (Korea Advanced Institute of Science and Technology (KAIST)), Donghyun Gouk (Korea Advanced Institute of Science and Technology (KAIST)), Sungjoon Koh (Korea Advanced Institute of Science and Technology (KAIST)), Nam Sung Kim (University of Illinois Urbana-Champaign), Mahmut Taylan Kandemir (Pennsylvania State University), Myoungsoo Jung (Korea Advanced Institute of Science and Technology (KAIST))

Session 9B: Network Storage and Acceleration

[NASGuard: A Novel Accelerator Architecture for Robust Neural Architecture Search \(NAS\) Networks776](#)

Xingbin Wang (Institute of Information Engineering, CAS, China; University of Chinese Academy of Sciences, China), Boyan Zhao (Institute of Information Engineering, CAS, China), Rui Hou (Institute of Information Engineering, CAS, China), Amro Awad (NC State University, USA),

Zhihong Tian (Guangzhou University, China), Dan Meng (Institute of Information Engineering, CAS, China)

[NASA: Accelerating Neural Network Design with a NAS Processor790](#)

Xiaohan Ma (Institute of Computing Technology, CAS; University of Chinese Academy of Sciences), Chang Si (Institute of Computing Technology, CAS; University of Chinese Academy of Sciences), Ying Wang (Institute of Computing Technology, CAS), Cheng Liu (Institute of Computing Technology, CAS), Lei Zhang (Institute of Computing Technology, CAS)

[PMNet: In-Network Data Persistence804](#)

Korakit Seemakhupt (University of Virginia), Sihang Liu (University of Virginia), Yasas Senevirathne (University of Virginia), Muhammad Shahbaz (Stanford University; Perdue University), Samira Khan (University of Virginia)

Session 10A: Quantum/Photonics

[Exploiting Long Distance Interactions and Tolerating Atom Loss in Neutral Atom Quantum Architectures818](#)

Jonathan M. Baker (University of Chicago, USA), Andrew Littken (University of Chicago, USA), Casey Duckering (University of Chicago, USA), Henry Hoffmann (University of Chicago, USA), Hannes Bernien (University of Chicago, USA), Frederic T. Chong (University of Chicago, USA)

[Software-Hardware Co-Optimization for Computational Chemistry on Superconducting Quantum Processors832](#)

Gushu Li (University of California, Santa Barbara), Yunong Shi (University of Chicago, USA), Ali Javadi-Abhari (IBM Quantum, T. J. Watson Research Center, USA)

[Designing Calibration and Expressivity-Efficient Instruction Sets for Quantum Computing846](#)

Lingling Lao (University College London), Prakash Murali (Princeton University), Margaret Martonosi (Princeton University), Dan Browne (University College London)

[Albireo: Energy-Efficient Acceleration of Convolutional Neural Networks via Silicon Photonics860](#)

Kyle Shiflett (Ohio University), Avinash Karanth (Ohio University), Razvan Bunescu (The University of North Carolina at Charlotte), Ahmed Louri (George Washington University)

Session 10B: Reliability and Security

[IntroSpectre: A Pre-Silicon Framework for Discovery and Analysis of Transient Execution Vulnerabilities874](#)

Moein Ghaniyoun (The Ohio State University, USA), Kristin Barber (The Ohio State University, USA), Yinqian Zhang (Southern University of Science and Technology, China), Radu Teodorescu (The Ohio State University, USA)

[Maya: Using Formal Control to Obfuscate Power Side Channels888](#)

Raghavendra Pradyumna Pothukuchi (University of Illinois at Urbana-Champaign), Sweta Yamini Pothukuchi (University of Illinois at Urbana-Champaign), Petros G. Voulgaris (University of Nevada, Reno), Alexander Schwing (University of Illinois at Urbana-Champaign), Josep Torrellas (University of Illinois at Urbana-Champaign)

[Demystifying the System Vulnerability Stack: Transient Fault Effects Across the Layers902](#)

George Papadimitriou (University of Athens, Greece), Dimitris Gizopoulos (University of Athens, Greece)

[No-FAT: Architectural Support for Low Overhead Memory Safety Checks916](#)

Mohamed Tarek Ibn Ziad (Columbia University), Miguel A. Arroyo (Columbia University), Evgeny Manzhosov (Columbia University), Ryan Piersma (Columbia University), Simha Sethumadhavan (Columbia University)

Session 11A: DRAM/IO/Network

[Ghost Routing to Enable Oblivious Computation on Memory-Centric Networks930](#)

Yeonju Ro (KAIST, South Korea), Seongwook Jin (KAIST, South Korea), Jaehyuk Huh (KAIST, South Korea), John Kim (KAIST, South Korea)

[QUAC-TRNG: High-Throughput True Random Number Generation Using Quadruple Row Activation in Commodity DRAM Chips944](#)

Ataberk Olgun (ETH Zürich; TOBB University of Economics and Technology), Minesh Patel (ETH Zürich), A. Giray Yağlıkçı (ETH Zürich), Haocong Luo (ETH Zürich), Jeremie S. Kim (ETH Zürich), F. Nisa Bostancı (ETH Zürich; TOBB University of Economics and Technology), Nandita Vijaykumar (ETH Zürich; University of Toronto), Oğuz Ergin (TOBB University of Economics and Technology), Onur Mutlu (ETH Zürich)

[A RISC-V in-Network Accelerator for Flexible High-Performance Low-Power Packet Processing958](#)

Salvatore Di Girolamo (ETH Zürich, Switzerland), Andreas Kurth (ETH Zürich, Switzerland), Alexandru Calotoiu (ETH Zürich, Switzerland), Thomas Benz (ETH Zürich, Switzerland), Timo Schneider (ETH Zürich, Switzerland), Jakub Beránek (Technical University of Ostrava), Luca Benini (ETH Zürich, Switzerland), Torsten Hoefler (ETH Zürich, Switzerland)

Session 11B: Security-2

[Leaky Buddies: Cross-Component Covert Channels on Integrated CPU-GPU Systems972](#)

Sankha Baran Dutta (University of California, USA), Hoda Naghibijouybari (Binghamton University, USA), Nael Abu-Ghazaleh (University of California, USA), Andres Marquez (Pacific Northwest National Laboratory, USA), Kevin Barker (Pacific Northwest National Laboratory, USA)

[ICannels: Exploiting Current Management Mechanisms to Create Covert Channels in Modern Processors985](#)

Jawad Haj-Yahya (ETH Zürich), Jeremie S. Kim (ETH Zürich), A. Giray Yaglıkçı (ETH Zürich), Ivan Puddu (ETH Zürich), Lois Orosa (ETH Zürich), Juan Gómez Luna (ETH Zürich), Mohammed Alser (ETH Zürich), Onur Mutlu (ETH Zürich)

[ZeRØ: Zero-Overhead Resilient Operation Under Pointer Integrity Attacks999](#)

Mohamed Tarek Ibn Ziad (Columbia University), Miguel A. Arroyo (Columbia University), Evgeny Manzhosov (Columbia University), Simha Sethumadhavan (Columbia University)

Session 12A: Accelerators-3

[NN-Baton: DNN Workload Orchestration and Chiplet Granularity Exploration for Multichip Accelerators1013](#)

Zhanhong Tan (Tsinghua University), Hongyu Cai (Tsinghua University), Runpei Dong (Xi'an Jiaotong University), Kaisheng Ma (Tsinghua University)

[SNAFU: An Ultra-Low-Power, Energy-Minimal CGRA-Generation Framework and Architecture1027](#)

Graham Gobieski (Carnegie Mellon University), Ahmet Oguz Atli (Carnegie Mellon University), Kenneth Mai (Carnegie Mellon University), Brandon Lucia (Carnegie Mellon University), Nathan Beckmann (Carnegie Mellon University)

[SARA: Scaling a Reconfigurable Dataflow Accelerator1041](#)

Yaqi Zhang (Stanford University), Nathan Zhang (Stanford University), Tian Zhao (Stanford University), Matt Vilim (Stanford University), Muhammad Shahbaz (Stanford University), Kunle Olukotun (Stanford University)

[HASCO: Towards Agile HARDware and Software CO-design for Tensor Computation1055](#)

Qingcheng Xiao (Peking University), Size Zheng (Peking University), Bingzhe Wu (Peking University), Pengcheng Xu (Peking University), Xuehai Qian (University of Southern California), Yun Liang (Peking University)

Session 12B: Sparse Processing

[SpZip: Architectural Support for Effective Data Compression In Irregular Applications1069](#)

Yifan Yang (Massachusetts Institute of Technology), Joel S. Emer (Massachusetts Institute of Technology; NVIDIA), Daniel Sanchez (Massachusetts Institute of Technology)

[Dual-Side Sparse Tensor Core1083](#)

Yang Wang (University of Electronic Science and Technology of China; Microsoft Research), Chen Zhang (Microsoft Research), Zhiqiang Xie (ShanghaiTech University), Cong Guo (Shanghai Jiao Tong University), Yunxin Liu (Microsoft Research), Jingwen Leng (Shanghai Jiao Tong University)

[RingCNN: Exploiting Algebraically-Sparse Ring Tensors for Energy-Efficient CNN-Based Computational Imaging](#)[1096](#)

Chao-Tsung Huang (National Tsing Hua University, Taiwan)

[GoSPA: An Energy-Efficient High-Performance Globally Optimized Sparse Convolutional Neural Network Accelerator](#)[1110](#)

Chunhua Deng (Rutgers University - New Brunswick, USA), Yang Sui (Rutgers University - New Brunswick, USA), Siyu Liao (Amazon Research, USA), Xuehai Qian (University of Southern California, USA), Bo Yuan (Rutgers University - New Brunswick, USA)