

# **2025 IEEE 43rd International Conference on Computer Design (ICCD 2025)**

**Richardson, Texas, USA  
10-12 November 2025**

**Pages 1-440**



**IEEE Catalog Number: CFP25ICD-POD  
ISBN: 979-8-3315-0347-5**

**Copyright © 2025 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:	CFP25ICD-POD
ISBN (Print-On-Demand):	979-8-3315-0347-5
ISBN (Online):	979-8-3315-0346-8
ISSN:	1063-6404

**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](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

# 2025 IEEE 43rd International Conference on Computer Design (ICCD)

## ICCD 2025

### Table of Contents

Message from the General Chair .....	xxiv
Organizing Committee .....	xxv
Technical Program Committee .....	xxvi

### Session 1A - Security and Privacy for AI Hardware

HElix: Genome Similarity Detection in the Encrypted Domain .....	1
<i>Rostin Shokri (University of Delaware, USA), Charles Gouert (University of Delaware, USA), and Nektarios Georgios Tsoutsos (University of Delaware, USA)</i>	
Targeted Fault Injection Attack on Semantic Segmentation Models .....	9
<i>Jhon Ordoñez (University of Delaware, USA) and Chengmo Yang (University of Delaware, USA)</i>	
Towards Low-Latency and Adaptive Ransomware Detection Using Contrastive Learning .....	17
<i>Zhixin Pan (Florida State University, USA), Ziyu Shu (Stony Brook University, USA), and Amberbir Alemayoh (Florida State University, USA)</i>	
SecNPU: Securing LLM Inference on NPU .....	25
<i>Xuanyao Peng (Chinese Academy of Sciences, China; Southern University of Science and Technology, China; University of Chinese Academy of Sciences, China), Yinghao Yang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Shangjie Pan (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; Zhongguancun Laboratory, China), Junjie Huang (Southern University of Science and Technology, China), Yujun Liang (Southern University of Science and Technology, China), Hang Lu (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; Zhongguancun Laboratory, China), Fengwei Zhang (Southern University of Science and Technology, China), and Xiaowei Li (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; Zhongguancun Laboratory, China)</i>	

## Session 1B - Logic and Circuit Design

DMP-BFP: Dynamic Mixed-Precision Block Floating-Point and Exponent-Guided Precision Adjustment .....	33
<i>Yu-Chih Tsai (National Tsing Hua University, Taiwan), Chia-Cheng Chang (National Tsing Hua University, Taiwan), and Ren-Shuo Liu (National Tsing Hua University, Taiwan)</i>	
Hardware Efficient Multiplier Design using an Optimal Mix of Approximate Booth Encodings .....	41
<i>Chandan Kumar N S (International Institute of Information Technology-Bangalore), S Bhavana (International Institute of Information Technology-Bangalore), Ajitesh Kumar Singh (International Institute of Information Technology-Bangalore), and Madhav Rao (International Institute of Information Technology-Bangalore)</i>	
PolyPE: An Efficient Multi-Precision Multi-Mode Floating-Point Processing Element for HPC and AI .....	49
<i>Zhenzhen Jia (National University of Defense Technology, China), Hongbing Tan (National University of Defense Technology, China), Ling Yang (National University of Defense Technology, China), Hui Guo (National University of Defense Technology, China), Kun Zeng (National University of Defense Technology, China), JunSheng Chang (National University of Defense Technology, China), Yongwen Wang (National University of Defense Technology, China), and Libo Huang (National University of Defense Technology, China)</i>	
CHQ-SC: Compact and High-Quality Stochastic Computing Framework Using Magnetic Tunnel Junction .....	57
<i>Yu Ma (National University of Defense Technology, China), Jianmin Zhang (National University of Defense Technology, China), Yan Sun (National University of Defense Technology, China), and Siqing Fu (National University of Defense Technology, China)</i>	

## Session 2A - Scalable AI Training

RAM-Wafer: RL-Based Automatic Mapping Framework for Large-Scale AI Training on Wafer-Scale Computing .....	61
<i>Xu Dai (Shanghai Artificial Intelligence Laboratory, China), Dehao Kong (Tsinghua University, China), Xufeng He (Shanghai Artificial Intelligence Laboratory, China), Zijun Xu (Shanghai Artificial Intelligence Laboratory, China), Shaopeng Zhai (Shanghai Artificial Intelligence Laboratory, China), Yang Hu (Tsinghua University, China), and Shouyi Yin (Tsinghua University, China)</i>	

DHeLlam: General-Purpose, Automatic Micro-batch Co-execution for Distributed LLM Training ....	70
<i>Haiquan Wang (University of Science and Technology of China, China),      Chaoyi Ruan (National University of Singapore, Singapore), Jia He      (University of Science and Technology of China, China; Mohamed bin      Zayed University of Artificial Intelligence, United Arab Emirates),      Jiaqi Ruan (University of Science and Technology of China, China;      Mohamed bin Zayed University of Artificial Intelligence, United Arab      Emirates), Chengjie Tang (Shanxi University, China; Institute of      Artificial Intelligence, China), Xiaosong Ma (Mohamed bin Zayed      University of Artificial Intelligence, United Arab Emirates), and      Cheng Li (University of Science and Technology of China, China;      Institute of Artificial Intelligence, China)</i>	
A Co-Design Framework for Graph Processing on CPU-GPU Heterogeneous Platforms .....	79
<i>Yuan Zhang (CAS, China), Huawei Cao (CAS, China; University of Chinese      Academy of Sciences, China), Yiming Sun (CAS, China), Ming Dun (CAS,      China), Jie Zhang (CAS, China), and Xiaochun Ye (CAS, China)</i>	
Towards Affordable, Adaptive and Automatic GNN Training on CPU-GPU Heterogeneous Platforms.	
87	
<i>Tong Qiao (Beihang University, China), Ao Zhou (Beihang University,      China), Yingjie Qi (Beihang University, China), Yiou Wang (Beihang      University, China), Han Wan (Beihang University, China), Jianlei Yang      (Beihang University, China), and Chunming Hu (Beihang University,      China)</i>	
<b>Session 2B - Emerging Trends</b>	
Enhancing Transformer Inference Efficiency on FPGA through Fully Fusion and Integer-Only Quantization Techniques .....	95
<i>Zhenqi Li (National University of Defense Technology, China), Yuan Li      (National University of Defense Technology, China), Mingche Lai      (National University of Defense Technology, China), Puguang Liu      (National University of Defense Technology, China), Qiang Wang      (National University of Defense Technology, China), Yankang Zhao      (National University of Defense Technology, China), Hanyuan Li      (National University of Defense Technology, China), and Xingyun Qi      (National University of Defense Technology, China)</i>	
RACE-IT: A Reconfigurable Analog Computing Engine for In-Memory Transformer Acceleration .	103
<i>Lei Zhao (Hewlett Packard Labs, USA), Aishwarya Natarajan (Hewlett      Packard Labs, USA), Luca Buonanno (Hewlett Packard Labs, USA), Archit      Gajjar (Hewlett Packard Labs, USA), Ron Roth (Technion - Israel      Institute of Technology, Israel), Sergey Serebryakov (Hewlett Packard      Labs, USA), John Moon (Hewlett Packard Labs, USA), Omar Eldash      (Hewlett Packard Labs, USA), Jim Ignowski (Hewlett Packard Labs, USA),      and Giacomo Pedretti (Hewlett Packard Labs, USA)</i>	
PACE-Lite: Compact and Efficient Piecewise Polynomial Approximation for Transformer Nonlinearity Acceleration .....	111
<i>Arpan Suravi Prasad (ETH Zurich, Switzerland), Gamze İslamoğlu (ETH      Zurich, Switzerland), Luca Bertaccini (ETH Zurich, Switzerland),      Davide Rossi (University of Bologna, Italy; Chips-IT, Italy),      Francesco Conti (University of Bologna, Italy), and Luca Benini (ETH      Zurich, Switzerland; University of Bologna, Italy )</i>	

QuFi: Adaptive Tiled Gustavson Output Reuse for Edge Sparse DNN Accelerators .....	119
<i>Adrián Navarro (University of Murcia, Spain), José Cano (University of Glasgow, United Kingdom), José L. Abellán (University of Murcia, Spain), and Manuel E. Acacio (University of Murcia, Spain)</i>	
HBM-Aware Number Theoretic Transform Accelerator for Zero-Knowledge Proof .....	127
<i>Sangwon Shin (Korea University, Republic of Korea), Ngoc-Son Pham (Korea University, Republic of Korea), Lei Xu (Kent State University, USA), Weidong Shi (University of Houston, USA), and Taeweon Suh (Korea University, Republic of Korea)</i>	

## Session 3A - Memory Management and Performance Modeling

Repo: Proactive Swapping Exploiting Loop Patterns in Modern Applications .....	131
<i>Jiahui Zhang (Huazhong University of Science and Technology, China), Qiang Cao (Huazhong University of Science and Technology, China), Yekang Zhan (Huazhong University of Science and Technology, China), Yuchen Hu (Huazhong University of Science and Technology, China), and Jie Yao (Huazhong University of Science and Technology, China)</i>	
RT-PMalloc: Optimizing Persistent Memory Allocation for Soft Real-Time Systems .....	139
<i>Yuquan Chi (Sun Yat-sen University, China), Yinjin Fu (Sun Yat-sen University, China), and Nong Xiao (Sun Yat-sen University, China)</i>	
A Scalable and Overflow-Tolerant Mechanism for Minimum Virtual Time Tracking .....	147
<i>Gyusun Lee (Yonsei University, South Korea), Seungwoo Jin (Sungkyunkwan University, South Korea), Jiwon Woo (Sungkyunkwan University, South Korea), and Jinkyu Jeong (Yonsei University, South Korea)</i>	
CAST: An Efficient Framework for Schedules Performance Prediction Based on Compact ASTs ....	155
<i>Qingqiu Lan (Chongqing University, China), Ao Ren (Chongqing University, China), Zhenyu Wang (Chongqing University, China), Wei Li (Chongqing University, China), Hongbin Zhu (Chongqing University, China), Yujuan Tan (Chongqing University, China), Duo Liu (Chongqing University, China), Kan Zhong (Chongqing University, China), and Chaoxia Qin (Chongqing University, China)</i>	

## Session 3B - Yield Analysis & 2.5D/3D Physical Design

STAMP-2.5D: Structural and Thermal Aware Methodology for Placement in 2.5D Integration .....	159
<i>Varun Darshana Parekh (The Pennsylvania State University, USA), Zachary Wyatt Hazenstab (The Pennsylvania State University, USA), Srivatsa Rangachar Srinivasa (Intel, USA), Krishnendu Chakrabarty (Arizona State University, USA), Kai Ni (University of Notre Dame, USA), and Vijaykrishnan Narayanan (The Pennsylvania State University, USA)</i>	

OpenYield: An Open-Source SRAM Yield Analysis and Optimization Benchmark Suite .....	167
<i>Shan Shen (Nanjing University of Science and Technology, China), Xingyang Li (Beihang University, China), Zhuohua Liu (Beihang University, China), Junhao Ma (Nanjing University of Science and Technology, China), Yikai Wang (Nanjing University of Science and Technology, China), Yiheng Wu (Nanjing University of Science and Technology, China), Yuquan Sun (Beihang University, China), and Wei W. Xing (University of Sheffield, United Kingdom)</i>	
3DPX - An Open-Source Methodology for 3D Physical Design Exploration .....	176
<i>George Goudroumanis (University of Thessaly, Greece), Maria Pantazi-Kypraiou (University of Thessaly, Greece), George Floros (Trinity College Dublin, Ireland), Athanasios Tziouvaras (University of Thessaly, Greece), George Stamoulis (University of Thessaly, Greece), and Alberto García-Ortiz (University of Bremen, Germany)</i>	
Declarative Synthesis and Multi-Objective Optimization of Stripboard Circuit Layouts Using Answer Set Programming .....	180
<i>Fang Li (Oklahoma Christian University, USA)</i>	

## Session 4A - Memory System Innovations

R2Hash: A Read-Optimized and Resize-Friendly Hashing Index for Persistent Memory .....	184
<i>Jinlei Hu (Huazhong University of Science and Technology, China), Bo Chen (Huazhong University of Science and Technology, China), Miao Song Zhang (Huazhong University of Science and Technology, China), Jing Hu (Huazhong University of Science and Technology, China), Jianxi Chen (Huazhong University of Science and Technology, China), and Dan Feng (Huazhong University of Science and Technology, China)</i>	
ALPHA: A Scalable Lock-Free Partitioned Hash Index for Persistent Memory on NUMA Architectures .....	193
<i>Qiyang Zheng (Harbin Institute of Technology, China), Hao Hu (Harbin Institute of Technology, China), Hao Huang (Harbin Institute of Technology, China), Yanqi Pan (Harbin Institute of Technology, China), Yifeng Zhang (Harbin Institute of Technology, China), Wen Xia (Harbin Institute of Technology, China), Xiangrui Meng (Jinan Inspur Data Technology Co., Ltd., China), and Xudong Li (Jinan Inspur Data Technology Co., Ltd., China)</i>	
DDLM: Demand-Aware Dynamic Link Width Management for Energy-Efficient CXL Memory .....	201
<i>Taejeong Kim (Sungkyunkwan University, Republic of Korea; Samsung Electronics Co., Ltd., Republic of Korea), Junbum Park (Sungkyunkwan University, Republic of Korea; Samsung Electronics Co., Ltd., Republic of Korea), Yongho Lee (Sungkyunkwan University, Republic of Korea), and Seokin Hong (Sungkyunkwan University, Republic of Korea)</i>	
Computing-In-Memory Dataflow for Minimal Buffer Traffic .....	209
<i>Choongseok Song (Hanyang University, Republic of Korea) and Doo Seok Jeong (Hanyang University, Republic of Korea)</i>	

PIMFY: Eliminating Remote Page Walks in MCM GPUs .....	217
<i>Junsung Kim (Yonsei University, Republic of Korea), Sungwoo Kim (Yonsei University, Republic of Korea), Seunghyun Jin (Yonsei University, Republic of Korea), and Won Woo Ro (Yonsei University, Republic of Korea)</i>	

## Session 4B - Emerging Memory Technologies

PriME: PIM-Aware Efficient Compression for Memory-Bound Embedding Layers in sLLMs .....	221
<i>Junghyeok Lee (Seoul National University of Science and Technology, Korea), Jhoon Jang (Seoul National University of Science and Technology, Korea), and Hyun Kim (Seoul National University of Science and Technology, Korea)</i>	
CMC: Compound Memory-Computing Architecture for Energy-Efficient CNN Accelerators .....	229
<i>Ming Han (Harbin Institute of Technology), Jin Wu (Harbin Institute of Technology), Jian Dong (Harbin Institute of Technology), Ye Wang (Harbin Institute of Technology), and Gang Qu (University of Maryland)</i>	
Dissecting and Re-Architecting 3D NAND Flash PIM Arrays for Efficient Single-Batch Token Generation in LLMs .....	237
<i>Yongjoo Jang (Korea University, South Korea), Sangwoo Hwang (Korea University, South Korea), Hojin Lee (Korea University, South Korea), Sangwoo Jung (Korea University, South Korea), Donghun Lee (Korea University, South Korea), Wonbo Shim (Seoul National University of Science and Technology, South Korea), and Jaeha Kung (Korea University, South Korea)</i>	
MamCIMFlow: An Integrated Co-Design of RRAM-Based CIM and Selective State-Space Streaming for Efficient Mamba Model Acceleration .....	245
<i>Mingzi Li (The University of Hong Kong, China; Hong Kong Productivity Council, China), Zhongrui Wang (Southern University of Science and Technology, China), Zhongwen Ye (Hong Kong Productivity Council, China), Tao Pan (Hong Kong Productivity Council, China), and Han Wang (The University of Hong Kong, China)</i>	
PIM-SUM: Fast and Reliable In-Memory Summation for Recommendation Systems .....	249
<i>Fan Li (University of Central Florida, USA), Ruizhi Zhu (University of Central Florida, USA), Huize Li (University of Central Florida, USA), Di Wu (University of Central Florida, USA), and Xin Xin (University of Central Florida, USA)</i>	

## Session 5A - Large-Scale Model Inference

Ghidorah: Fast LLM Inference on Edge with Speculative Decoding and Hetero-Core Parallelism....	253
<i>Jinhui Wei (Sun Yat-sen University, China), Ye Huang (Sun Yat-sen University, China), Yuhui Zhou (Sun Yat-sen University, China), Jiazhi Jiang (Beijing Normal University, China), and Jiangsu Du (Sun Yat-sen University, China)</i>	

DualSpar: A Dual-Granularity Memory Framework with Adaptive Sparsity for Efficient LLM Inference .....	261
<i>Yujuan Tan (Chongqing University, China; Institute of Computing Technology, China), Jiayi Guo (Chongqing University, China), Zhuoxin Bai (Chongqing University, China), Sanle Zhao (Chongqing University, China), Yujiao Wang (Chongqing University, China), Zongjie Wang (Chongqing University, China), Ao Ren (Chongqing University, China), Kan Zhong (Chongqing University, China), Lin Huang (Chongqing University, China; Inspur Yunzhou Industrial Internet Co., Ltd., China), and Jun Liu (IEIT SYSTEMS(Beiing)Co.,Ltd., China)</i>	
Throughput-Oriented LLM Inference via KV-Activation Hybrid Caching with a Single GPU .....	269
<i>Sanghyeon Lee (KAIST, Republic of Korea), Hongbeen Kim (KAIST, Republic of Korea), Soojin Hwang (KAIST, Republic of Korea), Guseul Heo (KAIST, Republic of Korea), Minwoo Noh (KAIST, Republic of Korea), and Jaehyuk Huh (KAIST, Republic of Korea)</i>	
AuLoRA: Fine-Grained Loading and Computation Orchestration for Efficient LoRA LLM Serving	277
<i>Xiao Shi (Sun Yat-sen University, China), Jiangsu Du (Sun Yat-sen University, China), Zhiguang Chen (Sun Yat-sen University, China), and Yutong Lu (Sun Yat-sen University, China)</i>	
Taming Sparse Giants: Deploying Mixture-of-Experts on 3D Heterogeneous Compute-in-Memory Systems .....	285
<i>Pragya Sharma (Arizona State University, USA), Ashish Reddy Bommana (Arizona State University, USA), Farshad Firouzi (Arizona State University, USA), and Krishnendu Chakrabarty (Arizona State University, USA)</i>	

## Session 5B - Design, Testing, and Verification

μSTT: Microarchitecture Design for Speculative Taint Tracking .....	289
<i>Boru Chen (University of California, USA), Rutvik Choudhary (University of Illinois Urbana-Champaign, USA), Kaustubh Khulbe (University of Illinois Urbana-Champaign, USA), Archie Lee (University of California, USA), Adam Morrison (Tel Aviv University, Israel), and Christopher W. Fletcher (University of California, USA)</i>	
Hot-FV: A Semi-Formal Test Generation Framework for RTL Functional Coverage using Warm Starting States .....	298
<i>Ziyue Zheng (The Hong Kong University of Science and Technology, China), Zhiyuan Yan (The Hong Kong University of Science and Technology, China), Xiangchen Meng (The Hong Kong University of Science and Technology, China), Guangyu Hu (The Hong Kong University of Science and Technology, China), Hongce Zhang (The Hong Kong University of Science and Technology, China), and Yangdi Lyu (The Hong Kong University of Science and Technology, China)</i>	

ATPG-Based Weighted Scan Chain Control for Programmable Low-Power LBIST .....	306
<i>Yumei Hu (Huazhong University of Science and Technology, China),      Hairui Cai (Huazhong University of Science and Technology, China),      Xiaohui Xue (Huazhong University of Science and Technology, China),      Yaning Wang (Huawei Technologies Co., Ltd, China), Yu Huang (Huawei Technologies Co., Ltd, China), Zhipeng Lv (Huazhong University of Science and Technology, China), Zhouxing Su (Huazhong University of Science and Technology, China), Zezhong Wang (Huawei Technologies Co., Ltd, China), and Xing Wang (Huawei Technologies Co., Ltd, China)</i>	
FitFuzz: Depth-Oriented Coverage-Guided Fuzzing via Fitness-Based Seed Scheduling .....	315
<i>Venkat Nitin Patnala (George Mason University, USA) and Sai Manoj Pudukotai Dinakarrao (George Mason University, USA)</i>	
DASICS: Efficient In-Process Protection with Hardware-Assisted Dynamic Compartmentalization .....	319
<i>Yue Jin (CAS, China; University of Chinese Academy of Sciences (UCAS), China; Zhongguancun Laboratory, China), Yibin Xu (CAS, China; University of Chinese Academy of Sciences (UCAS), China), Han Wang (CAS, China; University of Chinese Academy of Sciences (UCAS), China), Chengyuan Zhang (CAS, China; University of Chinese Academy of Sciences (UCAS), China), Tianyi Huang (CAS, China; University of Chinese Academy of Sciences (UCAS), China), Tianyue Lu (CAS, China; University of Chinese Academy of Sciences (UCAS), China), and Mingyu Chen (CAS, China; University of Chinese Academy of Sciences (UCAS), China; Zhongguancun Laboratory, China)</i>	

## Session 5C - Quantum Systems Security: Threats, Forensics & Defenses

Security Evaluation of Quantum Circuit Split Compilation under an Oracle-Guided Attack .....	323
<i>Hongyu Zhang (Lehigh University, USA) and Yuntao Liu (Lehigh University, USA)</i>	
Forensics of Error Rates of Quantum Hardware .....	329
<i>Rupshali Roy (Penn State University, USA) and Swaroop Ghosh (Penn State University, USA)</i>	
QTIME: A Machine Learning Framework for Timing Side-Channel Analysis in Quantum Circuit Simulators .....	335
<i>Ben Dong (University of California, Merced, USA), Hui Feng (University of California, Merced, USA), and Qian Wang (University of California, Merced, USA)</i>	
Recovering QSVT Polynomials from Side-Channel Information on Quantum Computers .....	342
<i>Kidus Tessma (Northwestern University, USA), Hrvoje Kukina (TU Wien, Austria), and Jakub Szefer (Northwestern University, USA)</i>	
Concolic Testing for Quantum Compilers .....	348
<i>Navnil Choudhury (Rensselaer Polytechnic Institute, USA), Ameya Bhave (University of Texas at Dallas, USA), and Kanad Basu (Rensselaer Polytechnic Institute, USA)</i>	

UQ-VarQA: Benchmarking and Characterizing NISQ Computers Through Uncertainty Quantification of Variational Quantum Algorithms .....	356
---	-----

*Priyabrata Senapati (Kent State University, USA; Pacific Northwest National Laboratory, USA), Shengye Zhu (George Washington University, USA), Bo Peng (Pacific Northwest National Laboratory, USA), Bo Fang (University of Texas at Arlington, USA), and Qiang Guan (Kent State University, USA; Miami University, USA)*

## Session 6A - Architecture-Aware Compilation and Scheduling

NaviMap: Partial Order-Guided Neural Architecture via Deep Q-Networks for Efficient CGRA Mapping .....	364
--	-----

*Mingyang Kou (University of Science and Technology Beijing, China), Jun Zeng (Tsinghua University, China), Xinyu Peng (University of Science and Technology Beijing, China), Weiqing Ji (University of Science and Technology Beijing, China), and Hailong Yao (University of Science and Technology Beijing, China; Ministry of Education, China)*

IasRT: Interference-Aware and SLO-Driven GPU Scheduling for Real-Time DNN Inference .....	372
---	-----

*Heming Zhong (Sun Yat-sen University, China), Jinhui Wei (Sun Yat-sen University, China), Yujia Fu (Sun Yat-sen University, China), Dan Huang (Sun Yat-sen University, China), and Yutong Lu (Sun Yat-sen University, China)*

AICAWS: Arithmetic Intensity Based Cache-Conscious Adaptive Warp Scheduler .....	380
--	-----

*Bo Yuan (National University of Defense Technology, China), Sheng Liu (National University of Defense Technology, China), Zekun Jiang (National University of Defense Technology, China), Jianfeng Cui (National University of Defense Technology, China), and Yang Guo (National University of Defense Technology, China)*

A Dynamic Virtual Memory Management System for LLMs on AI Chips .....	389
---	-----

*Gaolin Wei (Hong Kong Polytechnic University, Hong Kong), Zhaorui Zhang (Hong Kong Polytechnic University, Hong Kong), Jiaqi Xu (Hong Kong Polytechnic University, Hong Kong), Chen Jason Zhang (Hong Kong Polytechnic University, Hong Kong), Xin Yao (The University of Hong Kong, Hong Kong), and Benben Liu (The University of Hong Kong, Hong Kong)*

## Session 6B - Timing-Driven & Physically-Aware Optimization

CPA-Remap: Critical-Path-Based Physically Aware Remapping Framework for Timing Optimization .....	393
---	-----

*Mingxiao He (National University of Defense Technology(NUDT), China), Pengcheng Huang (National University of Defense Technology(NUDT), China), Zhenyu Zhao (National University of Defense Technology(NUDT), China), and Peiyun Bian (National University of Defense Technology(NUDT), China)*

Threshold Voltage Tuning Technique for Leakage Power Recovery .....	401
---	-----

*Jaejoon Yoon (Seoul National University, Republic of Korea) and Taewhan Kim (Seoul National University, Republic of Korea)*

Timing-Driven Global Placement with Entropy-Mobility Guided Pin-to-Pin Weighting .....	409
<i>Youzhi Zheng (Southwest University of Science and Technology, China),     Zhengjie Zhao (Southwest University of Science and Technology, China),     Linhao Lu (Southwest University of Science and Technology, China),     Xiaodong Zhu (Southwest University of Science and Technology, China),     Wenxin Yu (Southwest University of Science and Technology, China), and     Jingwei Lu (TikTok, United States)</i>	
Timing-Driven Multi-Bit Flip-Flop Allocation Utilizing Design-Technology Co-Optimization Techniques .....	416
<i>Yeongyeong Shin (Seoul National University, Republic of Korea),     Sehyeon Chung (Seoul National University, Republic of Korea), and     Taewhan Kim (Seoul National University, Republic of Korea)</i>	

## Session 6C - Efficient and Secure Generative AI

GAN-BiLSTM-HDC: A Hybrid Framework for Robust and Hardware-Efficient Malware Detection ....	
424	
<i>Emilien Meyer (University of Louisiana at Lafayette, USA), Abu Kaisar     Mohammad Masum (University of Louisiana at Lafayette, USA), Mehran     Moghadam (Case Western Reserve University, USA), Lida Kouhalvandi     (Dogus University, Turkiye), Gourav Datta (Case Western Reserve     University, USA), Sercan Aygun (University of Louisiana at Lafayette,     USA), and M. Hassan Najafi (Case Western Reserve University, USA)</i>	
LM-Fix: Lightweight Bit-Flip Detection and Rapid Recovery Framework for Language Models ....	432
425	
<i>Ahmad Tahmasivand (George Mason University, USA), Noureldin Zahran     (George Mason University, USA), Saba Al-Sayouri (The National     Institutes of Health, USA), Mohammed Fouad (Compumacy for Artificial     Intelligence Solutions, Egypt), and Khaled N. Khasawneh (George Mason     University, USA)</i>	
FaRAccel: FPGA-Accelerated Defense Architecture for Efficient Bit-Flip Attack Resilience in Transformer Models .....	441
426	
<i>Najmeh Nazari (University of California, USA), Banafsheh Saber     Latibari (University of Arizona, USA), Elahe Hosseini (University of     California, USA), Fatemeh Movafagh (Simon Fraser University, Canada),     Chongzhou Fang (Rochester Institute of Technology, USA), Hosein     Mohammadi Makrani (University of California, USA), Kevin Immanuel     Gubbi (University of California, USA), Abhijit Mahalanobis     (University of Arizona, USA), Setareh Rafatirad (University of     California, USA), Hossein Sayadi (California State University, USA),     and Houman Homayoun (University of California, USA)</i>	
Hammering the Diagnosis: Rowhammer-Induced Stealthy Trojan Attacks on ViT-Based Medical Imaging .....	450
427	
<i>Banafsheh Saber Latibari (University of Arizona, USA), Najmeh Nazari     (University of California, USA), Hossein Sayadi (California State     University Long Beach, USA), Houman Homayoun (University of     California, USA), and Abhijit Mahalanobis (University of Arizona, USA)</i>	

## Session 7A - Efficient Data Storage

Hybrid-Rewrite: A Rewriting Framework for Hybrid Deduplication and Delta Compression .....	458
<i>Qiao Li (Nanchang University, China), Hong Jiang (University of Texas at Arlington, USA), Zichen Xu (Nanchang University, China), Yucheng Zhang (Nanchang University, China), Junyun Wu (Nanchang University, China), and Puchen Lu (Nanchang University, China)</i>	
NatSep: Little-to-No Overhead Data Separation for Log-Structured Storage Using Native Information .....	466
<i>jinlong wang (Huazhong University of Science and Technology), zhipeng tan (Huazhong University of Science and Technology), yang xiao (Huazhong University of Science and Technology), wenjie qi (Huazhong University of Science and Technology), shikai tan (Huazhong University of Science and Technology), and ying yuan (Huazhong University of Science and Technology)</i>	
The Logic of Fingerprint Upgrade in Deduplicated Storage .....	474
<i>Cai Deng (Harbin Institute of Technology, China), Boju Chen (Harbin Institute of Technology, China), Philip Shilane (Dell Technologies, USA), Xiangyu Zou (Harbin Institute of Technology, China; Pengcheng Laboratory, China), Wen Xia (Harbin Institute of Technology, China; Pengcheng Laboratory, China), and Hao Hu (Harbin Institute of Technology, China)</i>	
Pixel-DNA: Increasing Robustness of Approximate DNA Storage for Images by Using Hierarchical Deduplication .....	482
<i>Alex Sensintaffar (The University of Texas at Dallas, USA), David Du (University of Minnesota, USA), and Bingzhe Li (The University of Texas at Dallas, USA)</i>	
TCFlash: In-Flash Bulk Bitwise Processing via Dynamic Sensing and TLC Encoding in 3D NAND .	491
<i>Habib Ur Rahman (Colorado State University, United States), Suresh Tharini (Colorado State University, United States), Sudeep Pasricha (Colorado State University, United States), and Biswajit Ray (Colorado State University, United States)</i>	
Minimizing Read Disturb via Localized Page Allocation for Modern NAND Flash-Based SSDs .....	495
<i>Joonseong Hwang (Sungkyunkwan University, Republic of Korea), Minkyu Choi (Sungkyunkwan University, Republic of Korea), Minjin Park (Sungkyunkwan University, Republic of Korea), Jihun Yoon (Sungkyunkwan University, Republic of Korea), Yoonho Jang (Sungkyunkwan University, Republic of Korea), and Seokin Hong (Sungkyunkwan University, Republic of Korea)</i>	

## Session 7B - Fault Tolerance and Resilience in Systems

Laser and Radiation Testing of Compiler-Based Protection for Multi-Bit Upsets .....	499
<i>Davide Baroffio (Politecnico di Milano, Italy), Tomas Antonio López (Politecnico di Milano, Italy), Federico Reghennzani (Politecnico di Milano, Italy), and William Fornaciari (Politecnico di Milano, Italy)</i>	

Masked Gadgets for Integer-Floating-Point Conversion with Applications to Falcon .....	507
<i>Shuyi Chen (Nanjing University of Science and Technology, China),     Jingdian Ming (Nanjing University of Science and Technology, China),     Yuejun Liu (Nanjing University of Science and Technology, China),     Yiwen Gao (Nanjing University of Science and Technology, China), and     Yongbin Zhou (Nanjing University of Science and Technology, China;     Chinese Academy of Sciences, China)</i>	
WSSR: Weight Set Segmentation and Recovery for Fault Resilient Transformers .....	515
<i>Ntsee Ndingwan (University of Delaware, USA) and Chengmo Yang (University of Delaware, USA)</i>	
ECOLogic: Enabling Circular, Obfuscated, and Adaptive Logic via eFPGA-Augmented SoCs .....	523
<i>Ishraq Tashdid (University of Central Florida, USA), Dewan Saiham (University of Central Florida, USA), Nafisa Anjum (Louisiana State University, USA), Tasnuva Farheen (Louisiana State University, USA), and Sazadur Rahman (University of Central Florida, USA)</i>	
Enhancing Key-Recovery Chosen-Ciphertext Side-Channel Attacks on NTRU Using LDPC .....	528
<i>Denis Nabokov (Lund University, Sweden), Xiaofei Tong (Beijing University of Posts and Telecommunications, China), and Qian Guo (Lund University, Sweden)</i>	

## Session 8A - Advanced Hardware Acceleration Trends

A Photonic Accelerator for Deep Learning Training .....	532
<i>Yuan Li (Singapore Institute of Technology, Singapore)</i>	
FINEA: An Efficient Neural Network Accelerator Exploiting Factorized Input Features .....	540
<i>Yujin Kim (Korea University, South Korea), Chanhung Jeong (Korea University, South Korea), Yunho Oh (Korea University, South Korea), Myung Kuk Yoon (Ewha Womans University, South Korea), and Gunjae Koo (Korea University, South Korea)</i>	
Flame: A Multiplier-Free LLM Accelerator with Dynamic Block Floating Point .....	549
<i>Ao Lyu (Chinese Academy of Sciences; University of Chinese Academy of Sciences), Haishuang Fan (Chinese Academy of Sciences; University of Chinese Academy of Sciences), and Guihai Yan (Chinese Academy of Sciences; YUSUR Technology Co., Ltd.)</i>	
Hermes: Accelerating Packet Processing in DPU with Neural Network .....	558
<i>Rui Meng (Chinese Academy of Sciences; University of Chinese Academy of Sciences), Xinyu Chen (The Hong Kong University of Science and Technology, (Guangzhou)), Hanyue Lin (Chinese Academy of Sciences; University of Chinese Academy of Sciences), Jingya Wu (Chinese Academy of Sciences), Wenyan Lu (Chinese Academy of Sciences; YUSUR Technology Co., Ltd.), Xiaowei Li (Chinese Academy of Sciences; Zhongguancun Laboratory), and Guihai Yan (Chinese Academy of Sciences; YUSUR Technology Co., Ltd.)</i>	

ASMA: An Anisotropy Scaling Memristor-Based Accelerator for LLM Inference .....	562
<i>Zijian Xiong (Huazhong University of Science and Technology), Xiangrui Yang (Huazhong University of Science and Technology), Yuhang Zhang (Huazhong University of Science and Technology), Yue Zhou (The Hong Kong Polytechnic University), Jianguo Yang (the Chinese Academy of Science), Yaoyu Tao (Peking University), Xiangshui Miao (Huazhong University of Science and Technology), and Yuhui He (Huazhong University of Science and Technology)</i>	

## Session 8B - AI-Assisted RTL & HLS Code Generation

RTL Bench: A Multi-Dimensional Benchmark Suite for Evaluating LLM-Generated RTL Code .....	566
<i>Zhigang Fang (National University of Defense Technology, China; Key Laboratory of Advanced Microprocessor Chips and Systems, China), Renzhi Chen (Academy of Military Sciences, China), Yang Guo (National University of Defense Technology, China; Key Laboratory of Advanced Microprocessor Chips and Systems, China), Huadong Dai (Academy of Military Sciences, China), and Lei Wang (Academy of Military Sciences, China; Qiyuan Lab, China)</i>	
SAGE-HLS: Syntax-Aware AST-Guided LLM for High-Level Synthesis Code Generation .....	574
<i>M Zafir Sadik Khan (University of Central Florida, USA), Nowfel Mashnoor (University of Central Florida, USA), Mohammad Akyash (University of Central Florida, USA), Kimia Azar (University of Central Florida, USA), and Hadi Kamali (University of Central Florida, USA)</i>	
LLM4MCU-Onto : Leveraging LLMs for Automated Ontology Generation from Microcontroller Reference Manual .....	582
<i>Asmita Asmita (University of California, USA), Grisha Bandodkar (Carnegie Mellon University, USA), Sujan Ghimire (The University of Arizona, USA), Shaurya Srivastav (University of California, USA), Soheil Salehi (The University of Arizona, USA), and Houman Homayoun (University of California, USA)</i>	
LLM-Driven Code Generation for Neural Networks on FPGAs: Bridging Python and HLS .....	590
<i>Rupesh Raj Karn (New York University, UAE), Johann Knechtel (New York University, UAE), Ramesh Karri (New York University, UAE), and Ozgur Sinanoglu (New York University, UAE)</i>	

## Session 9A - Processor-Based Solutions 1

TROOP: At-The-Roofline Performance for Vector Processors on Low Operational Intensity Workloads .....	594
<i>Navaneeth Kunhi Purayil (ETH Zürich, Switzerland), Diyou Shen (ETH Zürich, Switzerland), Matteo Perotti (ETH Zürich, Switzerland), and Luca Benini (ETH Zürich, Switzerland; Università di Bologna, Italy)</i>	

RVME: An Efficient Matrix Engine Design Based on Matrix Extension of RISC-V .....	602
<i>Wanqi Chen (Shanghai Jiao Tong University, China), Weidong Yang (Shanghai Jiao Tong University, China), Yiming Guo (Shanghai Jiao Tong University, China), Jing Qiu (Alibaba DAMO Academy, China), Renpei Wang (Shanghai Jiao Tong University, China), Jianfei Jiang (Shanghai Jiao Tong University, China), Naifeng Jing (Shanghai Jiao Tong University, China), and Qin Wang (Shanghai Jiao Tong University, China)</i>	
TeraNoC: A Multi-Channel 32-Bit Fine-Grained, Hybrid Mesh-Crossbar NoC for Efficient Scale-up of 1000+ Core Shared-L1-Memory Clusters .....	610
<i>Yichao Zhang (IIS, ETH Zürich; University of Bologna), Zexin Fu (IIS, ETH Zürich; University of Bologna), Tim Fischer (IIS, ETH Zürich), Yinrong Li (IIS, ETH Zürich), Marco Bertuletti (IIS, ETH Zürich), and Luca Benini (IIS, ETH Zürich; University of Bologna)</i>	
THENA: Torus Fully Homomorphic Encryption on Energy-Efficient Heterogeneous Architecture ..	618
<i>Yanze Wu (George Mason University, USA) and Md Tanvir Arafin (George Mason University, USA)</i>	
SSM-RDU: A Reconfigurable Dataflow Unit for Long-Sequence State-Space Models .....	626
<i>Sho Ko (Machines, Inc., USA) and Kunle Olukotun (Stanford University, USA)</i>	

## Session 9B - Advanced Hardware Design Flows & Synthesis Techniques

Optimization of Wire Pipelining and Channel Parallelism for 2D-Mesh NoC Physical Design .....	630
<i>Pei-Huan Tsai (Columbia University, USA), Maico Cassel dos Santos (Columbia University, USA), Joseph Zuckerman (Columbia University, USA), Kuan-Lin Chiu (Columbia University, USA), and Luca P. Carloni (Columbia University, USA)</i>	
Agile Design Flow for Cryptographic Hardware Accelerators .....	638
<i>Liming Deng (Fudan University, China), Guowei Zhu (Fudan University, China), Wei Cao (Fudan University, China), Xitian Fan (Fudan University, China), and Xuegong Zhou (Fudan University, China)</i>	
Decomposition Attack on Structural Logic Locking of Reversible Circuits .....	646
<i>Feng-Jie Chao (National Taiwan University of Science and Technology, Taiwan) and Yung-Chih Chen (National Taiwan University of Science and Technology, Taiwan)</i>	
Supporting Pipelined Memory Accesses in Processor Synthesis .....	654
<i>Essien Taylor (Northwestern University, United States), Colin Schilf (Northwestern University, United States), Sebastian Phemister (Northwestern University, United States), and Russ Joseph (Northwestern University, United States)</i>	

## Session 10A - Processor-Based Solutions 2

BNRV: A Lightweight SIMD Extension for Efficient BitNet Inference on RISC-V CPUs .....	658
<i>Zijun Jiang (The Hong Kong University of Science and Technology, China) and Yangdi Lyu (The Hong Kong University of Science and Technology, China)</i>	

Design and Evaluation of an N-Trace Compliant Hardware Tracer for RISC-V Processors .....	666
<i>Omer Karslioglu (Ozyegin University, Turkiye) and Ismail Akturk (Ozyegin University, Turkiye)</i>	
FlexIO: A Scalable IO Chiplet Architecture with Flexible Memory Controller Mapping .....	674
<i>Junpei Huang (University of Science and Technology of China, China; Chinese Academy of Sciences, China), Haobo Xu (Chinese Academy of Sciences, China), Ying Wang (Chinese Academy of Sciences, China), and Yinhe Han (Chinese Academy of Sciences, China)</i>	
Register Bridging: A Lightweight Microarchitectural Approach for Skipping Overhead Instructions in Distance-Based ISA Processors .....	682
<i>Fan Yang (The University of Tokyo, Japan), Toru Koizumi (Nagoya Institute of Technology, Japan), Jun Li (Nanjing University of Posts and Telecommunications, China), Shu Sugita (The University of Tokyo, Japan), Yuriko Yamauchi (The University of Tokyo, Japan), Ryota Shioya (The University of Tokyo, Japan), Junichiro Kadomoto (The University of Tokyo, Japan), and Hidetsugu Irie (The University of Tokyo, Japan)</i>	
XDMA: A Distributed, Extensible DMA Architecture for Layout-Flexible Data Movements in Heterogeneous Multi-Accelerator SoCs .....	690
<i>Fanchen Kong (KU Leuven, Belgium), Yunhao Deng (KU Leuven, Belgium), Xiaoling Yi (KU Leuven, Belgium), Ryan Antonio (KU Leuven, Belgium), and Marian Verhelst (KU Leuven, Belgium)</i>	

## Session 10B - System Level AI Optimization

AceHomo: Accelerating Privacy Preserving Inference through Dynamic Level Adjustment .....	694
<i>Hongyan Li (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; Zhongguancun Laboratory, China), Jinkai Zhang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Hang Lu (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; Zhongguancun Laboratory, China), and Xiaowei Li (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; Zhongguancun Laboratory, China)</i>	
HyperDrone: an Accurate, Robust, Fast, and Energy-Efficient Approach for Drone Classification .....	702
<i>Shriniwas Kulkarni (University of California San Diego, California), Flavio Ponzina (University of California San Diego, California ), and Tajana Rosing (University of California San Diego, California )</i>	
Access Frequency-Aware Storage Reduction for Deep Learning Recommendation Model .....	710
<i>Chia-Chun Wang (National Tsing Hua University, Taiwan), Chuan-Yao Lai (National Tsing Hua University, Taiwan), and Ren-Shuo Liu (National Tsing Hua University, Taiwan)</i>	
Recommendation-Expert Framework for Fast and Adaptive Scheduling in Computing Power Network .....	718
<i>Yu Chen (Shanghai Jiao Tong University, China) and Wenli Zheng (Shanghai Jiao Tong University, China)</i>	

Oak: A Fault-Tolerant Shared-Memory System Atop Memory-Semantic Fabrics .....	726
<i>Zhaoxiang Huang (Xiamen University, China), Jianqin Yan (Xiamen University, China), Hao Chen (Xiamen University, China), Jiaxin Li (National University of Defense Technology, China), and Yiming Zhang (Xiamen University, China; Shanghai Jiao Tong University, China)</i>	

## Session 11A - Emerging Trends 2

TLV-HGNN: Thinking Like a Vertex for Memory-Efficient HGNN Inference .....	730
<i>Dengke Han (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Duo Wang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Mingyu Yan (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Xiaochun Ye (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), and Dongrui Fan (Chinese Academy of Sciences, China)</i>	
SageSC: Accelerating GraphSAGE Minibatch Inference on Memory-Intensive Graphs .....	738
<i>Yuchen Gui (University of Science and Technology of China, China), Wei Yuan (University of Science and Technology of China, China), Qizhe Wu (University of Science and Technology of China, China), Huawei Liang (University of Science and Technology of China, China), Letian Zhao (University of Science and Technology of China, China), Linfeng Tao (University of Science and Technology of China, China), Zhongguang Xu (University of Science and Technology of China, China), and Xi Jin (University of Science and Technology of China, China)</i>	
TIPS: Augment Memory Tagging to Defend Against Prefetcher Side Channels .....	746
<i>Yubiao Huang (University of Chinese Academy of Sciences), Peinan Li (University of Chinese Academy of Sciences), Huan Qiao (University of Chinese Academy of Sciences), Yunkai Bai (University of Chinese Academy of Sciences), Shiwen Wang (University of Chinese Academy of Sciences), Dan Meng (University of Chinese Academy of Sciences), and Rui Hou (University of Chinese Academy of Sciences)</i>	
In-DRAM True Random Number Generation Using Simultaneous Multiple-Row Activation: An Experimental Study of Real DRAM Chips .....	754
<i>Ismail Emir Yuksel (ETH Zurich), Ataberk Olgun (ETH Zurich), Fatma Nisa Bostanci (ETH Zurich), Oguzhan Canpolat (ETH Zurich), Geraldo F. Oliveira (ETH Zurich), Mohammad Sadrosadati (ETH Zurich), Abdullah Giray Yaglikci (ETH Zurich and CISPA), and Onur Mutlu (ETH Zurich)</i>	
Adaptive ML-KEM: A Configurable HW-SW Architecture for Post-Quantum Cryptography .....	764
<i>Wenkai Wang (Shandong University, China), Chao Liu (Shandong University, China), Zhe Sun (Shandong University, China), Lei Ju (Shandong University, China), and Zimeng Zhou (Shandong University, China)</i>	

## Session 11B - SS-2: GenAI Meets Silicon: LLMs in Hardware Design, Verification, and Security

FV-PAL: Scalable Formal Verification through Partitioning and LLM-Guided Property Generation .....	768
<i>Sudipta Paria (University of Florida, USA), Aritra Dasgupta (University of Florida, USA), Dinesh Reddy Ankireddy (University of Florida, USA), Prabuddha Chakraborty (University of Maine, USA), and Swarup Bhunia (University of Florida, USA)</i>	
Tracing the Logic: Evaluating LLM Reasoning Paths in RTL Generation .....	776
<i>Matthew DeLorenzo (Texas A&amp;M University, USA), Kevin Tieu (Texas A&amp;M University, USA), and Jeyavijayan Rajendran (Texas A&amp;M University, USA)</i>	
MALLS: Multi-Agent LLMs for Synthetic Hardware Vulnerability Generation and Detection .....	782
<i>Jonti Talukdar (Arizona State University), Agastya Seth (Arizona State University), Sanmitra Banerjee (Arizona State University), Farshad Firouzi (Arizona State University), and Krishnendu Chakrabarty (Arizona State University)</i>	
CircuitGuard: Mitigating LLM Memorization in RTL Code Generation Against IP Leakage .....	790
<i>Nowfel Mashnoor (University of Central Florida, USA), Mohammad Akyash (University of Central Florida, USA), Hadi Kamali (University of Central Florida, USA), and Kimia Azar (University of Central Florida, USA)</i>	

## Session 12A - Specialized High-Performance Computing

FlashMP: Fast Discrete Transform-Based Solver for Preconditioning Maxwell's Equations on GPUs .....	798
<i>Haoyuan Zhang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Yaqian Gao (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Xinxin Zhang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Jialin Li (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Runfeng Jin (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Yidong Chen (Tsinghua University, China), Feng Zhang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Wu Yuan (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Wenpeng Ma (Xinyang Normal University, China), Shan Liang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Jian Zhang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), and Zhonghua Lu (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China)</i>	
MH-SpGEMM: Efficient Sparse General Matrix-Matrix Multiplication on Modern GPUs via Masking and Hashing Cooperative Optimization .....	807
<i>Shuang Yang (Southwest University of Science and Technology, China), Yaobin Wang (Southwest University of Science and Technology, China), Ling Li (Southwest University of Science and Technology, China), Qian Peng (Southwest University of Science and Technology, China), and Qiong Yu (Southwest University of Science and Technology, China)</i>	

TensTFM: Efficient Total Focusing Method for Ultrasonic Array Imaging on Dataflow Accelerators .....	815
<i>Jieran Zhang (Peking University, China), Bizhao Shi (Peking University, China), and Guojie Luo (Peking University, China)</i>	
Design of an Online Surface Code Decoder Using Union-Find Algorithm .....	823
<i>Takuya Kasamura (The University of Tokyo, Japan), Junichiro Kadomoto (The University of Tokyo, Japan), and Hidetsugu Irie (The University of Tokyo, Japan)</i>	
Early Termination with Activation Sign Prediction for Energy-Efficient CNN Inference Using Sum-of-Power-of-Two Quantization .....	831
<i>Emir Mehmet Eryilmaz (Ozyegin University, Turkiye), Selim Sandal (Ozyegin University, Turkiye), and Ismail Akturk (Ozyegin University, Turkiye)</i>	

## Session 12B - Sustainable Hardware Accelerators with Integrated Electro-Photonics

Sustainable Acceleration of Generative AI Neural Network Models with Silicon Photonics .....	835
<i>Tharini Suresh (Colorado State University, CO), Salma Afifi (Colorado State University, CO ), and Sudeep Pasricha (Colorado State University, CO)</i>	
Toward Lifelong-Sustainable Electronic-Photonic AI Systems via Extreme Efficiency, Reconfigurability, and Robustness .....	843
<i>Ziang Yin (Arizona State University), Hongjian Zhou (Arizona State University), Chetan Choppali Sudarshan (Arizona State University), Vidya Chhabria (Arizona State University), and Jiaqi Gu (Arizona State University)</i>	
SUSTAINPHOT: Sustainable Large-Scale AI Training using Analog Silicon Photonic Accelerators .....	851
<i>Dharanidhar Dang (University of Texas at San Antonio, USA)</i>	
Scaling Up the Sustainability of Photonic Tensor Cores with Device-Circuit-Signaling Co-Design .....	859
<i>Ishan Thakkar (University of Kentucky, USA), Sairam Sri Vatsavai (University of Kentucky, USA), Venkata Sai Praneeth Karempudi (University of Kentucky, USA), and Oluwaseun Adewunmi Alo (University of Kentucky, USA)</i>	

## Tutorials

Representation Learning for Digital Integrated Circuit Design Automation .....	867
<i>Pratik Shrestha (Drexel University, Pennsylvania ) and Ioannis Savidis (Drexel University, Pennsylvania)</i>	
Engineering Privacy at the Edge: A Practical Guide to Differential Privacy in System Architectures .....	872
<i>Olivera Kotevska (Oak Ridge National Laboratory), Wenjun Yang (University of Washington Tacoma), and Eyhab Al-Masri (University of Washington Tacoma)</i>	

## **Author Index**