Technical Program: 28 October 2024

7:00 – 8:00
Registration
Room: Grand Ballroom Foyer

8:00 – 8:30
Opening Ceremony
Room: Salons 4-8

8:30 – 9:30
Keynote: Design and Design Automation for a Future Generation of Chips
Philip Wong, Stanford University
Room: Salons 4-8

9:30 – 10:00
Coffee Break | Exhibits
Room: Salons D-H

10:00 – 12:00
Tutorial: Security of Quantum Computing Hardware and Architectures
Room: Salons 1-3

10:00 – 12:00
Tutorial: Security of Quantum Computing Hardware and Architectures
Room: Salons 1-3

10:00 – 12:00
Special Session: The Dawn of Domain-Specific Hardware System for Autonomous Machines
Room: Lincoln/Holland/Columbia

10:00 – 12:00
LLM4HWDesign Contest
Room: Salons A-C

10:00 – 12:00
Special Session: Exploring Attack Vectors and Resilient Defense Strategies in Microelectronics A Special Session on Hardware Security
Room: Skylands/Gateway

10:00 – 12:00
Special Session: Computing over Encrypted Data: Novel Acceleration of Fully Homomorphic Encryption on Hardware Platforms
Room: Essex/Liberty

10:00 - 11:00
Advanced Partitioning and Floorplanning
Room: Salons F-H

In this session, we explore cutting-edge techniques for partitioning and floorplanning. In the first paper, authors introduce GenPart that introduces a hypergraph partitioner using a generative model, outperforming traditional methods. Then in the second paper, TopoOrderPart is introduced that focuses on scheduling-driven partitioning, balancing topological order while minimizing cut size. The third paper delves into analytical-based approaches for rectilinear floorplanning, allowing shape-adjustable modules, and finally the fourth paper introduces JigsawPlanner, that handles complex-shaped rectilinear modules. Join us to discover the latest innovations in VLSI design planning.

10:00
1338: A Hypergraph Partitioner Utilizing a Novel Graph Generative Model
Magi Chen (National Tsing Hua University); Ting-Chi Wang (National Tsing Hua University)

10:15
1554: TopoOrderPart: a Multi-level Scheduling-Driven Partitioning Framework for Processor-Based Emulation
shunyang bi (Xidian University); jing tang (Xidian University); Hailong You (Xidian University); haonan wu (Xidian University); Cong Li (Xidian University); richard sun (S2C Inc.)

10:30
1440: Modern Fixed-Outline Floorplanning with Rectilinear Soft Modules
Yu-Yang Chen (National Taiwan University); Yi-Chen Lin (National Taiwan University); Tzu-Han Hsu (National Taiwan University); Iris Hui-Ru Jiang (National Taiwan University); Tung-Chieh Chen (Synopsys); Tai-Chen Chen (Synopsys, Inc.); Hua-Yu Chang (Synopsys, Inc.)

10:45
805: JigsawPlanner: Jigsaw-like Floorplanner for Eliminating Whitespace and Overlap among Complex Rectilinear Modules
Xingbo Du (Shanghai Jiao Tong University); Ruizhe Zhong (Shanghai Jiao Tong University); Shixiong Kai (Huawei Noah's Ark Lab); Zhentao Tang (Huawei Noah's Ark Lab); Siyuan Xu (Huawei Noah's Ark Lab); Jianye Hao (Tianjin University); Mingxuan Yuan (Huawei Noah's Ark Lab); Junchi Yan (Shanghai Jiao Tong University)

Technical Program: 28 October 2024

11:00 - 12:00
State-of-the-Art Placement
Room: Salons F-H

In this session, we explore innovative techniques for the placement problem. The first paper introduces a novel design flow that integrates gate sizing with global placement, leveraging differentiable timing and leakage power objectives. In the second paper, an approach named SysMix to tailor a mixed-size placement for systolic array in hardware accelerator design. It has scalable speedup and wirelength reduction.

Then, in the third paper, the authors delve into a multilevel framework addressing fence region constraints during cell placement by eliminating inappropriate clustering and gradually refining placement regions. Finally, in the fourth paper, a timing-driven global placement framework is introduced that considers both graph and path information.

11:00
650: Fusion of Global Placement and Gate Sizing with Differentiable Optimization
Yufan Du (Peking University); Zizheng Guo (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

11:15
734: SysMix: Mixed-Size Placement for Systolic-Array-Based Hierarchical Designs
Donghao Fang (Texas A&M University); Hailiang Hu (Texas A&M University); Wuxi Li (AMD); Bo Yuan (Rutgers University); Jiang Hu (Texas A&M University)

11:30
728: An Effective Analytical Placement Approach to Handle Fence Region Constraint
Jai-Ming Lin (Department of Electrical Engineering, National Cheng Kung University); Wei-Yuan Lin (Department of Electrical Engineering, National Cheng Kung University); Yung-Chen Chen (Department of Electrical Engineering, National Cheng Kung University); Pin-Yu Chen (Academy of Innovative Semiconductor and Sustainable Manufacturing Program on Integrated Circuit Design, National Cheng Kung University); Chen-Fa Tsai (Design Service Division of GUC); De-Shiun Fu (Design Service Division of GUC); Che-Li Lin (Design Service Division of GUC)

11:45
1329: Hybrid Modeling and Weighting for Timing-driven Placement with Efficient Calibration
Bangqi Fu (The Chinese University of Hong Kong); Lixin Liu (The Chinese University of Hong Kong); Martin Wong (The Chinese University of Hong Kong); Evangeline Young (The Chinese University of Hong Kong)

12:00 – 13:30
Synopsys Invited Speaker Lunch
Room: Salons 4-8

13:30 – 15:30
Tutorial: Hardware Security Trust and Verification
Room: Salons 1-3

Technical Program: 28 October 2024

13:30 – 15:30
Special Session: Advancing AI: Cross-disciplinary Insights into Next-Gen Tools, Tech & Architectures
Room: Lincoln/Holland/Columbia

13:30 – 15:30
Special Session: AI4HLS: New Frontiers in High-Level Synthesis Augmented with Artificial Intelligence
Room: Salons A-C

13:30 - 14:30
EDA for Quantum
Room: Skylands/Gateway

13:30
730: Barber: Balancing Thermal Relaxation Deviations of NISQ Programs by Exploiting Bit-Inverted Circuits
Enhyeok Jang (Yonsei University); Seungwoo Choi (Yonsei University); Youngmin Kim (Yonsei University); Jeewoo Seo (Yonsei University); Won Woo Ro (Yonsei University)

13:45
510: Quantum State Preparation Circuit Optimization Exploiting Don't Cares
Hanyu Wang (ETH Zurich); Daniel Bochen Tan (University of California, Los Angeles); Jason Cong (UCLA)

14:00
765: ReCon: Reconfiguring Analog Rydberg Atom Quantum Computers for Quantum Generative Adversarial Networks
Nicholas S. DiBrita (Rice University); Daniel Leeds (Rice University); Yuqian Huo (Rice University); Jason Ludmir (Rice University); Tirthak Patel (Rice University)

14:15
1444: SMT-based Layout Synthesis for Silicon-based Quantum Computing with Crossbar Architecture
Sheng-Tan Huang (National Taiwan University of Science and Technology); Ying-Jie Jiang (National Taiwan University of Science and Technology); Shao-Yun Fang (National Taiwan University of Science and Technology); Chung-Kuan Cheng (UCSD)

13:30 - 14:30
Techniques for reliability modeling and analys
Room: Essex/Liberty

This session focuses on modeling and analysis techniques aimed at improving modeling accuracy and reliability.

The first paper demonstrates improved process modeling using a combination of neural networks and ordinary differential equations.

The second paper proposes a customized large-scale multimodal model, called FabGPT, designed to improve IC manufacturing processes.

The third paper proposes a rigorous analysis framework for yield estimation based on variational importance sampling, and the last paper proposes a robust ASIC placer that can handle hybrid region constraints via a unified multi-electrostatic formulation.

13:30
514: A Neural-Ordinary-Differential-Equations Based Generic Approach for Process Modeling in DTCO: A Case Study in Chemical-Mechanical Planarization and Copper Plating
Yue Qian (EDA Center, Institute of Microelectronics, Chinese Academy of Sciences and University of Chinese Academy of Sciences); Lan Chen (EDA Center, Institute of Microelectronics, Chinese Academy of Sciences and University of Chinese Academy of Sciences)

13:45
999: FabGPT: An Efficient Large Multimodal Model for Complex Wafer Defect Knowledge Queries
Yuqi Jiang (Zhejiang University); Xudong Lu (Zhejiang University); Qian Jin (Zhejiang University); QI SUN (Zhejiang University); Hanming Wu (Zhejiang University); Cheng Zhuo (Zhejiang University)

14:00
656: Beyond the Yield Barrier: Variational Importance Sampling Yield Analysis
Yanfang Liu (Beihang University); Lei He (University of California, Los Angeles); Wei W. Xing (The University of Sheffield)

14:15
803: BPINN-EM: Fast Stochastic Analysis of Electromigration Damage using Bayesian Physics-Informed Neural Networks
Subed Lamichhane (University of California, Riverside); Mohammadamir Kavousi (University of California, Riverside); Sheldon Tan (University of California at Riverside)

13:30 - 14:30
Layout and Cell Optimization
Room: Salons F-H

In this session, we explore cutting-edge techniques for VLSI design and cell optimization. In the first paper, the authors discuss an optimal method for synthesizing area-optimal multi-row standard cells, integrating transistor folding, row partitioning, and transistor placement. The second work introduces ATPlace2.5D, an analytical thermal-aware chiplet placement framework for large-scale 2.5D-ICs. It balances wirelength and temperature. In the third paper, the authors present novel approaches for 3D SRAM arrays: wordline and bitline folding. These designs achieve remarkable reductions in footprint, improved speed, and energy efficiency. Finally, the fourth paper proposes MAXCell, a PPA-directed standard cell layout optimization framework using anytime MaxSAT, surpassing wire length optimization studies.

13:30
1232: Optimal Layout Synthesis of Multi-Row Standard Cells for Advanced Technology Nodes
Sehyeon Chung (Seoul National University); Hyunbae Seo (Seoul National University); Handong Cho (Seoul National University); Kyumyung Choi (Seoul National University); Taewhan Kim (Seoul National University)

13:45
520: ATPlace2.5D: Analytical Thermal-Aware Chiplet Placement Framework for Large-Scale 2.5D-IC
Qipan Wang (Peking University); Xueqing Li (Peking University); Tianyu Jia (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:00
1386: Multi-Tier 3D SRAM Module Design: Targeting Bit-Line and Word-Line Folding
Aditya S. Iyer (Georgia Institute of Technology); Daehyun Kim (Georgia Institute of Technology); Saibal Mukhopadhyay (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech)

14:15
804: MAXCell: PPA-Directed Multi-Height Cell Layout Routing Optimization using Anytime MaXSAT with Constraint Learning
Jiun-Cheng Tsai (Mediatek); Wei-Min Hsu (Mediatek); Yun-Ting Hsieh (Mediatek); Yu-Ju Li (Mediatek); Wei Huang (Mediatek); CN Ho (Mediatek); Hsuan-Ming Huang (Mediatek); Jen-Hang Yang (Mediatek); Heng-Liang Huang (Mediatek); Aaron C. -W. Liang (National Yang Ming Chiao Tung University); Charles H. -P. Wen (National Yang Ming Chiao Tung University)

14:30 - 15:30
Quantum Simulation and Quantum Cloud
Room: Skylands/Gateway

14:30
827: Accelerating Quantum Circuit Simulation with Symbolic Execution and Loop Summarization
Tian-Fu Chen (Graduate School of Advanced Technology, National Taiwan University); Yu-Fang Chen (Academia Sinica); Jie-Hong Roland Jiang (National Taiwan University); Sara Jobranova (Brno University of Technology); Ondrej Lengal (Brno University of Technology)

14:45
1387: Detecting Fraudulent Services on Quantum Cloud Platforms via Dynamic Fingerprinting
Jindi Wu (William & Mary); Tianjie Hu (William & Mary); Qun Li (William & Mary)

15:00
682: On Reducing the Execution Latency of Superconducting Quantum Processors via Quantum Job Scheduling
Wenjie Wu (Shanghai Jiao Tong University); Yiquan Wang (Shanghai JiaoTong University); Ge Yan (Shanghai Jiao Tong University); Yuming Zhao (Shanghai Jiao Tong University); Bo Zhang (Shanghai Artificial Intelligence Laboratory); Junchi Yan (Shanghai Jiao Tong University)

15:15
862: A Hardware-Aware Gate Cutting Framework for Practical Quantum Circuit Knitting
Xiangyu Ren (University of Edinburgh); Mengyu Zhang (Tencent Quantum Laboratory); Antonio Barbalace (University of Edinburgh)

14:30 - 15:30
Optimizations in lithography and physical design
Room: Essex/Liberty

This session will showcase the latest developments relevant to the ever-important topics of lithography and physical design.

The first paper proposes a multi-objective mask design optimization considering process variations.

The second paper presents a differentiable edge-based OPC framework that combines the manufacturability of EBOPC with the performance of ILT.

The third paper proposes a co-optimization framework to reduce the number of design rule violations in the legalization and filler insertion stages.

The last paper proposes a robust ASIC placer that can handle hybrid region constraints via a unified multi-electrostatic formulation.

14:30
1098: Enabling Robust Inverse Lithography with Rigorous Multi-Objective Optimization
Yang Luo (The Hong Kong University of Science and Technology (GuangZhou)); Xiaoxiao Liang (The Hong Kong University of Science and Technology (Guangzhou)); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou))

14:45
1107: Differentiable Edge-based OPC
Guojin Chen (The Chinese University of HongKong); Haoyu Yang (NVIDIA Corp.); Haoxing Ren (NVIDIA Corporation); Bei Yu (The Chinese University of Hong Kong); David Z. Pan (University of Texas at Austin)

15:00
851: A Co-optimization Framework with Multi-layer Constraints for Manufacturability
Guohao Chen (Fudan University); Chang Liu (Fudan University); Xingyu Tong (Fudan University); Peng Zou (Shanghai LEDA Technology Co., Ltd); Jianli Chen (Fudan University)

15:15
986: MORPH: More Robust ASIC Placement for Hybrid Region Constraint Management
Jing Mai (Peking University); Zuodong Zhang (School of Integrated Circuits, Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:30 - 15:30
When Diverse Architectures Meet Diverse Ais
Room: Salons F-H

This session explores the essential strategies for optimizing deployment on various backends, including CIM architectures, MCUs, and edge GPUs, catering to a wide array of applications such as RAG, MoE, BERT, and CNNs. The importance of this topic lies in its potential to enhance performance, efficiency, and scalability in diverse computing environments. Attendees will benefit from key papers presenting cutting-edge research and practical solutions for deploying these advanced models effectively. Join us to gain insights into the latest techniques and innovations driving the optimization of deployment across multiple hardware platforms.

14:30
662: Robust Implementation of Retrieval-Augmented Generation on Edge-based Computing-in-Memory Architectures
Ruiyang Qin (University of Notre Dame); Zheyu Yan (University of Notre Dame); Dewen Zeng (University of Notre Dame); Zhenge Jia (Shandong University); Dancheng Liu (SUNY Buffalo); Jianbo Liu (University of Notre Dame); Ahmed Abbasi (University of Notre Dame); Zhi Zheng (University of Notre Dame); Ningyuan Cao (University of Notre Dame); Kai Ni (University of Notre Dame); Jinjun Xiong (University at Buffalo); Yiyu Shi (University of Notre Dame)

14:45
977: AdapMoE: Adaptive Sensitivity-based Expert Gating and Management for Efficient MoE Inference
Shuzhang Zhong (Peking University); LING LIANG (Peking University); Yuan Wang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University)

15:00
995: MCUBERT: Memory-Efficient BERT Inference on Commodity Microcontrollers
Zebin Yang (Peking University); Renze Chen (Peking University); Taiqiang Wu (The University of Hong Kong); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University); Ngai Wong (The University of Hong Kong); Yun (Eric) Liang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

15:15
1467: TSB: Tiny Shared Block for Efficient DNN Deployment on NVCIM Accelerators
Yifan Qin (University of Notre Dame); Zheyu Yan (University of Notre Dame); Zixuan Pan (University of Notre Dame); Wujie Wen (North Carolina State University); X. Sharon Hu (University of Notre Dame); Yiyu Shi (University of Notre dame)

15:30 – 16:00
Coffee Break | Exhibits
Room: Salons D-H

16:00 – 18:00
Tutorial: Heterogeneous Integration: From physical layer to architecture and packaging
Room: Salons 1-3

16:00 – 18:00
Special Session: Towards Democratized and Reproducible AI for EDA Research: Open Datasets and Benchmarks in Various Aspects
Room: Lincoln/Holland/Columbia

16:00 – 18:00
Special Session: Exploring Quantum Technologies in Practical Applications
Room: Salons A-C

16:00 - 17:00
Timing Prediction and Acceleration
Room: Skylands/Gateway

This session discusses the acceleration of timing analysis and the enhancement of timing prediction accuracy. The first paper in this session presents a CPU-GPU heterogeneous STA engine that handles generalized timing exceptions. The second paper presents a learning-based cross-corner timing signoff framework requiring only one RC corner. The last two papers address delay prediction and timing prediction problems leveraging deep learning and network models.

16:00
526: HeteroExcept: A CPU-GPU Heterogeneous Algorithm to Accelerate Exception-aware Static Timing Analysis
Zizheng Guo (Peking University); Zuodong Zhang (School of Integrated Circuits, Peking University); Wuxi Li (AMD); Tsung-Wei Huang (University of Wisconsin at Madison); Xizhe Shi (Peking University); Yufan Du (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

16:15
557: One-for-All: An Unified Learning-based Framework for Efficient Cross-Corner Timing Signoff
Linyu Zhu (Shanghai Jiao Tong University); Yichen Cai (Shanghai Jiao Tong University); Xinfei Guo (Shanghai Jiao Tong University)

16:30
630: CircuitSeer: RTL Post-PnR Delay Prediction via Coupling Functional and Structural Representation
Sanjay Gandham (University of Central Florida); Joe Walston (Synopsys); Sourav Samanta (Synopsys); Lingxiang Yin (University of Central Florida); Hao Zheng (University of Central Florida); Mingjie Lin (University of Central Florida); Stelios Diamantidis (Synopsys)

16:45
1590: Explainable and Layout-Aware Timing Prediction
Zhengyang Lyu (University of Science and Technology of China; Institute of Computing Technology, CAS); Xiaqing Li (Institute of Computing Technology, CAS); Zidong Du (Institute of Computing Technology, CAS); Qi Guo (Institute of Computing Technology, CAS); Huaping Chen (University of Science and Technology of China); Yunji Chen (Institute of Computing Technology, CAS)

16:00 - 17:00
How Much ML Can You Squeeze into Your Edge Device?
Room: Essex/Liberty

Applying ML in edge devices has many important applications such as autonomous driving and medical diagnosis. This track includes four presentations considering how to efficiently and effectively apply ML in the edge. The first paper considers cooperative inferences on multiple edge devices. The second paper presents efficient HDR generation with the help of the user's visual attention. The third paper applies deep ensemble in edge devices with constrained resources. The last paper proposes a hardware-friendly softmax for power saving when applying ML in edge devices.

16:00
1551: RACI: A Resource-Aware Cooperative Inference Framework on Heterogeneous Edge Devices
zhenyu wang (Chongqing University); Ao Ren (Chongqing University); Duo Liu (Chongqing University); Haining Fang (Chongqing University); Jiaxing Shi (Chongqing University); Yujuan Tan (Chongqing University); Xianzhang Chen (Chongqing University)

16:15
927: Foveated HDR: Efficient HDR Content Generation on Edge Devices Leveraging User's Visual Attention
Ziyu Ying (Penn State); Sandeepa Bhuyan (The Pennsylvania State University); Yingtian Zhang (The Pennsylvania State University); Yan Kang (The Pennsylvania State University); Mahmut Taylan Kandemir (Penn State); Chita R. Das (Penn State University)

16:30
1337: Tiny Deep Ensemble: Uncertainty Estimation in Edge AI Accelerators via Ensembling Normalization Layers with Shared Weights
Soyed Tuhin Ahmed (KIT - Karlsruhe Institute of Technology, Karlsruhe, Germany); Mehdi Tahoori (Karlsruhe Institute of Technology)

16:45
1117: ConSmax: Hardware-Friendly Alternative Softmax with Learnable Parameters
Shiwei Liu (Google Research); Guanchen Tao (Department of Electrical Engineering and Computer Sciences, University of Michigan); Yifei Zou (Department of Electrical Engineering and Computer Sciences, University of Michigan); Derek Chow (Google Research); Zichen Fan (Department of Electrical Engineering and Computer Sciences, University of Michigan); Kauna Lei (Department of Electrical Engineering and Computer Sciences, University of Michigan); Bangfei Pan (Google Research); Dennis Sylvester (Department of Electrical Engineering and Computer Sciences, University of Michigan); Gregory Kielian (Google Research); Mehdi Saligane (Department of Electrical Engineering and Computer Sciences, University of Michigan)

16:00 - 17:00
Reliable emerging technologies
Room: Salons F-H

This session presents emerging technologies from 3D integration to wavelength routed optical network on chip (NOC) and NAND Flash, and the reliability issues associated with them.

16:00
1080: Efficient Ultra-Dense 3D IC Power Delivery and Cooling Using 3D Thermal Scaffolding
Dennis Rich (Stanford University); Tathagata Srimani (Stanford University); Mohamadali Malakoutian (Stanford University); Srabanti Chowdhury (Stanford University); Subhasish Mitra (Stanford University)

16:15
1512: Three Guides for Efficient Automatic Post-Fabrication Optimization of Modern NAND Flash Memory
Earl Kim (Samsung Electronics); Hyunuk Cho (POSTECH); Sungjun Cho (POSTECH); Myungsuk Kim (Kyungpook National University); Jisung Park (POSTECH (Pohang University of Science and Technology)); Jaeyong Jeong (Samsung Electronics); Eunkyoung Kim (Samsung Electronics); Sunghoi Hur (Samsung Electronics)

16:30
1323: Minimizing Worst-Case Data Transmission Cycles in Wavelength-Routed Optical NoC through Bandwidth Allocation
Liaoyuan Cheng (Technical University of Munich); Mengchu Li (Technical University of Munich); Tsun-Ming Tseng (Technical University of Munich); Ulf Schlichtmann (Technical University of Munich)

16:45
847: REMNA: Variation-Resilient and Energy-Efficient MLC FeFET Computing-in-Memory Using NAND Flash-Like Read and Adaptive Control
Taixin Li (Tsinghua University); Hongtao Zhong (Tsinghua University); Yixin Xu (The Pennsylvania State University); Vijaykrishnan Narayanan (Penn State University); Kai Ni (University of Notre Dame); Huazhong Yang (Tsinghua University); Thomas Kämpfe (Fraunhofer IPMS); Xueqing Li (Tsinghua University)

17:00 - 18:00
Innovative Approaches in Circuit Simulation: High-Fidelity Modeling, Optimization, and Parallelization
Room: Skylands/Gateway

This session highlights groundbreaking advancements in circuit simulation. The first paper introduces a high-fidelity 2D warpage model for advanced packaging, achieving significant speedups and accuracy. The second paper presents an optimization technique that accelerates SPICE simulations using neural ODEs and graph convolution networks. The third paper extends spectral sparsification to nonlinear circuits, enhancing solver performance. The final paper offers a parallel-in-time exponential integrator method for faster transient simulations. Together, these works represent the forefront of high-fidelity modeling, optimization, and parallelization in circuit simulation.

17:00
1006: Efficient High-Fidelity Two-Dimensional Warpage Modeling for Advanced Packaging Analysis
Shao-Yu Lo (National Taiwan University); MaoZe Liu (National Taiwan University); Yao-Wen Chang (National Taiwan University)

17:15
1245: Pseudo Adjoint Optimization: Harnessing the Solution Curve for SPICE Acceleration
Jiatai Sun (China University of Petroleum, Beijing); Xiaru Zha (China University of Petroleum, Beijing); Chao Wang (Southesat University); Xiao Wu (Huada Empyrean Software Co. Ltd); Dan Niu (Southeast University); Wei W. Xing (The University of Sheffield); Zhou Jin (Super Scientific Software Laboratory, China University of Petroleum-Beijing)

17:30
1266: CSP: Comprehensive Sparsification Preconditioning for Nonlinear Circuit Simulation
Yuxuan Zhao (Super Scientific Software Laboratory, China University of Petroleum,Beijing); Xiaoyu Yang (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Yinuo Bai (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Lijie Zeng (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Dan Niu (Southeast University); Weifeng Liu (China University of Petroleum-Beijing); Zhou Jin (Super Scientific Software Laboratory, China University of Petroleum-Beijing)

17:45
1236: EI-PIT: A Parallel-in-Time Exponential Integrator Method for Transient Linear Circuit Simulation
Hang Zhou (School of Microelectronics, Southern University of Science and Technology); Quan Chen (School of Microelectronics, Southern University of Science and Technology)

17:00 - 18:00
Analog, Analog, and More Analog Design using Your Favorite AI Algorithms
Room: Essex/Liberty

Applying AI in analog design has great potential for efficiency and effectiveness. This track includes four papers in this domain. The first paper combines Bayesian optimization with a large language model to utilize domain-specific knowledge for analog circuit design. The second paper applies an invertible graph generative model in the design of operational amplifiers. The third paper introduces a new physics-inspired NN model that can be implemented in low-power analog circuits. The last paper proposes new physics-informed neural networks for TSV electromigration analysis.

17:00
1432: ADO-LLM: Analog Design Bayesian Optimization with In-Context Learning of Large Language Models
Yuxuan Yin (University of California, Santa Barbara); Yu Wang (University of California, Santa Barbara); Boxun Xu (University of California, Santa Barbara); Peng Li (University of California, Santa Barbara)

17:15
753: TSO-Flow: A Topology Synthesis and Optimization Workflow for Operational Amplifiers with Invertible Graph Generative Model
Jinglin Han (Beihang University); Yuhao Leng (Beihang University); Xiuli Zhang (Beihang University); Peng Wang (Beihang University)

17:30
598: KirchhoffNet: A Scalable Ultra Fast Analog Neural Network
Zhengqi Gao (Dept. of EECS, MIT); Fan-Keng Sun (EECS, MIT); Ron Rohrer (CMU); Duane Boning (MIT)

17:45
942: Enforcing hard constraints in physics-informed learning for transient TSV electromigration analysis
Xiaoman Yang (Shanghai Jiao Tong University); Hai-Bao Chen (Department of Micro/Nano-electronics, Shanghai Jiao Tong University); Wenjie Zhu (Shanghai Jiao Tong University); Yuhan Zhang (University of Michigan - Ann Arbor); Yongkang Xue (Shanghai Jiao Tong University); Pengpeng Ren (Shanghai Jiao Tong University); Runsheng Wang (Peking University); Zhigang Ji (Shanghai Jiaotong University); Ru Huang (Peking University)

17:00 - 18:00
Emerging Technologies enabling Content Addressable Memories
Room: Salons F-H

This session delves into the efficient design of Content Addressable Memories (CAM) using a computation in memory style using emerging technologies.

17:00
1263: TReCiM: Lower Power and Temperature-Resilient Multibit 2FeFET-1T Compute-in-Memory Design
Yifei Zhou (Zhejiang University); Thomas Kämpfe (Fraunhofer IPMS); Kai Ni (University of Notre Dame); Hussam Amrouch (Technical University of Munich (TUM)); Cheng Zhuo (Zhejiang University); Xunzhao Yin (Zhejiang University)

17:15
763: CAMSHAP: Accelerating Machine Learning Model Explainability with Analog CAM
John Moon (Hewlett Packard Labs); Giacomo Pedretti (Hewlett Packard Enterprise); Pedro Bruel (Hewlett Packard Labs); Sergey Serebryakov (Hewlett Packard Labs); Omar Eldash (Hewlett Packard Labs); Luca Buonanno (Hewlett Packard Labs); Catherine E. Graves (Hewlett Packard Labs); Paolo Faraboschi (Hewlett Packard Labs); Jim Ignowski (Hewlett Packard Labs)

17:30
1314: ShiftCAM: A Time-Domain Content Addressable Memory Utilizing Shifted Hamming Distance for Robust Genome Analysis
Peiyi He (The University of Hong Kong); Ruibin Mao (The University of Hong Kong); Keyi Shan (The University of Hong Kong); Yunwei Tong (The University of Hong Kong); Zhicheng XU (The University of Hong Kong); Muyuan Peng (The University of Hong Kong); Ruibang Luo (The University of Hong Kong); Can Li (The University of Hong Kong)

17:45
782: TAP-CAM: A Tunable Approximate Matching Engine based on Ferroelectric Content Addressable Memory
Chenyu Ni (Zhejiang University); Sijie Chen (Zhejiang University); Liu Liu (University of Notre Dame); Mohsen Imani (University of California, Irvine); Thomas Kämpfe (Fraunhofer IPMS); Kai Ni (University of Notre Dame); Michael Niemier (University of Notre Dame); Xiaobo Sharon Hu (University of Notre Dame); Cheng Zhuo (Zhejiang University); Xunzhao Yin (Zhejiang University)

18:00 – 19:30
Welcome Reception

Technical Program: 29 October 2024

7:30 – 8:30
Registration
Room: Grand Ballroom Foyer

8:30 – 9:30
Keynote: The Future of Chip Design
Leon Stokm IBM
Room: Salons 4-8

9:30 – 10:00
Coffee Break | Exhibits
Room: Salons D-H

10:00 - 10:45
Processor, Memory, and Storage Designs
Room: Salons 1-3

This session contains three papers on system-level design and exploration, a classic problem that always needs new solutions for emerging devices and applications. The first one investigates the sustainability-performance trade-off of the dynamic instruction selection logic in superscalar processors. The second paper presents a tunable framework for generating memory-centric workloads that can be used to evaluate and validate memory subsystem designs. Finally, the third paper presents an approach for coupling heterogeneous management of modern high-density SSDs into the zone management at the host.

10:00
1058: Sustainable High-Performance Instruction Selection for Superscalar Processors
Saeideh Sheikhpour (Ghent University); David Christoph Metz (Norwegian University of Science and Technology); Erling Jullum (Norwegian University of Science and Technology (NTNU)); Magnus Själander (Norwegian University of Science and Technology); Lieven Eeckhout (Ghent University)

10:15
1350: A Framework for Explainable, Comprehensive, and Customizable Memory-Centric Workloads
Mohamed Abuelala (McMaster University); Mohamed Hassan (McMaster University)

10:30
529: ZnH2: Augmenting ZNS-based Storage System with Host-managed Heterogeneous Zones
Yingjia Wang (The Chinese University of Hong Kong); Lok Yin Chow (The Chinese University of Hong Kong); Xirui Nie (The Chinese University of Hong Kong); Yuhong Liang (The Chinese University of Hong Kong); Ming-Chang Yang (The Chinese University of Hong Kong)

10:00 - 10:45
Innovating Data Storage: Exploring Adaptive Indexing, Access Pattern Optimization, and Memory Longevity Enhancement for SSDs
Room: Lincoln/Holland/Columbia

10:00
679: ALISA: An Adaptive Learned Index Structure for Spatial Data on Solid-State Drives
Che-Wei Lin (National Yang Ming Chiao Tung University); Chun-Feng Wu (National Yang Ming Chiao Tung University)

10:15
1114: An Access Pattern-aware Hybrid Learning-based and Conventional Mapping for Solid-State Drives
Qian Wei (Shandong University); Xiaosu Guo (University of Texas at Dallas); Jie Wang (Shandong University); Zhaoyan Shen (Shandong University); Dongxiao Yu (Shandong University); Zhiping Jia (Shandong University); Bingzhe Li (University of Texas at Dallas)

10:30
799: CellRejuvo: Rescuing the Aging of 3D NAND Flash Cells with Dense-Sparse Cell Reprogramming
Han-Yu Liao (National Taiwan University of Science and Technology); Yi-Shen Chen (Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology); Jen-Wei Hsieh (National Taiwan University of Science and Technology); Yuan-Hao Chang (Academia Sinica); Hung-Pin Chen (Innodisk Corporation)

10:00 - 10:45
Enhancing Simulation Efficiency through Multi-Core/GPU-Acceleration and Instruction-level Fault Injection
Room: Salons A-C

This session focuses on simulation efficiency. The first paper introduces a powerful fault simulator for multi-core systems that first parallelizes simulation in the fault dimension and subsequently in the pattern dimension. The second paper introduces a GPU-accelerated logic simulator that uses two-phase simulation to enhance parallelism, achieving two to five times speed-ups over the state-of-the-art. The third paper explores innovative techniques for accurately and efficiently simulating processor faults by dynamically shifting the simulation from a register transfer level to a higher instruction level.

10:00
795: DDP-Fsim: Efficient and Scalable Fault Simulation for Deterministic Patterns with Two-Dimensional Parallelism
Feng Gu (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing); Mingjun Wang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing); Jianan Mu (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing); Zizhen Liu (Institute of Computing Technology,Chinese Academy of Sciences); Jiaping Tang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing); Hui Wang (CASTEST, Beijing); Yonghao Wang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing); Jing Ye (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing); Huawei Li (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing); Xiaowei Li (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; CASTEST, Beijing)

10:15
668: GL0AM: GPU Logic Simulation Using 0-Delay and Re-simulation Acceleration Method
Yanqing Zhang (NVIDIA); Haoxing Ren (NVIDIA Corporation); Brucek Khailany (NVIDIA)

10:30
1075: Accelerating Fault Injection for Validating Processor RTL Implementations
Yi Yuan (The University of Texas at Austin); Derek Chiou (Microsoft/UT Austin)

10:00 - 10:45
Let LLMs Generate Your RTL Code!
Room: Skylands/Gateway

This session showcases papers that aim to enhance the productivity and efficiency of the design automation flows leveraging LLMs. The first two papers address code generation and the final paper addresses debugging. The first paper in the lineup utilizes techniques such as multi-modal program analysis, a search engine, and a cost-aware search algorithm to optimize RTL in terms of area-delay while incurring lower synthesis and verification time. The second paper focuses on establishing a dataset that can most effectively support code generation. OriGen, an open-source system with self-reflection and dataset augmentation is presented to improve the quality of open-source RTL datasets. The final paper of the session introduces the Make Each Iteration Count (MEIC) framework. Syntax and functional errors are detected and mitigated within this framework via the underlying LLM structure.

10:00
1188: RTLRewriter: Methodologies for Large Models aided RTL Code Optimization
Xufeng Yao (Chinese University of HongKong); Yiwen Wang (Huawei); Xing Li (Huawei); Yingzhao Lian (Huawei); Ran Chen (Huawei); Lei Chen (Huawei); Mingxuan Yuan (Huawei Noah's Ark Lab); Hong Xu (CUHK); Bei Yu (The Chinese University of Hong Kong)

10:15
1539: OriGen: Enhancing RTL Code Generation with Code-to-Code Augmentation and Self-Reflection
Fan Cui (Peking University); Chenyang Yin (Peking University); Kexing Zhou (Peking University); Youwei Xiao (School of Integrated Circuits, Peking University); Guangyu Sun (Peking University); Qiang Xu (The Chinese University of Hong Kong); Qipeng Guo (Shanghai Artificial Intelligence Laboratory); Demin Song (Shanghai Artificial Intelligence Laboratory); Dahua Lin (Shanghai Artificial Intelligence Laboratory); Xingcheng Zhang (Shanghai Artificial Intelligence Laboratory); Yun (Eric) Liang (Peking University)

10:30
1315: MEIC: Re-thinking RTL Debug Automation using LLMs
Ke Xu (Southeast University); Jialin Sun (Southeast University); Yuchen HU (Southeast University); Xinwei Fang (University of York); Weiwei Shan (Southeast University); Xi Wang (Tsinghua University); Zhe Jiang (South East University)

10:00 - 10:45
Architectural Mapping
Room: Essex/Liberty

10:00
962: DISC: Exploiting Data Parallelism of Non-Stencil Computations on CGRAs via Dynamic Iteration Scheduling
Yue Liang (Chongqing University); Di Mou (Chongqing University); Dajiang Liu (Chongqing University)

10:15
1220: MatFactory: A Framework for High-performance Matrix Factorization on FPGAs
Mingzhe Zhang (Tsinghua University); Xiaochen Hao (Peking University); Hongbo Rong (Intel Parallel Computing Lab); Wenguang Chen (Tsinghua University)

10:30
687: EasyPart: An Effective and Comprehensive Hypergraph Partitioner for FPGA-based Emulation
Shengbo Tong (Tsinghua University); Haoyuan Li (Tsinghua University); Jiahao Xu (Tsinghua University); Chunyan Pei (Tsinghua University); Wenjian Yu (Tsinghua University); Shengjun Liu (HyperSilicon); Jian Shen (HyperSilicon)

10:00 - 10:45
IR Drop and High-speed Link Analysis
Room: Salons F-H

This session discusses modeling and estimation methods for addressing waveform distortion in high-speed links and dynamic/static IR drops. The first paper proposes a semi-vector-based assessment flow aimed at providing a more accurate estimation of worst-case peak power and IR drop. The second paper presents a CNN-based approach, along with comprehensive feature extraction, to speed up static IR drop estimation. The session concludes with LiTformer, a more accurate transformer-based model for high-speed link transmitters.

10:00
1121: Peak Power and Dynamic IR-drop Assessment via Waveform Augmenting
Yihan Wen (Beijing University of Technology); Juan Li (Beijing University of Technology); Bei Yu (The Chinese University of Hong Kong); Xiaoyi Wang (Unaffiliated Scholar)

10:15
1034: CFIRSTNET: Comprehensive Features for Static IR Drop Estimation with Neural Network
Yu-Tung Liu (National Yang Ming Chiao Tung University); Yu-Hao Cheng (National Yang Ming Chiao Tung University); Shao-Yu Wu (National Yang Ming Chiao Tung University); Hung-Ming Chen (National Yang Ming Chiao Tung University)

10:30
1212: LiTformer: Efficient Modeling and Analysis of High-Speed Link Transmitters Using Non-Autoregressive Transformer
Songyu Sun (Zhejiang University); Xiao Dong (Zhejiang University); YanLiang Sha (School of Microelectronics, Southern University of Science and Technology); Quan Chen (School of Microelectronics, Southern University of Science and Technology); Cheng Zhuo (Zhejiang University)

10:45 - 11:30
Efficient Machine Learning: from Cloud to Edge
Room: Salons 1-3

As machine learning continues to advance, its deployment in different computation platform becomes increasingly complex. This session combines three interesting papers on this topic. The first one proposes an automated optimization framework to support the concurrent execution of multiple neural network models on GPUs. The second work manages multiple DNN workloads on heterogenous embedded devices, aiming to co-optimize the throughput and power efficiency. The third paper aims to achieve energy-efficient Federated Learning (FL), considering the energy constraints in both heterogeneous IoT devices and heterogeneous deep learning models.

10:45
856: GACER: Granularity-Aware ConcurrEncy Regulation for Multi-Tenant Deep Learning
Yongbo Yu (George Mason University); Fuxun Yu (Microsoft); Zhi Tian (George Mason University); Xiang Chen (George Mason University)

11:00
893: MapFormer: Attention-based multi-DNN manager for throughout & power co-optimization on embedded devices
Andreas Karatzas (Southern Illinois University Carbondale); Iraklis Anagnostopoulos (Southern Illinois University Carbondale)

11:15
1428: Towards Energy-Aware Federated Learning via MARL: A Dual-Selection Approach for Model and Client
Jun Xia (University of Notre Dame); Yi Zhang (east china normal university); Yiyu Shi (University of Notre dame)

10:45 - 11:30
EdgeML: Efficient and Private ML for the Edge
Room: Lincoln/Holland/Columbia

10:45
755: AdaPI: Facilitating DNN Model Adaptivity for Efficient Private Inference in Edge Computing
Tong Zhou (Northeastern University); Jiahui Zhao (University of Connecticut); Yukui Luo (University of Massachusetts Dartmouth); Xi Xie (University of Connecticut); Wujie Wen (North Carolina State University); Caiwen Ding (University of Connecticut); Xiaolin Xu (Northeastern University)

11:00
1218: EPipe: Pipeline Inference Framework with High-quality Offline Parallelism Planning for Heterogeneous Edge Devices
Yi Xiong (university of science and technology of China); Weihong Liu (University of Science and Technology of China); Rui Zhang (School of Software Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China); Yulong Zu (School of Software Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China); Zhu Zongwei (University of Science and Technology of China (USTC)); Xuehai Zhou (University of Science and Technology of China)

11:15
658: Residual-INR: Communication Efficient On-Device Learning Using Implicit Neural Representation
Hanqiu Chen (Georgia Institute of Technology); Xuebin Yao (Samsung Semiconductor, Inc); Pradeep Subedi (Samsung Semiconductor, Inc); Cong "Callie" Hao (Georgia Institute of Technology)

10:45 - 11:30
Advances in Verification through SAT Solving and Machine Learning
Room: Salons A-C

This session presents several advancements in automated verification. The first paper introduces a new framework for checking the equivalence of flow-based computing circuits for in-memory processing, using helper variables to transform undirected graphs into directed ones for more effective satisfiability (SAT)-based verification. The second paper presents a novel approach for SAT solving by transforming it into a minimization problem and iteratively refining variable assignments using gradient descent. The third paper focuses on improving formal verification by integrating reinforcement learning to automatically generate conjectures from simulation traces.

10:45
880: Equivalence Checking for Flow-Based Computing using Iterative SAT Solving
Sven Thijssen (University of Central Florida); Muhammad Rashedul Haq Rashed (University of Central Florida); Md Rubel Ahmed (University of Central Florida); Suraj S. Singireddy (University of Texas at San Antonio); Sumit K. Jha (Florida International University); Rickard Ewetz (University of Central Florida)

11:00
996: DiffSAT: Differential MaxSAT Layer for SAT Solving
Yu Zhang (The Chinese University of Hong Kong); Hui-Ling Zhen (Huawei); Mingxuan Yuan (Huawei Noah's Ark Lab); Bei Yu (The Chinese University of Hong Kong)

11:15
715: Word-Level Augmentation of Formal Proof by Learning from Simulation Traces
Zhiyuan Yan (The Hong Kong Univeristy of Science and Technology(Guangzhou)); Hongce Zhang (The Hong Kong Univeristy of Science and Technology(Guangzhou))

10:45 - 11:30
New Benchmarks and Understanding Benchmarks using LLMs
Room: Skylands/Gateway

This session introduces works in benchmark and dataset generation with AI technologies. The first paper in the session aims to improve the documentation process with a customized retrieval augmented generation and benchmarking tool to improve Q&A functionality, The second paper presents a benchmarking tool to assess Verilog generation A multi-modal generative model for code generation leverages both image and language to generate Verilog code benchmarks. The final paper in this session introduces benchmarking datasets for ML-driven floorplanning. These datasets reflect the complexities and hard constraints of SoCs effectively.

10:45
931: Customized Retrieval Augmented Generation and Benchmarking for EDA Tool Documentation QA
Yuan Pu (The Chinese University of Hong Kong); Zhuolun He (The Chinese University of Hong Kong); Tairu Qiu (ChatEDA Tech); Haoyuan WU (Shanghai AI Lab); Bei Yu (The Chinese University of Hong Kong)

11:00
685: Natural language is not enough: Benchmarking multi-modal generative AI for Verilog generation
Kaiyan Chang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing; University of Chinese Academy of Sciences); Zhirong Chen (Zhejiang University); Yunhao Zhou (Shanghai Jiao Tong University); Wenlong Zhu (Institute of Computing Technology,Chinese Academy of Sciences); kun wang (UCAS); Haobo Xu (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences); Cangyuan Li (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences); Mengdi Wang (Institute of Computing Technology, Chinese Academy of Sciences); Shengwen Liang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing; University of Chinese Academy of Sciences); Huawei Li (Institute of Computing Technology, Chinese Academy of Sciences); yinhe han (Institute of Computing Technology,Chinese Academy of Sciences); Ying Wang (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences)

11:15
1410: FloorSet - a VLSI Floorplanning Dataset with Design Constraints of Real-World SOCs.
Uday Mallappa (Intel Labs); Hesham Mostafa (Intel Labs); Mikhail Galkin (Intel Labs); Mariano J. Phielipp (Intel Labs); Somdeb Majumdar (Intel Labs)

10:45 - 11:30
Applications and Architectures
Room: Essex/Liberty

10:45
1357: DoS-FPGA: Denial of Service on Cloud FPGAs via Coordinated Power Hammering
Hassan Nassar (Karlsruher Institut für Technologie); Philipp Machauer (KIT); Lars Bauer (Karlsruhe Institute of Technology); Dennis R. E. Gnad (Karlsruhe Institute of Technology); Mehdi Tahoori (Karlsruhe Institute of Technology); Joerg Henkel (KIT)

11:00
693: HG-PIPE: Vision Transformer Acceleration with Hybrid-Grained Pipeline
Qingyu Guo (School of Integrated Circuits, Peking University); Jiayong Wan (School of Integrated Circuits, Peking University); Songqiang Xu (School of Integrated Circuits, Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University); Yuan Wang (Peking University)

11:15
1196: Sustainable Hardware Specialization
Pranav Dangi (National University of Singapore); Thilini Kaushalya Bandara (National University of Singapore); Saeideh Sheikhpour (Ghent University); Tulika Mitra (National University of Singapore); Lieven Eeckhout (Ghent University)

10:45 - 11:30
Machine Learning-based Design and Timing Optimization
Room: Salons F-H

This session discusses design and timing optimization powered by machine leaning techniques. The first paper in this session proposes RankTuner, which is a ranking-based tool parameter tuning framework learning the dominant relationship between parameters. The second paper presents a new gate sizer based on LeakGAN which incorporates GAN with reinforcement learning. The last one addresses the challenges in timing-driven placement during the physical design flow of VLSI circuits with GNN.

10:45
1223: RankTuner: When Design Tool Parameter Tuning Meets Preference Bayesian Optimization
Peng Xu (The Chinese University of Hong Kong); Su Zheng (The Chinese University of Hong Kong); Yuyang Ye (Southeast University); Chen BAI (The Chinese University of Hong Kong); Siyuan Xu (Huawei Noah's Ark Lab); Hao Geng (ShanghaiTech University); Tsung-Yi Ho (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong)

11:00
1296: LAG-Sizer: A Novel Gate Sizer Based on Leak Generative Adversarial Network with Feature Fusion
Zhanhua Zhang (Southeast University); Wenjie Ding (Southeast university); Guoqing He (Southeast University); Peng Cao (Southeast University)

11:15
1166: A Physical and Timing Aware Placement Optimization Framework Based on Graph Neural Network
Wenjie Ding (Southeast university); Zhanhua Zhang (Southeast University); Guoqing He (Southeast University); Peng Cao (Southeast University)

11:30 – 13:00
CEDA Awards Luncheon & Keynote
Room: Salons 4-8

13:00 – 14:30
Student Research Competition
Room: Salons 1-3

13:00 – 14:30
Special Session: Co-Designing NVM-based Systems for Machine Learning Applications
Room: Lincoln/Holland/Columbia

13:00 – 14:30
Special Session: Delocalizing AI with Emerging Edge Intelligence (IoT/Internet)
Room: Salons A-C

13:00 - 13:45
Let AI Power Your Synthesis and Defect Analysis!
Room: Skylands/Gateway

The advancement in AI is now boosting the EDA design flow. This track has three presentations that apply AI in chip placement, in HLS DSE, and in defect detection respectively.

13:00
632: The Power of Graph Signal Processing for Chip Placement Acceleration
Yiting Liu (Fudan University); Hai Zhou (Northwestern University); Jia Wang (Illinois Institute of Technology); Fan Yang (Fudan University); Xuan Zeng (Fudan University); Li Shang (fudan university)

13:15
885: Efficient Task Transfer for HLS DSE
Zijian Ding (University of California, Los Angeles); Atefeh Sohrabizadeh (University of California Los Angeles); Weikai Li (University of California, Los Angeles); Zongyue Qin (UCLA); Yizhou Sun (University of California, Los Angeles); Jason Cong (UCLA)

13:30
1004: SEM-CLIP: Precise Few-Shot Learning for Nanoscale Defect Detection in Scanning Electron Microscope Image
Qian Jin (Zhejiang University); Yuqi Jiang (Zhejiang University); Xudong Lu (Zhejiang University); Yumeng Liu (Zhejiang University); yining chen (Zhejiang University); Dawei Gao (Zhejiang University); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)

13:00 - 13:45
Revolutionizing AI with Low Power Accelerators: Emerging Design Trends
Room: Essex/Liberty

This session includes three papers focusing on different aspects of low-power approximate and stochastic computing designs. The first paper introduces a novel mixed-signal Compute-in-Entropy hardware primitive aimed at addressing the substantial resource demands of Bayesian Neural Networks hardware implementations. The second paper presents an innovative stochastic computing-based method for neural network acceleration, which resolves critical issues found in existing accelerators. The third paper offers a comprehensive exploration of printed neural network accelerators, beginning with the analog-to-digital interface—an important area and major power sink in sensor processing applications—and extending to networks of ternary neurons and their implementation.

13:00
1349: Towards Uncertainty-Quantifiable Biomedical Intelligence: Mixed-signal Compute-in-Entropy for Bayesian Neural Networks
Likai Pei (University of Notre Dame); Yifan Qin (University of Notre Dame); Zephan M. Enciso (University of Notre Dame); Boyang Cheng (University of Notre Dame); Jianbo Liu (University of Notre Dame); Steven Davis (University of Notre Dame); Zhenge Jia (Shandong University); Michael Niemier (University of Notre Dame); Yiyu Shi (University of Notre Dame); X. Sharon Hu (University of Notre Dame); Ningyuan Cao (University of Notre Dame)

13:15
527: OSCA: End-to-end Serial Stochastic Computing Neural Acceleration with Fine-grained Scaling and Piecewise Activation
Yixuan Hu (Peking University); Yikang Jia (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University); Yuan Wang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

13:30
923: Evolutionary Approximation of Ternary Neurons for On-sensor Printed Neural Networks
Vojtech Mrazek (Brno University of Technology); Argyris Kokkinis (Aristotle University of Thessaloniki); Panagiotis Papanikolaou (University of Michigan); Zdenek Vasicek (Brno University of Technology); Kostas Siozios (Department of Physics, Aristotle University of Thessaloniki); Georgios Tzimpragos (University of Michigan); Mehdi Tahoori (Karlsruhe Institute of Technology); Georgios Zervakis (University of Patras)

13:00 - 13:45
New Techniques in Analog Optimization: Bayesian Sensitivity, Hierarchical Placement, and AI-Driven 2.5D Chiplet Design
Room: Salons F-H

This session highlights new advancements in analog optimization methods. The first paper is focused on optimization of analog sizing using Bayesian optimization combined with very accurate adjoint sensitivity analysis. The second paper presents an new approach for optimization of joint placement which exploits hierarchical information for analog/mixed-sigmal designs. The final paper uses AI to evaluate and optimize 2.5D chiplet bump pitch effects. These papers each present novel approaches to the optimized design of analog circuits.

13:00
966: Revisiting sensitivity-based analog sizing with derivative-aware Bayesian optimization and error-suppressed adjoint analysis
Ruiyu Lyu (School of Microelectronics, State Key Laboratory of Integrated Chips & System, Fudan University); Aidong Zhao (Fudan University); Yuan Meng (School of Microelectronics, State Key Laboratory of Integrated Chips & System, Fudan University); Zhaori Bi (Fudan University); Keren Zhu (The Chinese University of Hong Kong); Changhao Yan (Fudan University); Fan Yang (Fudan University); Dian Zhou (Fudan University); Xuan Zeng (Fudan University)

13:15
602: Joint Placement Optimization for Hierarchical Analog/Mixed-Signal Circuits
Xiaohan Gao (Peking University); Haoyi Zhang (Peking University); Bingyang Liu (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

13:30
1436: AI-Driven Evaluation and Optimization of Bump Pitch Effects on Chiplet and Interposer Design Quality
Seungmin Woo (Georgia Institute of Technology); Pruek Vanna-iampikul (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech)

13:45 - 14:30
Dive into the Deisgn Space for Design Automation
Room: Skylands/Gateway

This session delves into the critical methodologies for Design Space Exploration (DSE) in ASIC and FPGA design, focusing on innovative techniques like Bayesian Optimization, Reinforcement Learning, and Design Space Mining. The discussion will highlight the significance of these approaches in achieving efficient and effective DSE, which is paramount for optimizing performance and resource utilization in semiconductor design. Key papers to be presented include groundbreaking research on these topics, showcasing their practical applications and impact on the future of hardware design. Attendees will gain valuable insights into the latest advancements and best practices in DSE.

13:45
988: Is Vanilla Bayesian Optimization Enough for High-Dimensional Architecture Design Optimization?
Yuanhang Gao (Zhejiang University); Donger Luo (Shanghaitech University); Chen BAI (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong); Hao Geng (ShanghaiTech University); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)

14:00
998: TransLib: An Extensible Graph-Aware Library Framework for Automated Generation of Transformer Operators on FPGA
Yang Liu (Fudan University); Tianchen Wang (Fudan University); Yuxuan Dong (Fudan University); Zexu Zhang (Fudan University); Shun Li (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

14:15
1084: MapTune: Advancing ASIC Technology Mapping via Reinforcement Learning Guided Library Tuning
Mingju Liu (University of Maryland College Park); Daniel Robinson (University of Utah); Yingjie Li (University of Maryland, College Park); Cunxi Yu (University of Maryland, College Park)

13:45 - 14:30
Machine Learning Innovations for Thermal and Power Optimization
Room: Essex/Liberty

As the complexity and integration density of modern 3D-ICs and chiplet designs continue to grow, efficient thermal and power management becomes crucial for maintaining performance and reliability. This session not only delves into the latest ML advancements that address these challenges but also highlights their practical benefits. The session features pioneering research that leverages ML techniques to revolutionize the thermal analysis and power management processes. These sophisticated ML models significantly accelerate transient thermal prediction for full-chip designs, achieving speedups of several orders of magnitude over traditional simulators while ensuring long-term stability and minimal prediction errors. Additionally, this session showcases hierarchical power management frameworks designed for LLM chiplet designs, which integrate scalable simulation techniques with dynamic power delivery networks to adapt power strategies in real-time. This results in substantial energy savings and enhanced power efficiency, addressing the unique demands of chiplet-based systems. Join us to explore the intersection of machine learning, thermal management, and power optimization and discover how these innovations can be practically applied in your work, shaping the future of EDA.

13:45
972: FaStTherm: Fast and Stable Full-Chip Transient Thermal Predictor Considering Nonlinear Effects
Tianxiang Zhu (Peking University); Qipan Wang (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:00
974: Hierarchical Power Co-Optimization and Management for LLM Chiplet Designs
Yanchi Dong (Peking University); Xueping Liu (Peking University); Xiaochen Hao (Peking University); Yun (Eric) Liang (Peking University); Ru Huang (Peking University); Le Ye (Peking University); Tianyu Jia (Peking University)

14:15
830: ARO: Autoregressive Operator Learning for Transferable and Multi-fidelity 3D-IC Thermal Analysis With Active Learning
Mingyue Wang (Beihang University); Yuanqing Cheng (Beihang University); Weiheng Zeng (Beihang University); Zhenjie Lu (Shenzhen University); Vasilis F. Pavlidis (Aristotle University of Thessaloniki); Wei W. Xing (The University of Sheffield)

13:45 - 14:30
Routing and ECO Routing
Room: Salons F-H

Routing takes effort. This session starts with competing efforts for Global Routing in the first two presentations, HeLEM-GR and InstantGR. Routing takes even more effort, the third paper details a routing effort at ECO for DRV mitigation.

13:45
523: HeLEM-GR: Heterogeneous Global Routing with Linearized Exponential Multiplier Method
Chunyuan Zhao (Peking University); Zizheng Guo (Peking University); Rui Wang (Southwestern University of Finance and Economics); Zaiwen Wen (Peking University); Yun (Eric) Liang (Peking University); Yibo Lin (Peking University)

14:00
1239: InstantGR: Scalable GPU Parallelization for Global Routing
Shiju Lin (The Chinese University of Hong Kong); Liang Xiao (The Chinses University of Hong Kong); Jinwei Liu (The Chinese University of Hong Kong); Evangeline Young (The Chinese University of Hong Kong)

14:15
1501: An Effective ECO Methodology for Reducing Back-side Design Rule Violations in Double-sided Signal Routing
Che-Ping Tsai (National Tsing Hua University); Fang-Yu Hsu (National Tsing Hua University); Wai-Kei Mak (National Tsing Hua University); Ting-Chi Wang (National Tsing Hua University)

14:30 - 15:00
Coffee Break | Exhibits
Room: Salons D-H

15:00 - 15:45
Application Specific Accelerations
Room: Salons 1-3

15:00
1280: Fast and Efficient 2-bit LLM Inference on GPU: 2/4/16-bit in a Weight Matrix with Asynchronous Dequantization
Jinhao Li (Shanghai Jiao Tong University); Jiaming Xu (Shanghai Jiao Tong University); Shiyao Li (Tsinghua University); Shan Huang (Shanghai Jiao Tong University); Jun Liu (Shanghai Jiao Tong University); Yaoxiu Lian (Shanghai Jiao Tong University); Guohao Dai (Shanghai Jiao Tong University)

15:15
1140: AGC: A Unified Architecture for Accelerating K-Nearest Neighbor Graph Construction in Vector Search
Lei Dai (SKLP,Institute of Computing Technology,Chinese Academy of Sciences;University of Chinese Academy of Sciences); Ziming Yuan (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing; University of Chinese Academy of Sciences); Shengwen Liang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing; University of Chinese Academy of Sciences); Wen Li (School of Computer and Information Technology, Shanxi University); Kaiwei Zou (Tsinghua University); Ying Wang (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences); Cheng Liu (Institute of Computing Technology, Chinese Academy of Sciences); Huawei Li (Institute of Computing Technology, Chinese Academy of Sciences); Xiaowei Li (ICT, Chinese Academy of Sciences)

15:30
965: Partial Differential Equation Acceleration by Exploiting Value Similarity
Zehua Li (Xi'an Jiaotong University); Kaisheng Ma (Tsinghua University)

15:00 - 15:45
Enabling Sustainable Next Generation IoT and CPS
Room: Lincoln/Holland/Columbia

15:00
789: A Sparsity-Aware Autonomous Path Planning Accelerator with Algorithm-Architecture Co-Design
Yanjun Zhang (Beijing Institute of Technology); Xiaoyu Niu (Beijing Institute of Technology); Yifan Zhang (University of California, Irvine); Hongzheng Tian (University of California, Irvine); Bo Yu (Shenzhen Institute of Artificial Intelligence and Robotics for Society); Shaoshan Liu (Shenzhen Institute of Artificial Intelligence and Robotics for Society); Sitao Huang (University of California, Irvine)

15:15
907: HDXpose: Harnessing Hyperdimensional Computing's Explainability for Adversarial Attacks
Fatemeh Asgarinejad (UCSD); Flavio Ponzina (University of California San Diego); Onat Gungor (UCSD); Tajana Rosing (UCSD); Baris Aksanli (San Diego State University)

15:30
546: Hybrid Power Failure Recovery for Intermittent Computing
Gan Fang (Purdue University); Jongouk Choi (University of Central Florida); Changhee Jung (Purdue University)

15:00 - 15:45
Cycle-Accurate Timing Models, RISC-V Test Failure Analysis, and Low-Power Design Verification
Room: Salons A-C

The papers from this session address several challenges in simulation and verification. The first paper introduces a technique for generating cycle-accurate timing models for hardware accelerators from their register transfer level descriptions based on dependency analysis and constraint solving. The second paper details a modular, open-source framework designed to automate verification for RISC-V extensions, which reduces the need for manual result analysis by isolating failing instructions. The third paper introduces automatic algorithms to verify the partial set of retention registers identified to maintain correct design functionality when low-power circuits are powered off and on.

15:00
568: Automatic Generation of Timing Models from RTL for Hardware Accelerators
Yu Zeng (Princeton University); Aarti Gupta (Princeton University); Sharad Malik (Princeton University)

15:15
1013: Single Instruction Isolation for RISC-V Vector Test Failures
Manfred Schlaegl (Johannes Kepler University); Daniel Grosse (Johannes Kepler University Linz)

15:30
1061: Automatic Verification and Identification of Partial Retention Register Sets for Low-Power Designs
Yu-An Shih (Princeton University); Sharad Malik (Princeton University)

15:00 - 15:45
Bayesian Techniques for Software-Hardware Co-Optimization and Routing
Room: Skylands/Gateway

15:00
699: Multi-Objective Software-Hardware Co-Optimization for HD-PIM via Noise-Aware Bayesian Optimization
Chien-Yi Yang (University of California San Diego); Minxuan Zhou (UCSD); Flavio Ponzina (University of California San Diego); Suraj Sathya Prakash (UCSD); Raid Ayoub (Intel corporation); Pietro Mercati (Intel Labs); Mahesh Subedar (Intel Labs); Tajana Rosing (UCSD)

15:15
985: RABER: Reliability-Aware Bayesian-Optimization-based Control Layer Escape Routing for Flow-based Microfluidics
Siyuan Liang (The Chinese University of Hong Kong); Rongliang Fu (The Chinese University of Hong Kong); Mengchu Li (Technical University of Munich); Tsun-Ming Tseng (Technical University of Munich); Ulf Schlichtmann (Technical University of Munich); Tsung-Yi Ho (The Chinese University of Hong Kong)

15:30
705: Bayesian-Informed Hyperdimensional Learning for Intelligent and Efficient Data Processing
Hamza Errahmouni Barkam (University Of California Irvine); Tamoghno Das (University of California, Irvine); Prathyush P. Poduval (ICS, UCI); SungHeon Jeong (UCI); Calvin Yeung (University of California Irvine); Mostafa A. Solitan (Alexandria university); Mohsen Imani (University of California Irvine)

15:00 - 15:45
Design Frameworks and Post-place Optimization
Room: Essex/Liberty

In this session, we explore two cutting-edge topics. In the first paper, the authors introduce SeGen, an automated framework for generating sequencing elements (such as flip-flops) in digital integrated circuits. SeGen covers a wide range of designs, outperforming human-crafted solutions. In the second paper, the authors discuss a novel reinforcement learning-based post-place optimization framework that simultaneously optimizes power, performance, and area. By dynamically selecting effective actions, this approach surpasses traditional co-optimization methods, achieving improved Pareto-frontier sets.

15:00
609: SeGen: Automatic Topology Generator for Sequencing Elements
Kyounghun Kang (Korea Advanced Institute of Science and Technology); Wanyeong Jung (KAIST)

15:15
796: Improving Timing & Power Trade-off in Post-place Optimization Using Multi-agent Reinforcement Learning
Jaemin Seo (Pohang University of Science and Technology (POSTECH)); Sejin Park (POSTECH); Seokhyeong Kang (Pohang University of Science and Technology)

15:00 - 15:45
Advances in Analog and RF Synthesis: Machine Learning Techniques and Thermal Analysis
Room: Salons F-H

This session presents the latest advances in analog and RF synthesis and modeling techniques. The first paper introduces a new machine learning-based synthesis flow, transitioning from S-parameters to layout for RF passive connector design. The second paper demonstrates a comprehensive analog hierarchical synthesis flow, supported by reinforcement learning in Python environments, applicable to both standard and low-temperature designs, and available as open source. The third paper discusses recent developments in machine learning-based full-chip thermal analysis, accounting for FinFET self-heating effects in RF power amplifiers.

15:00
1130: PulseRF: Physics Augmented ML Modeling and Synthesis for High-Frequency RFIC Design
Hyunsu Chae (University of Texas at Austin); Hao Yu (University of Texas at Austin); Sensen Li (University of Texas at Austin); David Z. Pan (University of Texas at Austin)

15:15
1500: Reinforcement Learning-Enhanced Cloud-Based Open Source Analog Circuit Generator for Standard and Cryogenic Temperatures in 130-nm and 180-nm OpenPDKs
Ali Hammoud (University of Michigan); Anhang Li (University of Michigan - Ann Arbor); Ayushman Tripathi (University of Michigan - Ann Arbor); Wen Tian (University of Michigan - Ann Arbor); Harsh Khandeparkar (University of Michigan - Ann Arbor); Ryan Wans (University of Michigan - Ann Arbor); Gregory Kielian (Google AI); Boris Murmann (University of Hawai'i); Dennis Sylvester (University of Michigan - Ann Arbor); Mehdi Saligane (University of Michigan - Ann Arbor)

15:30
1534: Analyzing the Impact of FinFET Self-Heating on the Performance of RF Power Amplifiers
Nibedita Karmokar (University of Minnesota); Sai-Wang Tam (NXP); Thanh Viet Dinh (NXP); Vidya A. Chhabria (Arizona State University); Ramesh Harjani (University of Minnesota); Sachin S. Sapatnekar (University of Minnesota)

19:00 - 21:45
ICCAD on Broadway
***Transportation will be provided

Technical Program: 30 October 2024

7:30 – 8:30
Registration
Room: Grand Ballroom Foyer

8:30 – 9:30
Keynote: Coming Soon!
Dilma Da Silva, US National Science Foundation
Room: Salons 4-8

9:30 – 10:00
Award Ceremony
Room: Salons 4-8

10:00 – 10:30
Coffee Break | Exhibits
Room: Salons D-H

10:30 – 12:00
Top Picks Workshop
Room: Essex/Liberty

10:30 - 11:15
Microarchitecture Support for Security
Room: Salons 1-3

10:30
707: On the Security Vulnerabilities of MRAM-based In-Memory Computing Architectures against Model Extraction Attacks
Saion K. Roy (University of Illinois at Urbana-Champaign); Naresh Shanbhag (University of Illinois at Urbana-Champaign)

10:45
777: An FPGA-based Key-Switching Accelerator with Ultra-High Throughput for FHE
Zhaojun Lu (School of Cyber Science and Engineering, Huazhong University of Science and Technology); Peng Xu (School of Cyber Science and Engineering, Huazhong University of Science and Technology); Yijie Wang (Huazhong University of Science and Technology); Yifan Yang (School of Cyber Science and Engineering, Huazhong University of Science and Technology); Qidong Chen (School of Cyber Science and Engineering, Huazhong University of Science and Technology); Weizong Yu (School of Cyber Science and Engineering, Huazhong University of Science and Technology); Gang Qu (University of Maryland, College Park)

11:00
1581: µLAM: A LLM-Powered Assistant for Real-Time Micro-architectural Attack Detection and Mitigation
Upasana Mandal (Indian Institute of Technology, Kharagpur); Shubhi Shukla (Indian Institute of Technology Kharagpur); Ayushi Rastogi (Indian Institute of Technology Kharagpur); Sarani Bhattacharya (Indian Institute of Technology Kharagpur); Debdeep Mukhopadhyay (Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur)

10:30 - 11:15
New Research Developments in Synthesis
Room: Lincoln/Holland/Columbia

This session presents novel methods to optimize technology mapping and physical synthesis to create better chips. The first paper presents an improvement to cut choices in technology mapping based on machine learning. The second paper introduces a physical synthesis infrastructure based on a flexible intermediate representation. The third paper embraces technology mapping and physical synthesis with gate placement based on a novel wirelength-driven mapping algorithm.

10:30
1288: A Machine Learning Guided Cut Choices for ASIC Technology Mapping
Chandan Karfa (Indian Institute of Technology Guwahati); Chandrabhushan Reddy Chigarapall (IIT Guwahati); Harshwardhan Bhakkad (IIT Guwahati); Sukanta Bhattacharjee (Iitg); Animesh Basak Chowdhury (New York University)

10:45
522: RapidIR: A Practical Infrastructure for FPGA High-Level Physical Synthesis
Jason Lau (UCLA); Yuanlong Xiao (RapidStream Design Automation, Inc.); Yutong Xie (RapidStream Design Automation, Inc.); Yuze Chi (RapidStream Design Automation, Inc.); Linghao Song (UCLA); Sihao Liu (UCLA); Shaojie Xiang (Cornell University); Michael Lo (UCLA); Zhiru Zhang (Cornell University); Jason Cong (UCLA); Licheng Guo (RapidStream Design Automation, Inc.)

11:00
670: Physically Aware Synthesis Revisited: Guiding Technology Mapping with Primitive Logic Gate Placement
Hongyang Pan (Fudan university); Cunqing Lan (Fudan university); Yiting Liu (Fudan University); Zhiang Wang (University of California San Diego); Li Shang (fudan university); Xuan Zeng (Fudan University); Fan Yang (Fudan University); Keren Zhu (The Chinese University of Hong Kong)

10:30 - 11:15
CTS and FPGA Routing
Room: Salons A-C

The session starts with the presentation of a unique CTS methodology, OCTS, that synthesizes optical clock trees with EDA. The StarRoute and Potter papers compete in a virtual FPGA routing contest, enamored with novelties in space and parallelism exploration.

10:30
980: OCTS: An Optical Clock Tree Synthesis Methodology for 2.5D Systems
Aristotelis Tsekouras (Aristotle University of Thessaloniki); Georgios Kyriazidis (Harvard University); Gage Hills (Harvard University); Vasilis Pavlidis (University of Manchester)

10:45
808: AceRoute: Adaptive Compute-Efficient FPGA Routing with Pluggable Intra-Connection Bidirectional Exploration
Xinming Wei (Peking University); Ziyun Zhang (Peking University); Sunan Zou (School of Computer Science, Peking University); Kaiwen Sun (Deepoly); Jiahao Zhang (Peking University); Jiaxi Zhang (Peking University); Ping Fan (DeePoly Technology Inc.); Guojie Luo (Peking University)

11:00
1226: Potter: A Parallel Overlap-Tolerant Router for UltraScale FPGAs
Xinshi Zang (The Chinese University of Hong Kong); Wenhao Lin (The Chinese University of Hong Kong); Jinwei Liu (The Chinese University of Hong Kong); Evangeline Young (The Chinese University of Hong Kong)

10:30 - 11:15
PIM PIM PIM
Room: Skylands/Gateway

10:30
614: NAND-Tree: A 3D NAND Flash Based Processing In Memory Accelerator for Tree-Based Models on Large-Scale Tabular Data
Hongtao Zhong (Tsinghua University); Taixin Li (Tsinghua University); Yiming Chen (Tsinghua University); Wenjun Tang (Tsinghua University); Juejian Wu (Tsinghua University); Huazhong Yang (Tsinghua University); Xueqing Li (Tsinghua University)

10:45
1156: A Processing-using-Memory Architecture for Commodity DRAM Devices with Enhanced Compatibility and Reliability
Hoon Shin (Seoul National University, Samsung Electronics); Rihae Park (Seoul National University); Jae W. Lee (Seoul National University)

11:00
1195: Towards Floating Point-Based Attention-Free LLM: Hybrid PIM with Non-Uniform Data Format and Reduced Multiplications
Lidong Guo (Tsinghua University); Zhenhua Zhu (Tsinghua University); Tengxuan Liu (Tsinghua University); Xuefei Ning (Tsinghua University); Shiyao Li (Tsinghua University); Guohao Dai (Shanghai Jiao Tong University); Huazhong Yang (Tsinghua University); Wangyang Fu (Tsinghua University); Yu Wang (Tsinghua University)

10:30 - 11:15
Real-Time AI: Co-Designing for the Edge
Room: Salons F-H

This session dives into methods for optimizing real-time object detection and transformer network inference on edge devices. It emphasizes automated deployment strategies, software-hardware co-design, and the integration of binarized neural networks with FPGA technology to enhance computational efficiency and accuracy. The papers discuss end-to-end solutions and co-design approaches that significantly improve the performance and deployment of AI applications on edge platforms.

10:30
550: AyE-Edge: Automated Deployment Space Search Empowering Accuracy yet Efficient Real-Time Object Detection on the Edge
Chao Wu (Northeastern University); Yifan Gong (Northeastern University); Liangkai Liu (University of Michigan); Mengquan Li (Hunan University); Yushu Wu (Northeastern University); Xuan Shen (Northeastern University); Zhimin Li (Northeastern University); Geng Yuan (University of Georgia); Weisong Shi (Wayne State University); Yanzhi Wang (Northeastern University)

10:45
615: Edge-BiT: Software-Hardware Co-design for Optimizing Binarized Transformer Networks Inference on Edge FPGA
Shuai Zhou (Fudan University); Sisi Meng (Fudan University); Huinan Tian (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

11:00
949: Co-Designing Binarized Transformer and Hardware Accelerator for Efficient End-to-End Edge Deployment
Yuhao Ji (Nanjing University); Chao Fang (Nanjing University); Shaobo Ma (Nanjing University); Haikuo Shao (Nanjing University); Zhongfeng Wang (Nanjing University)

11:15 - 12:00
Security by Design and Pre-silicon Security Assurance
Room: Salons 1-3

11:15
571: HybriDIFT: Scalable Memory-Aware Dynamic Information Flow Tracking for Hardware
Flavien Solt (ETH Zurich); Kaveh Razavi (ETH Zurich)

11:30
852: VeriCHERI: Exhaustive Formal Security Verification of CHERI at the RTL
Anna Lena Duque Antón (RPTU Kaiserslautern-Landau); Johannes Müller (RPTU Kaiserslautern-Landau); Philipp Schmitz (RPTU Kaiserslautern-Landau); Tobias Jauch (RPTU Kaiserslautern-Landau); Alex Wezel (RPTU Kaiserslautern-Landau); Lucas Deutschmann (University of Kaiserslautern-Landau); Mohammad Rahmani Fadiheh (Technische Universität Kaiserslautern); Dominik Stoffel (TU Kaiserslautern); Wolfgang Kunz (TU Kaiserslautern)

11:45
1578: eXpect: On the Security Implications of Violations in AXI Implementations
Melisande Zonta (ETH Zürich); Andres Meza (UCSD); Nora Hinderling (ETH); Lucas Deutschmann (University of Kaiserslautern-Landau); Francesco Restuccia (University of California San Diego); Ryan Kastner (UCSD); Shweta Shinde (ETH Zurich)

11:15 - 12:00
A New Life to Logic Synthesis
Room: Lincoln/Holland/Columbia

This session presents innovative methods and applications for logic synthesis. The first paper presents a new architecture for circuit representation learning based on transformers. The second paper improves the understanding of circuit functionality for AIGs. The third paper applies rarity-reducing logic synthesis for hardware security.

11:15
1249: DeepGate3: Towards Scalable Circuit Representation Learning
Zhengyuan Shi (The Chinese University of Hong Kong); Ziyang Zheng (The Chinese University of Hong Kong); Sadaf Khan (The Chinese University of Hong Kong); Jianyuan Zhong (The Chinese University of Hong Kong); Min Li (Huawei Noah's Ark Lab); Qiang Xu (The Chinese University of Hong Kong)

11:30
1560: PolarGate: Breaking the Functionality Representation Bottleneck of And-Inverter Graph Neural Network
Jiawei Liu (Beijing University of Posts and Telecommunications); Jianwang Zhai (Beijing University of Posts and Telecommunications); Mingyu Zhao (Beijing University of Posts and Telecommunications); Zhe Lin (Sun Yat-sen University); Bei Yu (The Chinese University of Hong Kong); Chuan Shi (Beijing University of Posts and Telecommunications)

11:45
947: RareLS: Rarity-Reducing Logic Synthesis for Mitigating Hardware Trojan Threats
Chang Meng (EPFL); Mingfei Yu (EPFL); Hanyu Wang (ETH Zurich); Wayne Burleson (U Massachusetts Amherst); Giovanni De Micheli (École Polytechnique Fédérale de Lausanne (EPFL))

11:15 - 12:00
Machine Learning for P&R and Post-P&R
Room: Salons A-C

What does machine learning have to do the P&R and post-layout optimization? This session brings together latest approaches in GNNs, attention networks and reinforcement learning to solve some of the hardest problems, old and new, in these domains.

11:15
702: GAT-Steiner: Rectilinear Steiner Minimal Tree Prediction Using GNNs
Bugra Onal (University of California Santa Cruz); Eren Dogan (University of California, Santa Cruz); Muhammad Hadir Khan (University of California Santa Cruz); Matthew Guthaus (UC Santa Cruz)

11:30
1530: Placement Tomography-Based Routing Blockage Generation for DRV Hotspot Mitigation
Andrew Kahng (UCSD); Sayak Kundu (University of California, San Diego); Dooseok Yoon (University of California, San Diego)

11:45
1278: RL-Fill: Timing-Aware Fill Insertion Using Reinforcement Learning
Jinoh Cho (Pohang University of Science and Technology (POSTECH)); Seonghyeon Park (POSTECH); Jakang Lee (Pohang University of Science and Technology); Sung-Yun Lee (Pohang University of Science and Technology (POSTECH)); Jinmo Ahn (Postech); Seokhyeong Kang (Pohang University of Science and Technology)

11:15 - 12:00
Bringing Device Flavours
Room: Skylands/Gateway

11:15
867: Multi-phase Coupled CMOS Ring Oscillator based Potts Machine
Yilmaz Ege Gonul (Drexel University); Baris Taskin (Drexel University)

11:30
1521: Accurate, Yet Scalable: A SPICE-based Design and Optimization Framework for eNVM based Analog In-memory Computing
S M Mojahidul Ahsan (The University of Kansas); Muhammad Sakib Shahriar (Ulkasemi Inc.); Mrittika Chowdhury (University of Mississippi); Tanvir Hossain (The University of Kansas); Md Sakib Hasan (University of Mississippi); Tamzidul Hoque (The University of Kansas)

11:15 - 12:00
IP, Side-Channels, and Acceleration
Room: Salons F-H

This session presents a set of hardware approaches to protect static design information as well as dynamic application data. The first paper describes the use of encryption to protect the IP of a Spiking Neural Network implementation. The second paper presents a logic synthesis approach to improve resilience against side-channel attacks. The final paper proposes a hardware accelerator to efficiently compute a hash function used for Zero-Knowledge proofs.

11:15
915: SNNGX: Securing Spiking Neural Networks with Genetic XOR Encryption on RRAM-based Neuromorphic Accelerator
Kwun Hang WONG (University of Hong Kong); Songqi Wang (University of Hong Kong); Wei Huang (University of Hong Kong); Xinyuan Zhang (University of Hong Kong); Yangu He (University of Hong Kong); Karl M.H. Lai (University of Hong Kong); Yuzhong Jiao (AI Chip Center for Emerging Smart Systems (ACCESS)); Ning Lin (The University of Hong Kong); Xiaojuan Qi (University of Hong Kong); Xiaoming Chen (Institute of Computing Technology, Chinese Academy of Sciences); Zhongrui Wang (University of Hong Kong)

11:30
1472: ASCENT: Amplifying Power Side-Channel Resilience via Learning & Monte-Carlo Tree Search
Jitendra Bhandari (New York University); Animesh Basak Chowdhury (New York University); Ozgur Sinanoglu (New York University Abu Dhabi); Siddharth Garg (New York University); Ramesh Karri (NYU); Johann Knechtel (New York University Abu Dhabi)

11:45
1361: AMAZE: Accelerated MiMC Hardware Architecture for Zero-Knowledge Applications on the Edge
Anees Ahmed (Arizona State University); Nojan Sheybani (UC San Diego); Davi De Almeida (Arizona State University); Tengkai Gong (University of California San Diego); Nges Njungle (Arizona State University); Michel Kinsy (Arizona State University); Farinaz Koushanfar (University of California San Diego)

12:00 – 13:30
Lunch
Room: Salons 4-8

Technical Program: 30 October 2024

13:30 – 15:30
Tutorial: Advanced Sparse Linear Solver for Transistor-Level Circuit Simulation
Room: Salons 1-3

13:30 – 15:30
Special Session: 2024 CAD Contests at ICCAD
Room: Lincoln/Holland/Columbia

13:30 – 15:30
Top Picks Workshop
Room: Essex/Liberty

13:30 - 14:30
Swift LLMs: Easier Design, Faster Inference
Room: Salons A-C

This session explores various innovations in hardware design tailored for large language model (LLM) optimization and acceleration. The topics discussed include developing agile frameworks for LLM accelerator creation, speculative scheduling to enhance LLM serving, and dynamic approaches for efficient token processing in parallel decoding. Additionally, there's an emphasis on integrating FPGA-based solutions to handle unstructured sparsity in large language models, showcasing advancements in both theoretical and practical aspects of LLM deployment.

13:30
1005: An Agile Framework for Efficient LLM Accelerator Development and Model Inference
Lvcheng Chen (Zhejiang University); Ying Wu (Zhejiang University); Chenyi Wen (Zhejiang University); Shizhang Wang (Hubei University of Technology); Li Zhang (Hubei University of Technology); Bei Yu (The Chinese University of Hong Kong); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)

13:45
575: ALISE: Accelerating Large Language Model Serving with Speculative Scheduling
Youpeng Zhao (University of Central Florida); Jun Wang (University of Central Florida)

14:00
757: ProPD: Dynamic Token Tree Pruning and Generation for LLM Parallel Decoding
Shuzhang Zhong (Peking University); Zebin Yang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University); Ruihao Gong (SenseTime); Runsheng Wang (Peking University); Ru Huang (Peking University)

14:15
1082: ChatOPU: An FPGA-based Overlay Processor for Large Language Models with Unstructured Sparsity
Tiandong Zhao (University of California, Los Angeles); Shaoqiang Lu (Shanghai Jiao Tong University, Shanghai, China; Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China); Chen Wu (Ningbo Institute of Digital Twin, Ningbo, China); Lei He (Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China; University of California, Los Angeles)

13:30 - 14:30
Time is Limited: Fast and Secure Neural Network Accelerators
Room: Skylands/Gateway

This session examines strategies for improving the reliability and performance of neural network systems in critical applications like self-driving vehicles. It discusses techniques such as latency-constrained scheduling to ensure timely and reliable perception, eager gradient prediction to enhance training efficiency of attention mechanisms, and methods for reinforcing DNN accelerator integrity through selective and permuted recomputation. These approaches aim to boost both the speed and accuracy of AI applications, ensuring safer and more efficient operations.

13:30
741: LACO: A Latency-Constraint Offline Neural Network Scheduler towards Reliable Self-Driving Perception
Zhanhong Tan (State Key Laboratory of Intelligent Vehicle Safety Technology); Zijian Zhu (Tsinghua University); Mengdi Wu (Tsinghua University); Kaisheng Ma (Tsinghua University)

13:45
1148: OFT: An accelerator with eager gradient prediction for attention training
Miao Wang (Northwestern Polytechnical University); Shengbing Zhang (Northwestern Polytechnical University); Sijia Wang (Northwestern Polytechnical University); Zhao Yang (Chang'an University); Meng Zhang (Northwestern Polytechnical University)

14:00
894: Enhancing DNN Accelerator Integrity via Selective and Permuted Recomputation
Jhon Ordoñez (University of Delaware); Chengmo Yang (University of Delaware)

13:30 - 14:30
Private Machine Learning Inference
Room: Salons F-H

This session presents a series of approaches to perform machine learning inference privately and efficiently. The first paper reduces communication costs in distributed inference by co-optimizing quantization communication protocol constraints. The second, third, and fourth papers all employ homomorphic encryption to establish private inference. The second paper describes an approach to automatically generate efficient kernels for homomorphic encryption. The solution presented in the third paper combines homomorphic encryption with Garbled Circuits. The fourth paper focuses on reducing the latency induced by homomorphism in the convolution layers of CNNs.

13:30
596: PrivQuant: Communication-Efficient Private Inference with Quantized Network/Protocol Co-Optimization
Tianshi Xu (Peking University); Shuzhang Zhong (Peking University); Wenxuan Zeng (Peking University); Runsheng Wang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University)

13:45
973: FlexHE: A flexible Kernel Generation Framework for Homomorphic Encryption-Based Private Inference
Jiangrui Yu (Peking University); Wenxuan Zeng (Peking University); Tianshi Xu (Peking University); Renze Chen (Peking University); Yun (Eric) Liang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University)

14:00
1237: APINT: A Full-Stack Framework for Acceleration of Privacy-Preserving Inference of Transformers based on Garbled Circuits
Hyunjun Cho (Korea Advanced Institute of Science and Technology (KAIST)); Jaeho Jeon (KAIST); Jaehoon Heo (KAIST); Joo-Young Kim (KAIST)

14:15
1552: Hyena: Optimizing Homomorphically Encrypted Convolution for Private CNN Inference
Hyeri Roh (Seoul National University); Woo-Seok Choi (Seoul National University)

14:30 - 15:30
CIM is on the Run: Sparser and More Robust Designs
Room: Salons A-C

This session focuses on architectures and strategies for enhancing in-memory computation for AI applications. Discussions include leveraging hybrid RRAM-SRAM designs for accelerating sparse transformers and eigen-decomposition, along with probabilistic approximation techniques to optimize sparsity-centric computation. Furthermore, the session explores variation-resilient memory solutions and specialized accelerators for processing complex data structures like voxel-based point clouds, pushing the boundaries of efficiency and performance in AI hardware.

14:30
594: AESHA: Accelerating Eigen-decomposition-based Sparse Transformer with Hybrid RRAM-SRAM Architecture
Xuliang Yu (Zhejiang University); Tianwei Ni (Zhejiang University); Xinsong Sheng (Zhejiang University); Yun Pan (Zhejiang University); Lei He (University of California, Los Angeles); Liang Zhao (Zhejiang University)

14:45
802: PACiM: A Sparsity-Centric Hybrid Compute-in-Memory Architecture via Probabilistic Approximation
Wenlun Zhang (Keio University); Shimpei Ando (Keio University); Yung-Chin Chen (National Taiwan University); Satomi Miyagi (Keio University); Shinya Takamaeda-Yamazaki (The University of Tokyo); Kentaro Yoshioka (Keio University)

15:00
1003: ReSCIM: Variation-Resilient High Weight-Loading Bandwidth In-Memory Computation Based on Fine-Grained Hybrid Integration of Multi-Level ReRAM and SRAM Cells
Xiaomeng WANG (The Hong Kong University of Science and Technology); Jingyu HE (The Hong Kong University of Science and Technology); Kunming SHAO (The Hong Kong University of Science and Technology); Jiakun ZHENG (The Hong Kong University of Science and Technology); Fengshi TIAN (The Hong Kong university of Science and Technology); Tim Kwang-Ting CHENG (The Hong Kong University of Science and Technology); Chi-Ying TSUI (The Hong Kong University of Science and Technology)

15:15
838: Voxel-CIM: An Efficient Compute-in-Memory Accelerator for Voxel-based Point Cloud Neural Networks
Xipeng Lin (The Hong Kong University of Science and Technology (Guangzhou)); Shanshi Huang (Hong Kong University of Science and Technology (Guangzhou)); Hongwu Jiang (The Hong Kong University of Science and Technology (Guangzhou))

14:30 - 15:30
Treasures in the Graphs: Efficient Designs for GNNs
Room: Skylands/Gateway

This session delves into innovative approaches for optimizing graph neural network (GNN) computations on FPGA platforms. It features discussions on leveraging advanced dataflow techniques and high-bandwidth memory to accelerate sparse matrix multiplications, algorithm-hardware co-design for efficient large-scale GNN training, and the development of memory-optimized processors tailored for parallel graph mining. These papers highlight the ongoing advancements in hardware configurations and algorithmic strategies to enhance the performance and scalability of GNN applications.

14:30
1169: Leda: Leveraging Tiling Dataflow to Accelerate SpMM on HBM-Equipped FPGAs for GNNs
Enxin Yi (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Jiarui Bai (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Yijie Nie (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Dan Niu (Southeast University); Zhou Jin (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Weifeng Liu (Super Scientific Software Laboratory, China University of Petroleum-Beijing)

14:45
1561: CoCoA: Algorithm-Hardware Co-Design for Large-Scale GNN Training using Compressed Graph
Yunki Han (KAIST); Jaekang Shin (Korea Advanced Institute of Science and Technology (KAIST)); Gunhee Park (Samsung Electronics); Lee-Sup Kim (KAIST)

15:00
968: FLOP: A Flexible Memory-Optimized Processor for Parallel Graph Mining on FPGA
Guoyu Li (Fudan University); Runzhou Zhang (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

14:30 - 15:30
More than Matrix Multiplication: Efficient Designs for Neural Networks
Room: Salons F-H

This session explores in-memory computing architectures designed to enhance the efficiency and throughput of deep neural networks and transformers. Topics include a reprogramming-free RRAM-based approach for optimizing deep neural network computations through basis combination, a transposable digital SRAM architecture tailored for energy-efficient transformer acceleration, and a matrix multiplication accelerator that supports various levels of sparsity. These advancements underscore efforts to reduce energy consumption while boosting the computational performance of AI accelerators.

14:30
828: BasisN: Reprogramming-Free RRAM-Based In-Memory-Computing by Basis Combination for Deep Neural Networks
Amro Eldebiky (Technical University of Munich); Grace Li Zhang (TU Darmstadt); Xunzhao Yin (Zhejiang University); Cheng Zhuo (Zhejiang University); Ing-Chao Lin (National Cheng Kung University); Ulf Schlichtmann (Technical University of Munich); Bing Li (Technical University of Munich)

14:45
1579: TP-DCIM: Transposable Digital SRAM CIM Architecture for Energy-Efficient and High Throughput Transformer Acceleration
Junwoo Park (Korea University); Kyeongho Lee (korea univ.); Jongsun Park (Korea University)

15:00
599: FSMM: An Efficient Matrix Multiplication Accelerator Supporting Flexible Sparsity
Yuxuan Qiao (Fudan University); Fan Yang (Fudan University); Yecheng Zhang (State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation); Xiankui Xiong (State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation); Xiao Yao (State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation); Haidong Yao (State Key Laboratory of Mobile Network and Mobile Multimedia Technology, ZTE Corporation)

16:00 - 17:00
Side Channels and Trojans
Room: Salons 1-3

16:00 - 17:00
Top Picks Workshop
Room: Essex/Liberty

16:00 - 17:00
Side Channels and Trojans
Room: Salons 1-3

16:00
724: RandOhm: Mitigating Impedance Side-channel Attacks using Randomized Circuit Configurations
Saleh Khalaj Monfared (Worcester Polytechnic Institute (WPI)); Domenic Forte (University of Florida); Shahin Tajik (Worcester Polytechnic Institute)

16:15
806: Layout-level Hardware Trojan Prevention in the Context of Physical Design
Xingyu Tong (Fudan University); Guohao Chen (Fudan University); Min Wei (Fudan University); Zhijie Cai (Fudan University); Peng Zou (Shanghai LEDA Technology Co., Ltd); Zhifeng Lin (Fuzhou University); Jianli Chen (Fudan University)

16:30
1403: A Built-In Integrated Rowhammer, Rowpress, and Leakage Detection Sensor for DRAM
Nezam Rohbani (Institute for Research in Fundamental Sciences (IPM)); Rouzbeh Pirayadi (Sharif Univerisyt of Technology); Mohammad Arman Soleimani (Sharif University of Technology); Adrian Cristal Kestelman (Barcelona Supercomputing Center); Osman Unsal (Barcelona supercomputing center); Hamid Sarbazi-Azad (Sharif U of Tech)

16:45
1492: LaserEscape: Detecting and Mitigating Optical Probing Attacks
Saleh Khalaj Monfared (Worcester Polytechnic Institute (WPI)); Kyle Mitard (Worcester Polytechnic Institute); Andrew Cannon (University of Florida); Domenic Forte (University of Florida); Shahin Tajik (Worcester Polytechnic Institute)

16:00 - 17:00
Advances in High-Level Synthesis and Optimized Components
Room: Lincoln/Holland/Columbia

This session presents the latest developments in high-level synthesis and the creation of optimized micro-architectural components. The first paper presents a novel intermediate representation to optimize the dynamic scheduling and optimization of memory operations. The second paper introduces an approach based on LLM for high-level synthesis. The third paper discusses a method to estimate circuit metrics at higher levels of abstraction. The fourth paper presents a framework to optimize the generation of multipliers and MACs.

16:00
652: R-HLS: An IR for Dynamic High-Level Synthesis and Memory Disambiguation based on Regions and State Edges
David Christoph Metz (Norwegian University of Science and Technology); Nico Reissmann (Independent Researcher); Magnus Själander (Norwegian University of Science and Technology)

16:15
1222: HLSPilot: LLM-based High-Level Synthesis
Chenwei Xiong (Institute of Computing Technology, Chinese Academy of Sciences); Cheng Liu (Institute of Computing Technology, Chinese Academy of Sciences); Huawei Li (Institute of Computing Technology, Chinese Academy of Sciences); Xiaowei Li (ICT, Chinese Academy of Sciences)

16:30
1240: Balor: HLS Source Code Evaluator Based on Custom Graphs and Hierarchical GNNs
Emmet Murphy (ETH Zurich); Lana Josipovic (ETH Zurich)

16:45
1265: UFO-MAC: A Unified Framework for Optimization of High-Performance Multipliers and Multiply-Accumulators
Dongsheng Zuo (The Hong Kong University of Science and Technology (Guangzhou)); Jiadong ZHU (The Hong Kong University of Science and Technology (Guangzhou)); Chenglin Li (The Hong Kong University of Science and Technology (Guangzhou)); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou))

16:00 - 17:00
Innovations in Neuromorphic Hardware and 3D Integration
Room: Salons A-C

16:00
1516: Spiking Transformer Hardware Accelerators in 3D Integration
Boxun Xu (University of California, Santa Barbara); Junyoung Hwang (Georgia Institute of Technology); Pruek Vanna-iampikul (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech); Peng Li (University of California, Santa Barbara)

16:15
1397: Neural Architecture Search for Highly Bespoke Robust Printed Neuromorphic Circuits
Priyanjana Pal (Karlsruhe Institute of Technology); Haibin Zhao (Karlsruhe Institute of Technology); Tara Gheshlaghi (Karlsruhe Institute of Technology); Michael Hefenbrock (RevoAI GmbH); Michael Beigl (Karlsruhe Institute of Technology (KIT)); Mehdi Tahoori (Karlsruhe Institute of Technology)

16:30
809: An O(m+n)-Space Spatiotemporal Denoising Filter with Cache-Like Memories for Dynamic Vision Sensors
Qinghang Zhao (Xidian University); Jiaqi Wang (Xidian University); Yixi Ji (Xidian University); Jinjian Wu (Xidian University); Guangming Shi (Xidian University)

16:45
739: LSMR: Synergy Randomness in Liquid State Machine and RRAM-based Analog-digital Accelerator
Ning Lin (The University of Hong Kong); Songqi Wang (The University of Hong Kong); Xinyuan Zhang (The University of Hong Kong); Shaocong Wang (the University of Hong Kong); Yangu He (The University of Hong Kong); Woyu Zhang (Institute of Microelectronics, Chinese Academy of Sciences); Bo Wang (The University of Hong Kong); Jiankun Li (The University of Hong Kong); Mingzi Li (The University of Hong Kong); Binbin Cui (The University of Hong Kong); Yi Li (The University of Hong Kong); Jia Chen (The Hong Kong University of Science and Technology); Chunwei Xia (University of Leeds); Wei Xuan (AI Chip Center for Emerging Smart Systems (ACCESS)); Xiaoming Chen (Institute of Computing Technology, Chinese Academy of Sciences); Dashan Shang (Institute of Microelectronics, Chinese Academy of Sciences); Zhongrui Wang (The University of Hong Kong)

16:00 - 17:00
Precision Matters: Improving the Robustness and Reconfigurability
Room: Skylands/Gateway

This session dives into the development of specialized architectures and computational formats to optimize neural network inference under constrained conditions. Discussions include a novel number format designed to enhance the robustness of sub-8-bit neural network operations, the design of a reconfigurable accelerator for dynamic adaptation in AI tasks, and the implementation of mixed-precision neural networks through ISA extensions for soft SIMD operations on RISC-V cores. These innovations are aimed at improving efficiency, flexibility, and performance in AI processing environments.

16:00
1248: FlexInt: A New Number Format for Robust Sub-8-Bit Neural Network Inference
Minuk Hong (Ulsan National Institute of Science and Technology (UNIST)); Hyeonuk Sim (Samsung Advanced Institute of Technology); Sugil Lee (Ulsan National Institute of Science and Technology (UNIST)); Jongeun Lee (Ulsan National Institute of Science and Technology (UNIST))

16:15
1293: MARCA: Mamba Accelerator with Reconfigurable Architecture
Jinhao Li (Shanghai Jiao Tong University); Shan Huang (Shanghai Jiao Tong University); Jiaming Xu (Shanghai Jiao Tong University); Jun Liu (Shanghai Jiao Tong University); Li Ding (Shanghai Jiao Tong University); Ningyi Xu (Shanghai Jiao Tong University); Guohao Dai (Shanghai Jiao Tong University)

16:30
1600: Mixed-precision Neural Networks on RISC-V Cores: ISA extensions for Multi-Pumped Soft SIMD Operations
Giorgos Armeniakos (National Technichal University of Athens); Alexis Maras (National Technical University of Athens); Sotirios Xydis (National Technical University of Athens); Dimitrios Soudris (NTUA)

16:00 - 17:00
Sparsity Matters: Sparse Computing Engines for Different Platforms
Room: Salons F-H

This session explores hardware designs and algorithmic adaptations to enhance the performance and energy efficiency of AI accelerators. It includes a co-sparse photonic accelerator that integrates algorithmic and circuit design for thermal tolerance and power-efficient light distribution, an approach for exploiting sparsity in feed-forward networks and attention mechanisms in transformers on FPGA, and a dedicated point cloud inference engine optimized for RISC-V processors. These papers highlight the intersection of innovative hardware solutions and algorithmic efficiencies to push the boundaries of AI computational frameworks.

16:00
881: SCATTER: Algorithm-Circuit Co-Sparse Photonic Accelerator with Thermal-Tolerant, Power-Efficient In-situ Light Redistribution
Ziang Yin (Arizona State University); Nicholas Gangi (Rensselaer Polytechnic Institute); Meng Zhang (Rensselaer Polytechnic Institute); Jeff Zhang (Arizona State University); Rena Huang (Rensselaer Polytechnic Institute); Jiaqi Gu (Arizona State University)

16:15
981: FAS-Trans: Fully Exploiting FFN and Attention Sparsity for Transformer on FPGA
Hongji Wang (Fudan University); Yifan Zhang (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

16:30
1179: RISCSparse: Point Cloud Inference Engine on RISC-V Processor
Shangran Lin (The Chinese University of Hong Kong, Shenzhen); Xinrui Zhu (Chinese University of Hong Kong, Shenzhen); baohui xie (Chinese University of Hong Kong (ShenZhen)); Tinghuan Chen (The Chinese University of Hong Kong, Shenzhen); Cheng Zhuo (Zhejiang University); QI SUN (Zhejiang University); Bei Yu (The Chinese University of Hong Kong)

17:00 - 19:00
Job Fair & SIGDA Dinner
Room: Salons 4-8