2023 56th IEEE/ACM International Symposium on Microarchitecture (MICRO 2023)

Toronto, Ontario, Canada
28 October - 1 November 2023

Pages 1-741

IEEE Catalog Number: CFP23071-POD
ISBN: 979-8-3503-3056-4

Copyright © 2023, Association for Computing Machinery (ACM)
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: CFP23071-POD
ISBN (Print-On-Demand): 979-8-3503-3056-4
ISBN (Online): 979-8-4007-0329-4

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

Table of contents

MICRO Steering Committee xix

MICRO-56 Organizing Committee xx

MICRO-56 Program Committee xxiii

MICRO-56 Light Load Program Committee Members xxxi

MICRO-56 Sponsors.. xxxviii

Message from the MICRO 2023 Program Chairs xxxix

Message from the MICRO 2023 General Chair.. xli

MICRO'23 Artifact Evaluation Report xliii

Keynote I xlvi

Amin Vahdat: Vice President of ML, Systems and Cloud AI at Google xlvi

Keynote II xlvii

Debbie Marr: Intel Fellow and Chief Architect...xlvii

Keynote III xlviii

Mark Horowitz: Yahoo! Founders Professor in the School of Engineering and
Professor of Computer Science, Stanford xlviii

Best Paper Session

Session Chair: Davide Basilio Bartolini (Huawei)

- **Clockhands: Rename-free Instruction Set Architecture for Out-of-order Processors.......1**
  Toru Koizumi (Nagoya Institute of Technology); Ryota Shioya, Shu Sugita, Taichi Amano, Yuya Degawa, Junichiro Kadomoto, Hidetsugu Irie, Shuichi Sakai (The University of Tokyo)
- **Decoupled Vector Runahead 17**
  Ajeya Naithani, Jaime Roelandts (Ghent University); Sam Ainsworth (University of Edinburgh); Timothy M. Jones (University of Cambridge); Lieven Eeckhout (Ghent University)

- **CryptoMMU: Enabling Scalable and Secure Access Control of Third-Party Accelerators 32**
  Faiz Alam, Hyokeun Lee (North Carolina State University); Abhishek Bhattacharjee (Yale University); Amro Awad (North Carolina State University)
- **Phantom: Exploiting Decoder-Detectable Mispredictions 49**
  Johannes Wikner, Daniël Trujillo, Kaveh Razavi (ETH Zürich)

Session 1A: Accelerators Based on HW/SW Co-Design Accelerators for Matrix Processing

Session Chair: Michael Pellauer (NVIDIA)

- **AuRORA: Virtualized Accelerator Orchestration for Multi-Tenant Workloads 62**
  Seah Kim, Jerry Zhao, Krste Asanović, Borivoje Nikolić, Yakun Sophia Shao (University of California)
- **UNICO: Unified Hardware Software Co-Optimization for Robust Neural Network Acceleration...77**
  Bahador Rashidi, Chao Gao, Shan Lu (Huawei Canda Research Center); Zhisheng Wang (Huawei Hisilicon); Chunhua Zhou (Huawei Canada Research Center); Di Niu (University of Alberta); Fengyu Sun (Huawei Hisilicon)
- **Spatula: A Hardware Accelerator for Sparse Matrix Factorization 91**
  Axel Feldmann, Daniel Sanchez (MIT CSAIL)

Session 1B: Architectural Support/Programming Languages, Case Study

Session Chair: Saugata Ghose (University of Illinois Urbana-Champaign)

- **Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices 105**
  Yan Sun (University of Illinois); Yifan Yuan (Intel Labs); Zeduo Yu, Reese Kuper, Chihun Song, Jinghan Huang, Houxiang Ji. Siddharth Agarwal, Jiaqi Lou, Ipoom Jeong (University of Illinois); Ren Wang (Intel Labs); Jung Ho Ahn (Seoul National University); Tianyin Xu, Nam Sung Kim (University of Illinois)
- **Memento: Architectural Support for Ephemeral Memory Management in Serverless Environments 122**
  Ziqi Wang, Kaiyang Zhao, Pei Li, Andrew Jacob (Carnegie Mellon University); Michael Kozuch (Intel Labs); Todd C. Mowry, Dimitrios Skarlatos (Carnegie Mellon University)

- **Simultaneous and Heterogenous Multithreading 137**
  Kuan-Chieh Hsu, Hung-Wei Tseng (University of California)

Session 1C: Design Automation, Synthesis, Hardware Generation

Session Chair: Mark Jeffrey (University of Toronto)

- **Accelerating RTL Simulation with Hardware-Software Co-Design.............................. 153**
  Fares Elsabbagh, Shabnam Sheikhha, Victor A. Ying, Quan M. Nguyen, Joel S. Emer, Daniel Sanchez (MIT CSAIL)
- **Fast, Robust and Transferable Prediction for Hardware Logic Synthesis 167**
  Ceyu Xu, Pragya Sharma, Tianshu Wang, Lisa Wu Wills (Duke University)
- **Khronos: Fusing Memory Access for Improved Hardware RTL Simulation 180**
  Kexing Zhou, Yun Liang, Yibo Lin, Runsheng Wang, Ru Huang (Peking University)

Session 2A: ML Design Space Exploration Generation

Session Chair: Tushar Krishna (Georgia Institute of Technology)

- **SecureLoop: Design Space Exploration of Secure DNN Accelerators.......................... 194**
  Kyungmi Lee, Mengjia Yan (Massachusetts Institute of Technology); Joel S. Emer (Massachusetts Institute of Technology/NVIDIA); Anantha P. Chandrakasan (Massachusetts Institute of Technology)
- **DOSA: Differentiable Model-Based One-Loop Search for DNN Accelerators 209**
  Charles Hong (University of California); Qijing Huang (NVIDIA); Grace Dinh (University of California); Mahesh Subedar (Intel Labs); Yakun Sophia Shao (University of California)
- **TorchSparse++: Efficient Training and Inference Framework for Sparse Convolution on GPUs 225**
  Haotian Tang (MIT); Shang Yang (MIT, Tsinghua University); Zhijian Liu (MIT); Ke Hong (Tsinghua University); Zhongming Yu (UCSD); Xiuyu Li (UC Berkeley); Guohao Dai (Shanghai Jiao Tong University); Yu Wang (Tsinghua University); Song Han (MIT, NVIDIA)

Session 2B: Microarchitecture

Session Chair: Daniel Sorin (Duke University)

- **Branch Target Buffer Organizations.. 240**
  Arthur Perais (Univ. Grenoble Alpes, CNRS); Rami Sheikh (Arm)

- **Warming Up a Cold Front-End with Ignite** 254
  David Schall (University of Edinburgh); Andreas Sandberg (Arm Ltd.); Boris Grot (University of Edinburgh)
- **ArchExplorer: Microarchitecture Exploration via Bottleneck Analysis** 268
  Chen Bai (The Chinese University of Hong Kong, Alibaba Group); Jiayi Huang (The Hong Kong University of Science and Technology (Guangzhou)); Xuechao Wei (Alibaba Group); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou)); Sicheng Li, Hongzhong Zheng (Alibaba Group); Bei Yu (The Chinese University of Hong Kong); Yuan Xie (The Hong Kong University of Science and Technology (Alibaba Group))

Session 2C: Accelerators for Graphs, Robotics

Session Chair: Sabrina Neuman (Boston University)

- **DF-GAS: A Distributed FPGA-as-a-Service Architecture towards Billion-Scale Graph-based Approximate Nearest Neighbor Search** 283
  Shulin Zeng, Zhenhua Zhu, Jun Liu, Haoyu Zhang (Tsinghua University); Guohao Dai (Shanghai Jiao Tong University); Zixuan Zhou (Tsinghua University); Shuangchen Li (Alibaba Group); Xuefei Ning (Tsinghua University); Yuan Xie (Alibaba Group); Huazhong Yang, Yu Wang (Tsinghua University)
- **Dadu-RBD: Robot Rigid Body Dynamics Accelerator with Multifunctional Pipelines** ... 297
  Yuxin Yang, Xiaoming Chen, Yinhe Han (Chinese Academy of Sciences)
- **MEGA Evolving Graph Accelerator** 310
  Chao Gao, Mahbod Afarin, Shafiur Rahman, Nael Abu-Ghazaleh, Rajiv Gupta (UC Riverside)

Session 3A: ML Sparsity

Session Chair: Biswabandan Panda (Indian Institute of Technology Bombay)

- **Eureka: Efficient Tensor Cores for One-Sided Unstructured Sparsity in DNN Inference** 324
  Ashish Gondimalla (Google); Mithuna Thottethodi, T. N. Vijaykumar (Purdue University)

- **RM-STC: Row-Merge Dataflow Inspired GPU Sparse Tensor Core for Energy-Efficient Sparse Acceleration 338**
  Guyue Huang, Zhengyang Wang (University of California); Po-An Tsai (NVIDIA); Chen Zhang (Shanghai Jiao Tong University); Yufei Ding (University of California); Yuan Xie (Hong Kong University of Science and Technology)
- **Sparse-DySta: Sparsity-Aware Dynamic and Static Scheduling for Sparse Multi-DNN Workloads 353**
  Hongxiang Fan (Samsung AI Cambridge and University of Cambridge); Stylianos I. Venieris, Alexandros Kouris (Samsung AI Center); Nicholas D. Lane (University of Cambridge & Flower Labs)

Session 3B: GPUs

Session Chair: Nandita Vijaykumar (University of Toronto); Sabrina Neuman (Boston University)

- **MAD MAcce: Supporting Multiply-Add Operations for Democratizing Matrix-Multiplication Accelerator 367**
  Seunghwan Sung, Sujin Hur, Sungwoo Kim, Dongho Ha (Yonsei University); Yunho Oh (Korea University); Won Woo Ro (Yonsei University)
- **Path Forward Beyond Simulators: Fast and Accurate GPU Execution Time Prediction for DNN Workloads 380**
  Ying Li, Yifan Sun (William & Mary); Adwait Jog (University of Virginia)
- **G10: Enabling an Efficient Unified GPU Memory and Storage Architecture with Smart Tensor Migrations 395**
  Haoyang Zhang, Yirui Eric Zhou, Yuqi Xue, Yiqi Liu, Jian Huang (UIUC)

Session 4A: ML Architecture

Session Chair: Po-An Tsai (NVIDIA)

- **MAICC : A Lightweight Many-Core Architecture with In-Cache Computing for Multi-DNN Parallel Inference 411**
  Renhao Fan, Yikai Cui, Qilin Chen (Tsinghua University); Mingyu Wang (Sun Yat-Sen University); Youhui Zhang, Weimin Zheng, Zhaolin Li (Tsinghua University)

- **Cambricon-U: A Systolic Random Increment Memory Architecture for Unary Computing 424**
  Hongrui Guo; Yongwei Zhao (ICT, CAS); Zhangmai Li (ICT, CAS, HUST); Yifan Hao; Chang Liu; Xinkai Song, Xiaqing Li (ICT, CAS); Zidong Du (ICT, CAS, SHIC); Rui Zhang, Qi Guo (ICT, CAS); Tianshi Chen (Cambricon Technologies); Zhiwei Xu (ICT, CAS)
- **Improving Data Reuse in NPU On-Chip Memory with Interleaved Gradient Order for DNN Training........ 438**
  Jungwoo Kim, Seonjin Na, Sanghyeon Lee, Sunho Lee, Jaehyuk Huh (KAIST)
- **TT-GNN: Efficient On-Chip Graph Neural Network Training via Embedding Reformation and Hardware Optimization 452**
  Zheng Qu (UC Santa Barbara); Dimin Niu, Shuangchen Li, Hongzhong Zheng, Yuan Xie (Alibaba Group)
- **Supporting Energy-Based Learning with an Ising Machine Substrate: A Case Study on RBM 465**
  Uday Kumar Reddy Vengalam, Yongchao Liu, Tong Geng, Hui Wu, Michael Huang (University of Rochester)

Session 4B: Quantum

Session Chair: Hiroaki Kobayashi (Tohoku University)

- **QuComm: Optimizing Collective Communication for Distributed Quantum Computing 479**
  Anbang Wu, Yufei Ding (University of California); Ang Li (Pacific Northwest National Laboratory)
- **QuCT: A Framework for Analyzing Quantum Circuit by Extracting Contextual and Topological Features........ 494**
  Siwei Tan, Congliang Lang, Liang Xiang, Shudi Wang, Xinghui Jia, Ziqi Tan, Tingting Li (Zhejiang University); Jieming Yin (Nanjing University of Posts and Telecommunications); Yongheng Shang, Andre Python, Liqiang Lu, Jianwei Yin (Zhejiang University)
- **ERASER: Towards Adaptive Leakage Suppression for Fault-Tolerant Quantum Computing........ 509**
  Suhas Vittal (Georgia Institute of Technology); Poulami Das (University of Texas); Moinuddin Qureshi (Georgia Institute of Technology)
- **Systems Architecture for Quantum Random Access Memory........ 526**
  Shifan Xu (Yale University); Connor T. Hann (AWS Center for Quantum Computing and California Institute of Technology); Ben Foxman, Steven M. Girvin, Yongshan Ding (Yale University)

- **HetArch: Heterogeneous Microarchitectures for Superconducting Quantum Systems 539**
  Samuel Stein (Pacific Northwest National Laboratory); Sara Sussman, Teague Tomesh, Charles Guinn, Esin Tureci (Princeton University); Sophia Fuhui Lin (University of Chicago); Wei Tang (Princeton University); James Ang (Pacific Northwest National Laboratory); Srivatsan Chakram (Rutgers University); Ang Li (Pacific Northwest National Laboratory); Margaret Martonosi (Princeton University); Fred T. Chong (University of Chicago); Andrew A. Houck (Princeton University); Isaac L. Chuang (Massachusetts Institute of Technology); Michael Austin DeMarco (Massachusetts Institute of Technology and Brookhaven National Laboratory)

Session 4C: Emerging Technologies: Superconducting, Photonics, DNA

Session Chair: Koji Inoue (Kyushu University)

- **Efficiently Enabling Block Semantics and Data Updates in DNA Storage 555**
  Puru Sharma, Cheng-Kai Lim, Dehui Lin, Yash Pote, Djordje Jevdjic (National University of Singapore)
- **ReFOCUS: Reusing Light for Efficient Fourier Optics-Based Photonic Neural Network Accelerator 569**
  Shurui Li, Hangbo Yang, Chee Wei Wong (Univerisity of California); Volker J. Sorger (University of Florida); Puneet Gupta (Univerisity of California)
- **SupeRBNN: Randomized Binary Neural Network Using Adiabatic Superconductor Josephson Devices 584**
  Zhengang Li (Northeastern University); Geng Yuan (University of Georgia); Tomoharu Yamauchi (Tokyo City University); Zabihi Masoud, Yanyue Xie, Peiyan Dong (Northeastern University); Xulong Tang (University of Pittsburgh); Nobuyuki Yoshikawa (Yokohama National University); Devesh Tiwari, Yanzhi Wang (Northeastern University); Olivia Chen (Tokyo City University)
- **SuperBP: Design Space Exploration of Perceptron-Based Branch Predictors for Superconducting CPUs 599**
  Haipeng Zha (University of Southern California); Swamit Tannu (University of Wisconsin-Madison); Murali Annavaram (University of Southern California)

- **SUSHI: Ultra-High-Speed and Ultra-Low-Power Neuromorphic Chip Using Superconducting Single-Flux-Quantum Circuits 614**
  Zeshi Liu, Shuo Chen, Peiyao Qu (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Science); Huanli Liu, Minghui Niu, Liliang Ying, Jie Ren (Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Science); Guangming Tang, Haihang You (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Science)

Session 5A: Security Encryption/Confidentiality Support

Session Chair: Gururaj Saileshwar (University of Toronto / NVIDIA Research)

- **AQ2PNN: Enabling Two-Party Privacy-Preserving Deep Neural Network Inference with Adaptive Quantization 628**
  Yukui Luo (Northeastern University); Nuo Xu (Lehigh University); Hongwu Peng (University of Connecticut); Chenghong Wang (Indiana University Bloomington); Shijin Duan (Northeastern University); Kaleel Mahmood (University of Connecticut); Wujie Wen (North Carolina State University); Caiwen Ding (University of Connecticut); Xiaolin Xu (Northeastern University)
- **CHERIoT: Complete Memory Safety for Embedded Devices 641**
  Saar Amar, David Chisnall, Tony Chen, Nathaniel Wesley Filardo (Microsoft); Ben Laurie (Google); Kunyan Liu, Robert Norton (Microsoft); Simon W. Moore (University of Cambridge); Yucong Tao (Microsoft); Robert N. M. Watson (University of Cambridge); Hongyan Xia (Arm Ltd.)
- **Accelerating Extra Dimensional Page Walks for Confidential Computing....................654**
  Dong Du, Bicheng Yang, Yubin Xia, Haibo Chen (Shanghai Jiao Tong University)
- **GME: GPU-Based Microarchitectural Extensions to Accelerate Homomorphic Encryption..670**
  Kaustubh Shivdikar, Yuhui Bao (Northeastern University); Rashmi Agrawal (Boston University); Michael Shen (Northeastern University); Gilbert Jonatan (KAIST); Evelio Mora (UCAM); Alexander Ingare, Neal Livesay (Northeastern University); José L. Abellán (Universidad de Murcia); John Kim (KAIST); Ajay Joshi (Boston University); David Kaeli (Northeastern University)
- **MAD: Memory-Aware Design Techniques for Accelerating Fully Homomorphic Encryption..685**
  Rashmi Agrawal (Boston University); Leo de Castro (MIT); Chiraag Juvekar (Analog Devices); Anantha ChandraKasan (MIT); Vinod Vaikuntanathan (MIT); Ajay Joshi (Boston University)

Session 5B: Prefetching

Session Chair: Leeor Peled (Toga Networks)

- **Micro-Armed Bandit: Lightweight & Reusable Reinforcement Learning for Microarchitecture Decision-Making 698**
  Gerasimos Gerogiannis, Josep Torrellas (University of Illinois at Urbana-Champaign)
- **CLIP: Load Criticality Based Data Prefetching for Bandwidth-Constrained Many-Core Systems 714**
  Biswabandan Panda (Indian Institute of Technology Bombay)
- **Snake: A Variable-Length Chain-Based Prefetching for GPUs 728**
  Saba Mostofi (Sharif University of Technology); Hajar Falahati (Institute for Research in Fundamental Sciences (IPM)); Negin Mahani (Shahid Bahonar Universuty); Pejman Lotfi-Kamran (Institute for Research in Fundamental Sciences (IPM)); Hamid Sarbazi-Azad (Sharif University of Technology)
- **Treelet Prefetching for Ray Tracing 742**
  Yuan Hsi Chou (University of British Columbia); Tyler Nowicki (Huawei); Tor M. Aamodt (University of British Columbia)

Session 5C: Processing-In-Memory

Session Chair: Dimitrios Skarlatos (Carnegie Mellon University)

- **NAS-SE: Designing a Highly-Efficient In-Situ Neural Architecture Search Engine for Large-Scale Deployment 756**
  Qiyu Wan (NVIDIA and University of Houston); Lening Wang (University of Houston); Jing Wang (Renmin University of China); Shuaiwen Leon Song (Microsoft and University of Sydney); Xin Fu (University of Houston)
- **XFM: Accelerated Software-Defined Far Memory 769**
  Neel Patel, Amin Mamandipoor, Derrick Quinn, Mohammad Alian (University of Kansas)
- **Affinity Alloc: Taming ~~Not So~~ Near-Data Computing 784**
  Zhengrong Wang; Christopher Liu (UCLA); Nathan Beckmann (CMU); Tony Nowatzki (UCLA)
- **MVC: Enabling Fully Coherent Multi-Data-Views through the Memory Hierarchy with Processing in Memory 800**
  Daichi Fujiki (Keio University)
- **AESPA: Asynchronous Execution Scheme to Exploit Bank-Level Parallelism of Processing-in-Memory 815**
  Hongju Kal, Chanyoung Yoo, Won Woo Ro (Yonsei University)

Session 6A: Security Hardware

Session Chair: Samira Mirbagher Ajorpaz (North Carolina State University)

- **ReCon: Efficient Detection, Management, and Use of Non-Speculative Information Leakage 828**
  Pavlos Aimoniotis (Uppsala University); Amund Bergland Kvalsvik (Norwegian University of Science and Technology); Xiaoyue Chen (Uppsala University); Magnus Själander (Norwegian University of Science and Technology); Stefanos Kaxiras (Uppsala University)
- **Uncore Encore: Covert Channels Exploiting Uncore Frequency Scaling 843**
  Yanan Guo (University of Pittsburgh); Dingyuan Cao (University of Illinois Urbana Champaign); Xin Xin, Youtao Zhang, Jun Yang (University of Pittsburgh)
- **Hardware Support for Constant-Time Programming 856**
  Yuanqing Miao, Mahmut Taylan Kandemir, Danfeng Zhang, Yingtian Zhang, Gang Tan, Dinghao Wu (The Pennsylvania State University)
- **AutoCC: Automatic Discovery of Covert Channels in Time-Shared Hardware 871**
  Marcelo Orenes-Vera, Hyunsung Yun (Princeton University); Nils Wistoff (ETH Zürich); Gernot Heiser (UNSW Sydney); Luca Benini (ETH Zürich); David Wentzlaff, Margaret Martonosi (Princeton University)

Session 6B: Datacenter Networks

Session Chair: Trevor E. Carlson (National University of Singapore)

- **NeuroLPM - Scaling Longest Prefix Match Hardware with Neural Networks 886**
  Alon Rashelbach, Igor de-Paula, Mark Silberstein (Technion)
- **Space Microdatacenters 900**
  Nathaniel Bleier, Muhammad Husnain Mubarik, Gary R. Swenson, Rakesh Kumar (University of Illinois Urbana-Champaign)
- **LogNIC: A High-Level Performance Model for SmartNICs 916**
  Zerui Guo (University of Wisconsin-Madison); Jiaxin Lin (The University of Texas at Austin); Yuebin Bai (Beihang University); Daehyeok Kim (The University of Texas at Austin and Microsoft); Michael Swift (University of Wisconsin-Madison); Aditya Akella (The University of Texas at Austin); Ming Liu (University of Wisconsin-Madison)
- **Heterogeneous Die-to-Die Interfaces: Enabling More Flexible Chiplet Interconnection Systems 930**
  Yinxiao Feng, Dong Xiang, Kaisheng Ma (Tsinghua University)

Session 6C: Reliability, Availability

Session Chair: Freddy Gabbay (Ruppin Academic College)

- **Predicting Future-System Reliability with a Component-Level DRAM Fault Model 944**
  Jeageun Jung, Mattan Erez (University of Texas at Austin)
- **Impact of Voltage Scaling on Soft Errors Susceptibility of Multicore Server CPUs 957**
  Dimitris Agiakatsikas (University of Piraeus); George Papadimitriou, Vasileios Karakostas, Dimitris Gizopoulos (University of Athens); Mihalis Psarakis (University of Piraeus); Camille Bélanger-Champagne, Ewart Blackmore (TRIUMF)
- **Si-Kintsugi: Towards Recovering Golden-Like Performance of Defective Many-Core Spatial Architectures for AI 972**
  Edward Hanson, Shiyu Li, Guanglei Zhou, Feng Cheng, Yitu Wang, Rohan Bose, Hai (Helen) Li, Yiran Chen (Duke University)
- **How to Kill the Second Bird with One ECC: The Pursuit of Row Hammer Resilient DRAM 986**
  Michael Jaemin Kim, Minbok Wi, Jaehyun Park, Seoyoung Ko, Jaeyoung Choi, Hwayoung Nam (Seoul National University); Nam Sung Kim (University of Illinois Urbana Champaign); Jung Ho Ahn (Seoul National University); Eojin Lee (Inha University)

Session 7A: Accelerators Various

Session Chair: Alex K. Jones (University of Pittsburgh)

- **Bucket Getter: A Bucket-Based Processing Engine for Low-bit Block Floating Point (BFP) DNNs........ 1002**
  Yun-Chen Lo, Ren-Shuo Liu (National Tsing Hua University)
- **ACRE: Accelerating Random Forests for Explainability 1016**
  Andrew McCrabb, Aymen Ahmed, Valeria Bertacco (University of Michigan)
- **δLTA: Decoupling Camera Sampling from Processing to Avoid Redundant Computations in the Vision Pipeline 1029**
  Raúl Taranco, José-María Arnau, Antonio González (Universitat Politècnica de Catalunya)

Session 7B: Caches, Intermitent Computing, Persistency

Session Chair: Rachata Ausavarungnirun (King Mongkut's University of Technology North Bangkok)

- **McCore: A Holistic Management of High-Performance Heterogeneous Multicores** **1044**
  Jaewon Kwon, Yongju Lee, Hongju Kal, Minjae Kim, Youngsok Kim, Won Woo Ro (Yonsei University)
- **SweepCache: Intermittence-Aware Cache on the Cheap** **1059**
  Yuchen Zhou, Jianping Zeng (Purdue University); Jungi Jeong (Google); Jongouk Choi (University of Central Florida); Changhee Jung (Purdue University)
- **Persistent Processor Architecture** **1075**
  Jianping Zeng, Jungi Jeong, Changhee Jung (Purdue University)

Session 8A: Accelerators for Neural Nets <br> Accelerators for Matrix Processing

Session Chair: Jason Clemons (NVIDIA)

- **ADA-GP: Accelerating DNN Training by Adaptive Gradient Prediction** **1092**
  Vahid Janfaza, Shantanu Mandal, Farabi Mahmud, Abdullah Muzahid (Texas A&M University)
- **HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity** **1106**
  Yannan Nellie Wu (MIT); Po-An Tsai, Saurav Muralidharan, Angshuman Parashar (NVIDIA); Vivienne Sze (MIT); Joel S. Emer (MIT/NVIDIA)
- **Exploiting Inherent Properties of Complex Numbers for Accelerating Complex Valued Neural Networks** **1121**
  Hyunwuk Lee, Hyungjun Jang, Sungbin Kim, Sungwoo Kim, Wonho Cho, Won Woo Ro (Yonsei University)
- **Point Cloud Acceleration by Exploiting Geometric Similarity** **1135**
  Cen Chen (South China University of Technology & Pazhou Lab); Xiaofeng Zou, Hongen Shao (South China University of Technology); Yangfan Li (Central South University); Kenli Li (Hunan University)
- **HARP: Hardware-Based Pseudo-Tiling for Sparse Matrix Multiplication Accelerator** **1148**
  Jinkwon Kim, Myeongjae Jang, Haejin Nam, Soontae Kim (KAIST)

Session 8B: Virtual Memory (Translation)

Session Chair: Mohammad Alian (University of Kansas)

- **IDYLL: Enhancing Page Translation in Multi-GPUs via Light Weight PTE Invalidations 1163**
  Bingyao Li, Yanan Guo, Yueqi Wang (University of Pittsburgh); Aamer Jaleel (NVIDIA); Jun Yang, Xulong Tang (University of Pittsburgh)
- **Victima: Drastically Increasing Address Translation Reach by Leveraging Underutilized Cache Resources 1178**
  Konstantinos Kanellopoulos, Hong Chul Nam, F. Nisa Bostanci, Rahul Bera, Mohammad Sadrosadati (ETH Zürich); Rakesh Kumar (Norwegian University of Science and Technology); Davide Basilio Bartolini (Huawei Zurich Research Center); Onur Mutlu (ETH Zürich)
- **Utopia: Fast and Efficient Address Translation via Hybrid Restrictive & Flexible Virtual-to-Physical Address Mapping 1196**
  Konstantinos Kanellopoulos, Rahul Bera, Kosta Stojiljkovic, Nisa Bostanci, Can Firtina (ETH Zürich); Rachata Ausavarungnirun (King Mongkut's University of Technology North Bangkok); Rakesh Kumar (Norwegian University of Science and Technology); Nastaran Hajinazar (Intel Labs); Mohammad Sadrosadati (ETH Zürich); Nandita Vijaykumar (University of Toronto); Onur Mutlu (ETH Zürich)
- **Architectural Support for Optimizing Huge Page Selection Within the OS 1213**
  Aninda Manocha (Princeton University); Zi Yan (NVIDIA); Esin Tureci (Princeton University); Juan L. Aragón (University of Murcia); David Nellans (NVIDIA); Margaret Martonosi (Princeton University)

Session 8C: Benchmarking and Methodology

Session Chair: Miquel Moretó (Universitat Politècnica de Catalunya/Barcelona Supercomputing Center)

- **Photon: A Fine-Grained Sampled Simulation Methodology for GPU Workloads 1227**
  Changxi Liu (National University of Singapore); Yifan Sun (College of William & Mary); Trevor E. Carlson (National University of Singapore)
- **Rigorous Evaluation of Computer Processors with Statistical Model Checking 1242**
  Filip Mazurek, Arya Tschand (Duke University); Yu Wang (University of Florida); Miroslav Pajic, Daniel J. Sorin (Duke University)

- **TeAAL: A Declarative Framework for Modeling Sparse Tensor Accelerators............. 1255**
  Nandeeka Nayak (University of Illinois Urbana-Champaign); Toluwanimi O. Odemuyiwa (University of California); Shubham Ugare, Christopher W. Fletcher (University of Illinois Urbana-Champaign); Michael Pellauer (NVIDIA); Joel S. Emer (MIT/NVIDIA)
- **TileFlow: A Framework for Modeling Fusion Dataflow via Tree-Based Analysis........ 1271**
  Size Zheng, Siyuan Chen, Siyuan Gao, Liancheng Jia (Peking University); Guangyu Sun, Runsheng Wang, Yun Liang (Peking University & Advanced Innovation Center for Integrated Circuits)
- **Learning to Drive Software-Defined Solid-State Drives 1289**
  Daixuan Li, Jinghan Sun, Jian Huang (University of Illinois Urbana-Champaign)

Session 9A: Accelerators in Processors

Session Chair: Sihang Liu (University of Waterloo)

- **Cambricon-R: A Fully Fused Accelerator for Real-Time Learning of Neural Scene Representation.. 1305**
  Xinkai Song, Yuanbo Wen (ICT, CAS); Xing Hu (ICT, CAS SHIC); Tianbo Liu (ICT, CAS USTC); Yifan Hao (ICT, CAS); Haoxuan Zhou (ICT, CAS UCAS); Husheng Han; Tian Zhi (ICT, CAS); Zidong Du (ICT, CAS SHIC); Wei Li, Rui Zhang (ICT, CAS); Chen Zhang (SJTU); Lin Gao, Qi Guo (ICT, CAS); Tianshi Chen (Cambricon Technologies)
- **Strix: An End-to-End Streaming Architecture with Two-Level Ciphertext Batching for Fully Homomorphic Encryption with Programmable Bootstrapping 1319**
  Adiwena Putra, Prasetiyo, Yi Chen, John Kim, Joo-Young Kim (KAIST)
- **A Tensor Marshaling Unit for Sparse Tensor Algebra on General-Purpose Processors.. 1332**
  Marco Siracusa, Víctor Soria-Pardos, Francesco Sgherzi (Barcelona Supercomputing Center and Universitat Politècnica de Catalunya); Joshua Randall (Arm); Douglas J. Joseph (Samsung); Miquel Moretó Planas, Adrià Armejach (Barcelona Supercomputing Center and Universitat Politècnica de Catalunya)
- **Tailors: Accelerating Sparse Tensor Algebra by Overbooking Buffer Capacity 1347**
  Zi Yu Xue, Yannan Nellie Wu (MIT); Joel S. Emer (MIT/NVIDIA); Vivienne Sze (MIT)

Session 9B: ML Compiler Optimizations/Reconfigurable Architectures

Session Chair: Jian Huang (University of Illinois Urbana-Champaign)

- **Grape: Practical and Efficient Graph-based Executions for Dynamic Deep Neural Networks on GPUs 1364**
  Bojian Zheng (University of Toronto, CentML); Cody Hao Yu, Jie Wang (Amazon); Yaoyao Ding (University of Toronto, CentML); Yizhi Liu, Yida Wang (Amazon); Gennady Pekhimenko (University of Toronto, CentML)
- **PockEngine: Sparse and Efficient Fine-Tuning in a Pocket 1381**
  Ligeng Zhu (MIT); Lanxiang Hu (UCSD); Ji Lin, Wei-Chen Wang, Wei-Ming Chen (MIT); Chuang Gan (MIT-IBM Watson AI Lab); Song Han (MIT, NVIDIA)
- **Towards Efficient Control Flow Handling in Spatial Architecture via Architecting the Control Flow Plane 1395**
  Jinyi Deng, Xinru Tang, Jiahao Zhang, Yuxuan Li, Linyun Zhang (Tsinghua University); Boxiao Han, Hongjun He (China Mobile Research Institute); Fengbin Tu (Hong Kong University of Science and Technology); Leibo Liu, Shaojun Wei, Yang Hu (Tsinghua University); Shouyi Yin (Tsinghua University and Shanghai AI Lab)
- **Pipestitch: An Energy-Minimal Dataflow Architecture with Lightweight Threads 1409**
  Nathan Serafin, Souradip Ghosh, Harsh Desai, Nathan Beckmann, Brandon Lucia (Carnegie Mellon University)

Session 9C: Domain Specific Genomics

Session Chair: Pradip Bose (IBM)

- **CASA: An Energy-Efficient and High-Speed CAM-Based SMEM Seeding Accelerator for Genome Alignment 1423**
  Yi Huang (Tsinghua University); Lingkun Kong (Rice University); Dibei Chen (Tsinghua University); Zhiyu Chen (Rice University); Xiangyu Kong, Jianfeng Zhu (Tsinghua University); Konstantinos Mamouras (Rice University); Shaojun Wei (Tsinghua University); Kaiyuan Yang (Rice University); Leibo Liu (Tsinghua University)

- **Swordfish: A Framework for Evaluating Deep Neural Network-Based Basecalling using Computation-In-Memory with Non-Ideal Memristors 1437**
  Taha Shahroodi (TU Delft); Gagandeep Singh (AMD Research and ETH Zürich); Mahdi Zahedi (TU Delft); Haiyu Mao, Joel Lindegger, Can Firtina (ETH Zürich); Stephan Wong (TU Delft); Onur Mutlu (ETH Zürich); Said Hamdioui (TU Delft)
- **DASH-CAM: Dynamic Approximate SearcH Content Addressable Memory for Genome Classification.. 1453**
  Zuher Jahshan, Itay Merlin (Bar Ilan University); Esteban Garzón (University of Calabria); Leonid Yavits (Bar Ilan University)
- **GMX: Instruction Set Extensions for Fast, Scalable, and Efficient Genome Sequence Alignment 1466**
  Max Doblas, Oscar Lostes-Cazorla (Barcelona Supercomputing Center); Quim Aguado-Puig (Universitat Autònoma de Barcelona); Nicholas Cebry (Cornell University); Pau Fontova-Musté (Barcelona Supercomputing Center); Christopher Batten (Cornell University); Santiago Marco-Sola (Barcelona Supercomputing Center and Universitat Politècnica de Catalunya); Miquel Moretó (Universitat Politècnica de Catalunya)