

2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA 2024)

**Edinburgh, United Kingdom
2-6 March 2024**

Pages 1-612



**IEEE Catalog Number: CFP24013-POD
ISBN: 979-8-3503-9314-9**

**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: | CFP24013-POD |
| ISBN (Print-On-Demand): | 979-8-3503-9314-9 |
| ISBN (Online): | 979-8-3503-9313-2 |
| ISSN: | 1530-0897 |

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2024 IEEE International Symposium on High- Performance Computer Architecture (HPCA) **HPCA 2024**

Table of Contents

| | |
|--|--------|
| Message from General Chairs | xviii |
| Message from Program Chairs | xix |
| Message from Industry Track Chairs | xxii |
| Organizing Committee | xxiii |
| Program Committee | xxiv |
| Industry Track Program Committee | xxviii |
| Sponsors | xxix |

Security: Side-Channel, Session 1/A

| | |
|--|----|
| Exploitation of Security Vulnerability on Retirement | 1 |
| <i>Ke Xu (Wuhan University, China), Ming Tang (Wuhan University, China), Quancheng Wang (Wuhan University, China), and Han Wang (Wuhan University, China)</i> | |
| GadgetSpinner: A New Transient Execution Primitive using the Loop Stream Detector | 15 |
| <i>Yun Chen (National University of Singapore, Singapore), Ali Hajiabadi (National University of Singapore, Singapore), and Trevor E. Carlson (National University of Singapore, Singapore)</i> | |
| Uncovering and Exploiting AMD Speculative Memory Access Predictors for Fun and Profit | 31 |
| <i>Chang Liu (Tsinghua University, China), Dongsheng Wang (Tsinghua University, Zhongguancun Laboratory, China), Yongqiang Lyu (Tsinghua University, China), Pengfei Qiu (Beijing University of Posts and Telecommunications, China), Yu Jin (Beijing University of Posts and Telecommunications, China), Zhuoyuan Lu (Beijing University of Posts and Telecommunications, China), Yinqian Zhang (Southern University of Science and Technology, China), and Gang Qu (University of Maryland, College Park, USA)</i> | |

Reconfigurable Architecture & FPGA, 1/B

| | |
|--|----|
| E2EMap: End-to-End Reinforcement Learning for CGRA Compilation via Reverse Mapping | 46 |
| <i>Dajiang Liu (Chongqing University, China), Yuxin Xia (Chongqing University, China), Jiaying Shang (Chongqing University, China), Jiang Zhong (Chongqing University, China), Peng Ouyang (TsingMicro Co. Ltd., China), and Shouyi Yin (Tsinghua University, China)</i> | |
| Revet: A Language and Compiler for Dataflow Threads | 61 |
| <i>Alexander Rucker (Stanford University), Shiv Sundram (Stanford University), Coleman Smith (Stanford University), Matthew Vilim (SambaNova Systems, Inc.), Raghu Prabhakar (SambaNova Systems, Inc.), Fredrik Kjolstad (Stanford University), and Kunle Olukotun (Stanford University)</i> | |
| An Optimizing Framework on MLIR for Efficient FPGA-Based Accelerator Generation | 75 |
| <i>Weichuang Zhang (Shanghai Jiao Tong University, China), Jieru Zhao (Shanghai Jiao Tong University, China), Guan Shen (Shanghai Jiao Tong University, China), Quan Chen (Shanghai Jiao Tong University, China), Chen Chen (Shanghai Jiao Tong University, China), and Minyi Guo (Shanghai Jiao Tong University, China)</i> | |

GNN, Session 1/C

| | |
|---|-----|
| Celeritas: Out-of-Core Based Unsupervised Graph Neural Network via Cross-Layer Computing 2024 | 91 |
| <i>Yi Li (University of Texas at Dallas), Tsun-Yu Yang (The Chinese University of Hong Kong), Ming-Chang Yang (The Chinese University of Hong Kong), Zhaoyan Shen (Shandong University), and Bingzhe Li (University of Texas at Dallas)</i> | |
| PruneGNN: Algorithm-Architecture Pruning Framework for Graph Neural Network Acceleration | 108 |
| <i>Deniz Gurevin (University of Connecticut, USA), Mohsin Shan (University of Connecticut, USA), Shaoyi Huang (University of Connecticut, USA), MD Amit Hasan (University of Connecticut, USA), Caiwen Ding (University of Connecticut, USA), and Omer Khan (University of Connecticut, USA)</i> | |
| MEGA: A Memory-Efficient GNN Accelerator Exploiting Degree-Aware Mixed-Precision Quantization | 124 |
| <i>Zeyu Zhu (Institute of Automation, Chinese Academy of Sciences, China; School of Future Technology, University of Chinese Academy of Sciences, China), Fanrong Li (Institute of Automation, Chinese Academy of Sciences, China), Gang Li (Shanghai Jiao Tong University, China), Zejian Liu (Institute of Automation, Chinese Academy of Sciences, China), Zitao Mo (Institute of Automation, Chinese Academy of Sciences, China), Qinghao Hu (Institute of Automation, Chinese Academy of Sciences, China), Xiaoyao Liang (Shanghai Jiao Tong University, China), and Jian Cheng (Institute of Automation, Chinese Academy of Sciences, China; School of Future Technology, University of Chinese Academy of Sciences, China; AiRiA, China; Maicro.ai, China)</i> | |

Accelerators & Memory Systems, Session 2/A

| | |
|--|-----|
| Bandwidth-Effective DRAM Cache for GPUs with Storage-Class Memory | 139 |
| <i>Jeongmin Hong (POSTECH, Republic of Korea), Sungjun Cho (POSTECH, Republic of Korea), Geonwoo Park (POSTECH, Republic of Korea), Wonhyuk Yang (POSTECH, Republic of Korea), Young-Ho Gong (Soongsil University, Republic of Korea), and Gwangsun Kim (POSTECH, Republic of Korea)</i> | |
| Gemini: Mapping and Architecture Co-Exploration for Large-Scale DNN Chiplet Accelerators | 156 |
| <i>Jingwei Cai (Tsinghua University), Zuotong Wu (Xi'an Jiaotong University; Institute for Interdisciplinary Information Core Technology), Sen Peng (Xi'an Jiaotong University; Institute for Interdisciplinary Information Core Technology), Yuchen Wei (Tsinghua University), Zhanhong Tan (Tsinghua University), Guiming Shi (Tsinghua University), Mingyu Gao (Tsinghua University; Shanghai AI Laboratory), and Kaisheng Ma (Tsinghua University)</i> | |
| STELLAR: Energy-Efficient and Low-Latency SNN Algorithm and Hardware Co-Design with Spatiotemporal Computation | 172 |
| <i>Ruixin Mao (University of Electronic Science and Technology of China, China), Lin Tang (University of Electronic Science and Technology of China, China), Xingyu Yuan (University of Electronic Science and Technology of China, China), Ye Liu (University of Electronic Science and Technology of China, China), and Jun Zhou (University of Electronic Science and Technology of China, China)</i> | |
| MIMDRAM: An End-to-End Processing-Using-DRAM System for High-Throughput, Energy-Efficient and Programmer-Transparent Multiple-Instruction Multiple-Data Computing | 186 |
| <i>Geraldo F Oliveira (ETH Zurich), Ataberk Olgun (ETH Zurich), Abdullah Giray Yaglikci (ETH Zurich), F. Nisa Bostanci (ETH Zurich), Juan Gómez-Luna (ETH Zurich), Saugata Ghose (University of Illinois Urbana-Champaign), and Onur Mutlu (ETH Zurich)</i> | |

Security, Session 2/B

| | |
|--|-----|
| Supporting Secure Multi-GPU Computing with Dynamic and Batched Metadata Management | 204 |
| <i>Seonjin Na (Georgia Institute of Technology), Jungwoo Kim (KAIST), Sunho Lee (KAIST), and Jaehyuk Huh (KAIST)</i> | |
| Data Enclave: A Data-Centric Trusted Execution Environment | 218 |
| <i>Yuanchao Xu (University of California Santa Cruz), James Pangia (North Carolina State University), Chencheng Ye (Huazhong University of Science and Technology), Yan Solihin (University of Central Florida), and Xipeng Shen (North Carolina State University)</i> | |
| Salus: Efficient Security Support for CXL-Expanded GPU Memory | 233 |
| <i>Rahaf Abdullah (North Carolina State University, USA), Hyokeun Lee (North Carolina State University, USA), Huiyang Zhou (North Carolina State University, USA), and Amro Awad (North Carolina State University, USA)</i> | |

| | |
|--|-----|
| Morphling: A Throughput-Maximized TFHE-Based Accelerator using Transform-Domain Reuse .. | 249 |
| <i>Prasetyo Prasetyo (Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea), Adiwena Putra (Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea), and Joo-Young Kim (Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea)</i> | |

Near Memory Processing, Session 2/C

| | |
|--|-----|
| Pathfinding Future PIM Architectures by Demystifying a Commercial PIM Technology | 263 |
| <i>Bongjoon Hyun (KAIST), Taehun Kim (KAIST), Dongjae Lee (KAIST), and Minsoo Rhu (KAIST)</i> | |
| Functionally-Complete Boolean Logic in Real DRAM Chips: Experimental Characterization and Analysis | 280 |
| <i>Ismail Emir Yuksel (ETH Zurich), Yahya Can Tugrul (ETH Zurich), Ataberk Olgun (ETH Zurich), F. Nisa Bostanci (ETH Zurich), A. Giray Yaglikci (ETH Zurich), Geraldo F. Oliveira (ETH Zurich), Haocong Luo (ETH Zurich), Juan Gómez-Luna (ETH Zurich), Mohammad Sadrosadati (ETH Zurich), and Onur Mutlu (ETH Zurich)</i> | |
| StreamPIM: Streaming Matrix Computation in Racetrack Memory | 297 |
| <i>Yuda An (Peking University, China), Yunxiao Tang (Peking University, China), Shushu Yi (Peking University, China), Li Peng (Peking University, China), Xiurui Pan (Peking University, China), Guangyu Sun (Peking University, China; Beijing Advanced Innovation Center for Integrated Circuits, China), Zhaochu Luo (Peking University, China), Qiao Li (Xiamen University, China), and Jie Zhang (Peking University, China)</i> | |
| SmartDIMM: In-Memory Acceleration of Upper Layer Protocols | 312 |
| <i>Neel Patel (University of Kansas), Amin Mamandipoor (University of Kansas), Mohammad Nouri (University of Kansas), and Mohammad Alian (University of Kansas)</i> | |

Computational SSD, Session 3/A

| | |
|---|-----|
| BeaconGNN: Large-Scale GNN Acceleration with Out-of-Order Streaming In-Storage Computing .. | 330 |
| <i>Yuyue Wang (UCLA), Xiurui Pan (Peking University), Yuda An (Peking University), Jie Zhang (Peking University), and Glenn Reinman (UCLA)</i> | |
| Smart-Infinity: Fast Large Language Model Training using Near-Storage Processing on a Real System | 345 |
| <i>Hongsun Jang (Seoul National University, Republic of Korea), Jaeyong Song (Seoul National University, Republic of Korea), Jaewon Jung (Seoul National University, Republic of Korea), Jaeyoung Park (University of Texas at Austin, United States of America), Youngsok Kim (Yonsei University, Republic of Korea), and Jinho Lee (Seoul National University, Republic of Korea)</i> | |

| | |
|---|-----|
| FlashGNN: An In-SSD Accelerator for GNN Training | 361 |
| <i>Fuping Niu (Huazhong University of Science and Technology, China), Jianhui Yue (Michigan Technological University, USA), Jiangqiu Shen (Michigan Technological University, USA), Xiaofei Liao (Huazhong University of Science and Technology, China), and Hai Jin (Huazhong University of Science and Technology, China)</i> | |
| DockerSSD: Containerized In-Storage Processing and Hardware Acceleration for Computational SSDs | 379 |
| <i>Donghyun Gouk (KAIST, South Korea), Miryeong Kwon (KAIST and Panmnesia, South Korea), Hanyeoreum Bae (KAIST, South Korea), and Myoungsoo Jung (KAIST and Panmnesia, South Korea)</i> | |

Side-Channel & Microarchitecture, Session 3/B

| | |
|--|-----|
| PrefetchX: Cross-Core Cache-Agnostic Prefetcher-Based Side-Channel Attacks | 395 |
| <i>Yun Chen (National University of Singapore, Singapore), Ali Hajiabadi (National University of Singapore, Singapore), Lingfeng Pei (National University of Singapore, Singapore), and Trevor E. Carlson (National University of Singapore, Singapore)</i> | |
| Modeling, Derivation, and Automated Analysis of Branch Predictor Security Vulnerabilities | 409 |
| <i>Quancheng Wang (Wuhan University, China), Ming Tang (Wuhan University, China), Ke Xu (Wuhan University, China), and Han Wang (Wuhan University, China)</i> | |
| SegScope: Probing Fine-Grained Interrupts via Architectural Footprints | 424 |
| <i>Xin Zhang (Peking University, China), Zhi Zhang (The University of Western Australia, Australia), Qingni Shen (Peking University, China), Wenhao Wang (Institute of Information Engineering, China), Yansong Gao (Data61, Australia), Zhuoxi Yang (Peking University, China), and Jiliang Zhang (Hunan University, China)</i> | |
| Differential-Matching Prefetcher for Indirect Memory Access | 439 |
| <i>Gelin Fu (Xi'an Jiaotong University), Tian Xia (Xi'an Jiaotong University), Zhongpei Luo (Xi'an Jiaotong University), Ruiyang Chen (Xi'an Jiaotong University), Wenzhe Zhao (Xi'an Jiaotong University), and Pengju Ren (Xi'an Jiaotong University)</i> | |

Accelerator (non-DNN), Session 3/C

| | |
|---|-----|
| SPADE: Sparse Pillar-Based 3D Object Detection Accelerator for Autonomous Driving | 454 |
| <i>Minjae Lee (Hanyang University), Seongmin Park (Hanyang University), Hyungmin Kim (Hanyang University), Minyong Yoon (Hanyang University), Janghwan Lee (Hanyang University), Junwon Choi (Hanyang University), Nam Sung Kim (University of Illinois at Urbana-Champaign), Mingu Kang (University of California San Diego), and Jungwook Choi (Hanyang University)</i> | |
| Rapper: A Parameter-Aware Repair-in-Memory Accelerator for Blockchain Storage Platform | 468 |
| <i>Chenlin Ma (Shenzhen University, China), Yingping Wang (Shenzhen University, China), Fuwen Chen (Shenzhen University, China), Jing Liao (Shenzhen University, China), Yi Wang (Shenzhen University, China), and Rui Mao (Shenzhen University, China)</i> | |

| | |
|--|-----|
| MOPED: Efficient Motion Planning Engine with Flexible Dimension Support | 483 |
| <i>Lingyi Huang (Rutgers University, USA), Yu Gong (Rutgers University, USA), Yang Sui (Rutgers University, USA), Xiao Zang (Rutgers University, USA), and Bo Yuan (Rutgers University, USA)</i> | |
| TALCO: Tiling Genome Sequence Alignment using Convergence of Traceback Pointers | 498 |
| <i>Sumit Walia (University of California San Diego, USA), Cheng Ye (University of California San Diego, USA), Arkid Bera (University of California San Diego, USA), Dhruvi Lodhavia (University of California San Diego, USA), and Yatish Turakhia (University of California San Diego, USA)</i> | |

Microarchitecture, Session 5/A

| | |
|--|-----|
| Effective Context-Sensitive Memory Dependence Prediction | 515 |
| <i>Sebastian S. Kim (University of Murcia, Spain) and Alberto Ros (University of Murcia, Spain)</i> | |
| A Two Level Neural Approach Combining Off-Chip Prediction with Adaptive Prefetch Filtering... 528 | |
| <i>Alexandre Valentin Jamet (Universitat Politecnica de Catalunya / Barcelona Supercomputing Center), Georgios Vavouliotis (Huawei Zurich Research Center), Daniel A. Jiménez (Texas A&M University), Lluc Alvarez (Universitat Politecnica de Catalunya / Barcelona Supercomputing Center), and Marc Casas (Universitat Politecnica de Catalunya / Barcelona Supercomputing Center)</i> | |
| gem5-MARVEL: Microarchitecture-Level Resilience Analysis of Heterogeneous SoC Architectures | 543 |
| <i>Odysseas Chatzopoulos (University of Athens, Greece), George Papadimitriou (University of Athens, Greece), Vasileios Karakostas (University of Athens, Greece), and Dimitris Gizopoulos (University of Athens, Greece)</i> | |

Rowhammer, Session 5/B

| | |
|--|-----|
| Spatial Variation-Aware Read Disturbance Defenses: Experimental Analysis of Real DRAM Chips and Implications on Future Solutions | 560 |
| <i>Abdullah Giray Yağlıkçı (ETH Zürich), Yahya Can Tuğrul (ETH Zürich), Geraldo Francisco de Oliveira (ETH Zurich), Ismail Emir Yüksel (ETH Zürich), Ataberk Olgun (ETH Zurich), Haocong Luo (ETH Zurich), and Onur Mutlu (ETH Zurich)</i> | |
| START: Scalable Tracking for Any Rowhammer Threshold | 578 |
| <i>Anish Saxena (Georgia Institute of Technology, USA) and Moinuddin Qureshi (Georgia Institute of Technology, USA)</i> | |
| CoMeT: Count-Min-Sketch-Based Row Tracking to Mitigate RowHammer at Low Cost | 593 |
| <i>Fatma Nisa Bostancı (ETH Zürich), İsmail Emir Yüksel (ETH Zürich), Ataberk Olgun (ETH Zürich), Konstantinos Kanellopoulos (ETH Zürich), Yahya Can Tuğrul (ETH Zürich), Abdullah Giray Yağlıkçı (ETH Zürich), Mohammad Sadrosadati (ETH Zürich), and Onur Mutlu (ETH Zürich)</i> | |

Best of CAL, Session 5/C

| | |
|--|-----|
| A Quantum Computer Trusted Execution Environment | 613 |
| Unleashing the Potential of PIM: Accelerating Large Batched Inference of Transformer-Based Generative Models | 614 |
| Computational CXL-Memory Solution for Accelerating Memory-Intensive Applications | 615 |

SSD, Session 6/A

| | |
|--|-----|
| LearnedFTL: A Learning-Based Page-Level FTL for Reducing Double Reads in Flash-Based SSDs .. | 616 |
| <i>Shengzhe Wang (Xiamen University, China), Zihang Lin (Xiamen University, China), Suzhen Wu (Xiamen University, China), Hong Jiang (University of Texas at Arlington, USA), Jie Zhang (Peking University, China), and Bo Mao (Xiamen University, China)</i> | |
| Are Superpages Super-Fast? Distilling Flash Blocks to Unify Flash Pages of a Superpage in an SSD | 630 |
| <i>Shih-Hung Tseng (National Central University, Taiwan), Tseng-Yi Chen (National Central University, Taiwan), and Ming-Chang Yang (The Chinese University of Hong Kong)</i> | |
| RiF: Improving Read Performance of Modern SSDs Using an On-Die Early-Retry Engine | 643 |
| <i>Myoungjun Chun (Seoul National University, Korea), Jaeyong Lee (Seoul National University, Korea), Myungsuk Kim (Kyungpook National University, Korea), Jisung Park (POSTECH, Korea), and Jihong Kim (Seoul National University, Korea)</i> | |
| Midas Touch: Invalid-Data Assisted Reliability and Performance Boost for 3D High-Density Flash | 657 |
| <i>Qiao Li (Xiamen University), Hongyang Dang (Xiamen University), Zheng Wan (Xiamen University), Congming Gao (Xiamen University), Min Ye (City University of Hong Kong), Jie Zhang (Peking University), Tei-Wei Kuo (National Taiwan University), and Chun Jason Xue (Mohamed bin Zayed University of Artificial Intelligence)</i> | |

Emerging Technology, Session 6/B

| | |
|---|-----|
| ECO-CHIP: Estimation of Carbon Footprint of Chiplet-Based Architectures for Sustainable VLSI | 671 |
| <i>Chetan Choppali Sudarshan (Arizona State University, USA), Nikhil Matkar (Arizona State University, USA), Sarma Vrudhula (Arizona State University, USA), Sachin S. Sapatnekar (University of Minnesota, USA), and Vidya A. Chhabria (Arizona State University, USA)</i> | |

| | |
|---|-----|
| Lightening-Transformer: A Dynamically-Operated Optically-Interconnected Photonic Transformer Accelerator | 686 |
| <i>Hanqing Zhu (The University of Texas at Austin, USA), Jiaqi Gu (The University of Texas at Austin, USA; Arizona State University, USA), Hanrui Wang (Massachusetts Institute of Technology, USA), Zixuan Jiang (The University of Texas at Austin, USA), Zhekai Zhang (Massachusetts Institute of Technology, USA), Rongxing Tang (The University of Texas at Austin, USA), Chenghao Feng (The University of Texas at Austin, USA), Song Han (Massachusetts Institute of Technology, USA), Ray T. Chen (The University of Texas at Austin, USA), and David Z. Pan (The University of Texas at Austin, USA)</i> | |
| MIRAGE: Quantum Circuit Decomposition and Routing Collaborative Design using Mirror Gates | 704 |
| <i>Evan McKinney (University of Pittsburgh), Michael Hatridge (University of Pittsburgh), and Alex K. Jones (University of Pittsburgh)</i> | |
| SACHI: A Stationarity-Aware, All-Digital, Near-Memory, Ising Architecture | 719 |
| <i>Siddhartha Raman Sundara Raman (The University of Texas At Austin), Lizy K. John (The University of Texas at Austin), and Jaydeep P. Kulkarni (The University of Texas at Austin)</i> | |

Accelerator, Session 6/C

| | |
|--|-----|
| BitWave: Exploiting Column-Based Bit-Level Sparsity for Deep Learning Acceleration | 732 |
| <i>Man Shi (KU Leuven, Belgium), Vikram Jain (KU Leuven, Belgium), Antony Joseph (NXP Semiconductor, Belgium), Maurice Meijer (NXP Semiconductor, Belgium), and Marian Verhelst (KU Leuven, Belgium)</i> | |
| LUTEin: Dense-Sparse Bit-Slice Architecture with Radix-4 LUT-Based Slice-Tensor Processing Units | 747 |
| <i>Dongseok Im (Korea Advanced Institute of Science and Technology) and Hoi-Jun Yoo (Korea Advanced Institute of Science and Technology)</i> | |
| FIGNA: Integer Unit-Based Accelerator Design for FP-INT GEMM Preserving Numerical Accuracy..... | 760 |
| <i>Jaeyong Jang (Seoul National University, Korea), Yulhwa Kim (Sungkyunkwan University, Korea), Juheun Lee (Seoul National University, Korea), and Jae-Joon Kim (Seoul National University, Korea)</i> | |
| ASADI: Accelerating Sparse Attention using Diagonal-Based In-Situ Computing | 774 |
| <i>Huize Li (National University of Singapore, Singapore), Zhaoying Li (National University of Singapore, Singapore), Zhenyu Bai (National University of Singapore, Singapore), and Tulika Mitra (National University of Singapore, Singapore)</i> | |

Distributed DNN & Training, 7/A

- Enabling Large Dynamic Neural Network Training with Learning-Based Memory Management ... 788
Jie Ren (William & Mary), Dong Xu (University of California, Merced), Shuangyan Yang (University of California, Merced), Jiacheng Zhao (University of Chinese Academy of Sciences), Zhicheng Li (University of Chinese Academy of Sciences), Christian Navasca (University of California, Los Angeles), Chenxi Wang (Chinese Academy of Sciences), Harry Xu (University of California, Los Angeles), and Dong Li (University of California, Merced)
- Tessel: Boosting Distributed Execution of Large DNN Models via Flexible Schedule Search 803
Zhiqi Lin (University of Science and Technology of China, China), Youshan Miao (Microsoft Research, China), Guanbin Xu (University of Science and Technology of China, China), Cheng Li (University of Science and Technology of China, China), Olli Saarikivi (Microsoft Research, United States), Saeed Maleki (Microsoft Research, United States), and Fan Yang (Microsoft Research, China)
- SpecFL: An Efficient Speculative Federated Learning System for Tree-Based Model Training 817
Yuhui Zhang (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS and School of Cyber Security, University of Chinese Academy of Sciences, China), Lutan Zhao (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS and School of Cyber Security, University of Chinese Academy of Sciences, China), Cheng Che (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS and School of Cyber Security, University of Chinese Academy of Sciences, China), XiaoFeng Wang (Indiana University Bloomington, America), Dan Meng (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS and School of Cyber Security, University of Chinese Academy of Sciences, China), and Rui Hou (Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS and School of Cyber Security, University of Chinese Academy of Sciences, China)
- Enhancing Collective Communication in MCM Accelerators for Deep Learning Training 832
Sabuj Laskar (Texas A&M University), Pranati Majhi (Texas A&M University), Sungkeun Kim (Texas A&M University), Farabi Mahmud (Texas A&M University), Abdullah Muzahid (Texas A&M University), and Eun Jung Kim (Texas A&M University)

IoT/Edge, Session 7/B

- TinyTS: Memory-Efficient TinyML Model Compiler Framework on Microcontrollers 848
Yu-Yuan Liu (National Yang Ming Chiao Tung University, Taiwan), Hong-Sheng Zheng (National Yang Ming Chiao Tung University, Taiwan), Yu-Fang Hu (National Yang Ming Chiao Tung University, Taiwan), Chen-Fong Hsu (National Yang Ming Chiao Tung University, Taiwan), and Tsung Tai Yeh (National Yang Ming Chiao Tung University, Taiwan)
- CAMEL: Co-Designing AI Models and eDRAMs for Efficient On-Device Learning 861
Sai Qian Zhang (New York University), Thierry Tambe (Harvard University), Nestor Cuevas (Harvard University), Gu-Yeon Wei (Harvard University), and David Brooks (Harvard University)

| | |
|--|-----|
| FlipBit: Approximate Flash Memory for IoT Devices | 876 |
| <i>Alexander Buck (University of Toronto, Canada), Karthik Ganesan (University of Toronto, Canada), and Natalie Enright Jerger (University of Toronto, Canada)</i> | |
| Usás: A Sustainable Continuous-Learning Framework for Edge Servers | 891 |
| <i>Cyan Subhra Mishra (The Pennsylvania State University, USA), Jack Sampson (The Pennsylvania State University, USA), Mahmut Taylan Kandemir (The Pennsylvania State University, USA), Vijaykrishnan Narayanan (The Pennsylvania State University, USA), and Chita Das (The Pennsylvania State University, USA)</i> | |

Datacenter & Networks, Session 7/C

| | |
|---|-----|
| Cepheus: Accelerating Datacenter Applications with High-Performance RoCE-Capable Multicast.. | 908 |
| <i>Wenxue Li (Hong Kong University of Science and Technology), Junyi Zhang (Huawei and University of Science and Technology of China), Yufei Liu (Huawei), Gaoxiong Zeng (Huawei), Zilong Wang (Hong Kong University of Science and Technology), Chaoliang Zeng (Hong Kong University of Science and Technology), Pengpeng Zhou (Huawei), Qiaoling Wang (Huawei), and Kai Chen (Hong Kong University of Science and Technology and University of Science and Technology of China)</i> | |
| LibPreemptible: Enabling Fast, Adaptive, and Hardware-Assisted User-Space Scheduling | 922 |
| <i>Yueying Li (Cornell University, USA), Nikita Lazarev (Massachusetts Institute of Technology, USA), David Koufaty (Intel Labs, USA), Tenny Yin (Cornell University, USA), Andy Anderson (Intel Labs, USA), Zhiru Zhang (Cornell University, USA), G. Edward Suh (Cornell University, USA), Kostis Kaffes (Columbia University, USA), and Christina Delimitrou (Massachusetts Institute of Technology, USA)</i> | |
| MINOS: Distributed Consistency and Persistency Protocol Implementation & Offloading to SmartNICs | 937 |
| <i>Antonis Psistakis (University of Illinois Urbana-Champaign, USA), Fabien Chaix (FORTH, Greece), and Josep Torrellas (University of Illinois Urbana-Champaign, USA)</i> | |
| Ursa: Lightweight Resource Management for Cloud-Native Microservices | 954 |
| <i>Yanqi Zhang (Cornell University), Zhuangzhuang Zhou Zhou (Cornell University), Sameh Elnikety (Microsoft Research), and Christina Delimitrou (MIT)</i> | |

Industrial Track, Session 9

| | |
|--|-----|
| An LPDDR-Based CXL-PNM Platform for TCO-Efficient Inference of Transformer-Based Large Language Models | 970 |
| <i>Sang-Soo Park (Samsung Electronics, South Korea), KyungSoo Kim (Samsung Electronics, South Korea), Jinin So (Samsung Electronics, South Korea), Jin Jung (Samsung Electronics, South Korea), Jonggeon Lee (Samsung Electronics, South Korea), Kyoungwan Woo (Samsung Electronics, South Korea), Nayeon Kim (Samsung Electronics, South Korea), Younghyun Lee (Samsung Electronics, South Korea), Hyungyo Kim (Samsung Electronics, South Korea), Yongsuk Kwon (Samsung Electronics, South Korea), Jinhyun Kim (Samsung Electronics, South Korea), Jieun Lee (Samsung Electronics, South Korea), YeonGon Cho (Samsung Electronics, South Korea), Yongmin Tai (Samsung Electronics, South Korea), Jeonghyeon Cho (Samsung Electronics, South Korea), Hoyoung Song (Samsung Electronics, South Korea), Jung Ho Ahn (Seoul National University, South Korea), and Nam Sung Kim (University of Illinois Urbana-Champaign, USA)</i> | |
| LightPool: A NVMe-oF-Based High-Performance and Lightweight Storage Pool Architecture for Cloud-Native Distributed Database | 983 |
| <i>Jiexiong Xu (Zhejiang University & Alibaba Group), Yiquan Chen (Alibaba Group), Yijing Wang (Alibaba Group), Wenhui Shi (Ant Group), Guoju Fang (Alibaba Group), Yi Chen (Zhejiang University), Huasheng Liao (Ant Group), Yang Wang (Ant Group), Hai Lin (Ant Group), Zhen Jin (Zhejiang University), Qiang Liu (Alibaba Group), and Wenzhi Chen (Zhejiang University)</i> | |
| Enterprise-Class Cache Compression Design | 996 |
| <i>Alper Buyuktosunoglu (IBM Research), David Trilla (IBM Research), Bulent Abali (IBM Research), Deanna Berger (IBM Infrastructure), Craig Walters (IBM Infrastructure), and Jang-Soo Lee (IBM Infrastructure)</i> | |

Accelerator, Session 10/A

| | |
|---|------|
| HotTiles: Accelerating SpMM with Heterogeneous Accelerator Architectures | 1012 |
| <i>Gerasimos Gerogiannis (Intel Corporation; University of Illinois at Urbana-Champaign), Sriram Ananthakrishnan (Intel Corporation), Josep Torrellas (University of Illinois at Urbana-Champaign), and Ibrahim Hur (Intel Corporation)</i> | |
| SPARK: Scalable and Precision-Aware Acceleration of Neural Networks via Efficient Encoding... | 1029 |
| <i>Fangxin Liu (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Ning Yang (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Haomin Li (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Zongwu Wang (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Zhuoran Song (Shanghai Jiao Tong University), Songwen Pei (University of Shanghai for Science and Technology), and Li Jiang (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute)</i> | |

| | |
|--|------|
| Data Motion Acceleration: Chaining Cross-Domain Multi Accelerators | 1043 |
| <i>Shu-Ting Wang (University of California San Diego), Hanyang Xu (University of California San Diego), Amin Mamandipoor (University of Kansas), Rohan Mahapatra (University of California San Diego), Byung Hoon Ahn (University of California San Diego), Soroush Ghodrati (University of California San Diego), Krishnan Kailas (IBM Research), Mohammad Alian (University of Kansas), and Hadi Esmaeilzadeh (University of California San Diego)</i> | |
| RELIEF: Relieving Memory Pressure In SoCs Via Data Movement-Aware Accelerator Scheduling | 1063 |
| <i>Sudhanshu Gupta (University of Rochester, USA) and Sandhya Dwarkadas (University of Virginia, USA)</i> | |

GPU, Session 10/B

| | |
|---|------|
| GRIT: Enhancing Multi-GPU Performance with Fine-Grained Dynamic Page Placement | 1080 |
| <i>Yueqi Wang (University of Pittsburgh), Bingyao Li (University of Pittsburgh), Aamer Jaleel (NVIDIA), Jun Yang (University of Pittsburgh), and Xulong Tang (University of Pittsburgh)</i> | |
| WASP: Exploiting GPU Pipeline Parallelism with Hardware-Accelerated Automatic Warp Specialization | 1095 |
| <i>Neal C. Crago (NVIDIA Corporation), Sana Damani (NVIDIA Corporation), Karthikeyan Sankaralingam (NVIDIA Corporation), and Stephen W. Keckler (NVIDIA Corporation)</i> | |
| Guser: A GPGPU Power Stressmark Generator | 1111 |
| <i>Yalong Shan (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Yongkui Yang (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Xuehai Qian (Purdue University, USA), and Zhibin Yu (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China)</i> | |
| GPU Scale-Model Simulation | 1125 |
| <i>Hossein SeyyedAghaei (Ghent University, Belgium), Mahmood Naderan-Tahan (Ghent University, Belgium), and Lieven Eeckhout (Ghent University, Belgium)</i> | |

Cache & Memory System, Session 10/C

| | |
|---|------|
| Agile-DRAM: Agile Trade-Offs in Memory Capacity, Latency, and Energy for Data Centers | 1141 |
| <i>Jaeyoon Lee (Sungkyunkwan University, Republic of Korea), Wonyeong Jung (Sungkyunkwan University, Republic of Korea), Dongwee Kim (Sungkyunkwan University, Republic of Korea), Daero Kim (Samsung Electronics), Junseung Lee (Sungkyunkwan University, Republic of Korea), and Jungrae Kim (Sungkyunkwan University, Republic of Korea)</i> | |

| | |
|--|------|
| CHROME: Concurrency-Aware Holistic Cache Management Framework with Online Reinforcement Learning | 1154 |
| <i>Xiaoyang Lu (Illinois Institute of Technology, USA), Hamed Najafi (Florida International University, USA), Jason Liu (Florida International University, USA), and Xian-He Sun (Illinois Institute of Technology, USA)</i> | |
| Prosper: Program Stack Persistence in Hybrid Memory Systems | 1168 |
| <i>Arun KP (Indian Institute of Technology, India), Debadatta Mishra (Indian Institute of Technology, India), and Biswabandan Panda (Indian Institute of Technology, India)</i> | |
| Mitigating Write Disturbance in Non-Volatile Memory via Coupling Machine Learning with Out-of-Place Updates | 1184 |
| <i>Ronglong Wu (Xiamen University), Zhirong Shen (Xiamen University), Zhiwei Yang (Xiamen University), and Jiwu Shu (Xiamen University)</i> | |

Author Index