Submissio	Title	Abstract	Authors with Affiliations
5	SMORE: Similarity-Based Hyperdimensional Domain Adaptation for Multi-Sensor Time Series Classification	Many real-world applications of the Internet of Things (IoT) employ machine learning (ML) algorithms to analyze time series information collected by interconnected sensors. However, distribution shift, a fundamental challenge in data-driven ML, arises when a model is deployed on a data distribution different from the training data and can substantially degrade model performance. Additionally, increasingly sophisticated deep neural networks (DNNs) are proposed to capture intricate spatial and temporal dependencies in multi-sensor time series data, often exceeding the capabilities of today's edge devices. In this paper, we propose SMORE, a novel resource-efficient domain adaptation (DA) algorithm for multi-sensor time series classification, leveraging the ultra-efficient operations of hyperdimensional computing. SMORE dynamically customizes test-time models with explicit consideration of the domain context of each sample to provide accurate predictions when confronted with domain shifts. Our evaluation on a variety of multi-sensor time series classification tasks shows that SMORE achieves on average 1.98% higher accuracy than state-of-the-art (SOTA) DNN-based DA algorithms with 18.81× faster training and 4.85× faster inference.	Junyao Wang (University of California, Irvine); Mohammad Al Faruque (University of California Irvine)
10	EnTurbo: Accelerate Confidential Serverless Computing via Parallelizing Enclave Startup Procedure	Serverless computing has gained widespread attention, and Trusted Execution Environments (TEEs) are well-suited for safeguarding user privacy. However, the additional startup procedure introduced by TEEs imposes considerable performance overhead on confidential serverless workloads. This paper introduces a novel parallelized enclave startup design, EnTurbo, which eliminates the integrity dependence of the enclave startup procedure, accelerating it while ensuring its security. Additionally, EnTurbo parallelizes the measurement procedure, enabling multi-thread measurement for acceleration with provable security. We evaluate EnTurbo by running confidential serverless workloads on SGX simulation mode. Results show that EnTurbo effectively speeds up enclave serverless by 1.42x-6.48x (SGXv1) and 1.33x-3.76x (SGXv2).	Yifan Zhu (Institute of Information Engineering, Chinese Academy of Sciences); Peinan Li (Institute of Information Engineering, CAS); Yunkai Bai (Institute of Information Engineering, CAS); Yubiao Huang (Chinese Academy of Sciences); Shiwen Wang (Institute of Information Engineering, Chinese Academy of Sciences); Xingbin Wang (Institute of Information Engineering, Chinese Academy of Sciences); Dan Meng (Institute of Information Engineering, Chinese Academy of Sciences); Rui Hou (Institute of Information Engineering, Chinese Academy of Sciences)
12	SecPaging: Secure Enclave Paging with Hardware-Enforced Protection against Controlled-Channel Attacks	As a prevalent privacy-preserving technology, Trusted Execution Environment has become widely adopted in numerous commercial processors. Nonetheless, they remain susceptible to various controlled-channel attacks. Untrusted operating systems can deduce enclave secrets by manipulating page tables or observing allocation- or swap-based page faults. In this paper, we propose SecPaging, a novel secure enclave paging mechanism based on hardware-enforced and microcode-supported protection to prevent these attacks. First, enclave PTEs are protected through hardware isolation, preventing privileged attackers from malicious tampering or observations. Second, Eager-Allocation mechanism is employed to prevent allocation-based controlled-channel attacks. Besides, Record-Reload mechanism is proposed to prevent swap-based controlled-channel attacks.	Yunkai Bai (Institute of Information Engineering, CAS); Peinan Li (Institute of Information Engineering, CAS); Yubiao Huang (Chinese Academy of Sciences); Shiwen Wang (Institute of Information Engineering, Chinese Academy of Sciences); Xingbin Wang (Institute of Information Engineering, Chinese Academy of Sciences); Dan Meng (Institute of Information Engineering, Chinese Academy of Sciences); Rui Hou (Institute of Information Engineering, Chinese Academy of Sciences)
16	GDR-HGNN: A Heterogeneous Graph Neural Networks Accelerator Frontend with Graph Decoupling and Recoupling	Heterogeneous Graph Neural Networks (HGNNs) have broadened the applicability of graph representation learning to heterogeneous graphs. However, the irregular memory access pattern of HGNNs leads to the buffer thrashing issue in HGNN accelerators. In this work, we identify an opportunity to address buffer thrashing in HGNN acceleration through an analysis of the topology of heterogeneous graphs. To harvest this opportunity, we propose a graph restructuring method and map it into a hardware frontend named GDR-HGNN. GDR-HGNN dynamically restructures the graph on the fly to enhance data locality for HGNN accelerators. Experimental results demonstrate that, with the assistance of GDR-HGNN, a leading HGNN accelerator achieves an average speedup of 14.6$\times$ and 1.78$\times$ compared to the state-of-the-art software framework running on A100 GPU and itself, respectively.	Runzhen Xue (State Key Lab of Processors, Institute of Computing Technology, CAS; School of Computer Science and Technology, University of Chinese Academy of Sciences); Mingyu Yan (Institute of Computing Technology, Chinese Academy of Sciences); Dengke Han (State Key Lab of Processors, Institute of Computing Technology, CAS; School of Computer Science and Technology, University of Chinese Academy of Sciences); Yihan Teng (State Key Lab of Processors, Institute of Computing Technology, CAS; School of Computer Science and Technology, University of Chinese Academy of Sciences); Zhimin Tang (Institute of Computing Technology, Chinese Academy of Sciences City:Beijing State/Province:Beijing Country/Region:China (CN)); Xiaochun Ye (Institute of Computing Technology, Chinese Academy of Sciences City:Beijing State/Province:Beijing Country/Region:China (CN)); Dongrui Fan (Institute of Computing Technology, Chinese Academy of Sciences City:Beijing State/Province:Beijing Country/Region:China (CN))
18	Garrison: A High-Performance GPU-Accelerated Inference System for Adversarial Ensemble Defense	Adversarial ensemble defense is one of the most effective techniques for defending against adversarial attacks, which constructs ensembles of multiple DNNs to improve the model's robustness. However, deploying ensemble defense methods on existing DNN inference systems is inefficient and impractical due to their dynamics and randomness. To this end, we propose an inference system for adversarial ensemble defense called Garrison, which can deliver robust and low-latency predictions using Multi-Instance GPUs. Our evaluations show that Garrison can improve adversarial robustness by up to 24.5% while accelerating ensemble inference by up to 6.6x compared to the state-of-the-art inference framework.	Yan Wang (Institute of Information Engineering, Chinese Academy of Sciences; School of Cyber Security, University of Chinese Academy of Sciences); Xingbin Wang (Institute of Information Engineering, Chinese Academy of Sciences); Zechao Lin (Institute of Information Engineering,Chinese Academy of Sciences); Yulan Su (Institute of Information Engineering, Chinese Academy of Sciences); Sisi Zhang (Institute of Information Engineering, Chinese Academy of Sciences); Rui Hou (Institute of Information Engineering, Chinese Academy of Sciences); Dan Meng (Institute of Information Engineering, Chinese Academy of Sciences)
21	PHD: Parallel Huffman Decoder on FPGA for Extreme Performance and Energy Efficiency	Huffman decoding is crucial in data compression, and the self-synchronization-based parallel decoding algorithm enables subsequence-level parallelism. This paper introduces PHD, the first accelerator designed for self-synchronization-based parallel Huffman decoding on the Field-Programmable Gate Array (FPGA). Designing PHD poses challenges, including managing fine-grained parallelism, addressing limited on-chip memory, and handling inter-codeword dependency. PHD incorporates bit-level, subsequence-level, and tile-level parallelism, utilizes hybrid memory to store the codebook efficiently, and introduces the ONCE MORE optimization to reduce decoding loop iterations. Experimental results demonstrate that PHD outperforms the state-of-the-art GPU-based baseline regarding latency (9.4X to 12.8X reduction) and energy consumption (12.4X to 18.2X reduction).	Yunkun Liao (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences; Zhongguancun National Laboratory); Jingya Wu (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences); Wenyan Lu (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; YUSUR Technology Co., Ltd); Xiaowei Li (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; Zhongguancun National Laboratory); Guihai Yan (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; YUSUR Technology Co., Ltd)
24	Explainable Fuzzy Neural Network with Multi-Fidelity Reinforcement Learning for Micro-Architecture Design Space Exploration	With the continuous advancement of processors, modern micro-architecture designs have become increasingly complex. The vast design space presents significant challenges for human designers, making design space exploration (DSE) algorithms a significant tool for μ-arch design. In recent years, efforts have been made in the development of DSE algorithms, and promising results have been achieved. However, the existing DSE algorithms, e.g., Bayesian Optimization and ensemble learning, suffer from poor interpretability, hindering designers' understanding of the decision-making process. To address this limitation, we propose utilizing Fuzzy Neural Networks to induce and summarize knowledge and insights from the DSE process, enhancing the interpretability and controllability of DSE results. Furthermore, to improve efficiency, we introduce a multi-fidelity reinforcement learning approach, which primarily conducts exploration using inexpensive but imprecise data, thereby substantially diminishing the reliance on costly data. Experimental results show that our method achieved excellent results with a very limited sample budget and successfully surpasses the current state-of-the-art.	Hanwei FAN (HKUST); Ya Wang (Hong Kong University of Science and Technology); Sicheng Li (Alibaba Group); Tingyuan Liang (The Hong Kong University of Science and Technology); Wei Zhang (Hong Kong University of Science and Technology)

Submissio	Title	Abstract	Authors with Affiliations
53	Levioso: Efficient Compiler-Informed Secure Speculation	Spectre-type attacks have demonstrated a major class of vulnerabilities arising from speculative execution of instructions, the main performance enabler of modern CPUs. These attacks speculatively leak secrets that have been either speculatively loaded (seen in sandboxed programs) or non-speculatively loaded (seen in constant-time programs). Various hardware-only defenses have been proposed to mitigate both speculative and non-speculative secrets via all potential transmission channels. However, these solution rely on limited knowledge of the hardware about the program to conservatively restrict the execution of all instructions that can potentially leak information. In this work, we discuss that not all instructions depend on older unresolved branches and they can safely execute without leaking speculative information. We present Levioso, a novel hardware/software co-design, that provides comprehensive secure speculation guarantees while reducing performance overhead compared to the existing methodologies. Levioso informs the hardware about true branch dependencies in order to apply restrictions only when necessary. Our evaluations demonstrate that Levioso is able to significantly reduce the performance overhead compared to two state-of-the-art defenses from 51% and 43% to just 23%.	Ali Hajiabadi (National University of Singapore); Archit Agarwal (University of California San Diego); Andreas Diavastos (National University of Singapore); Trevor E. Carlson (National University of Singapore)
54	Conjuring: Leaking Control Flow via Speculative Fetch Attacks	In this work, we propose a new attack called Conjuring that exploits one of the main features of CPUs' front-end: speculative fetch of instructions. We show that the Pattern History Table (PHT) in modern CPUs are a great channel to learn and leak the control-flow of victim applications. Unlike prior work, Conjuring does not require to prime the PHT or interfere with the victim execution enabling a realistic and unprivileged attacker to leak control flow information. By improving the branch predictors, our attack becomes even more serious and practical. We demonstrate the feasibility of our attack on different existing Intel, AMD, and Apple CPUs.	Ali Hajiabadi (National University of Singapore); Trevor E. Carlson (National University of Singapore)
55	INSPIRE: Accelerating Deep Neural Networks via Hardware-friendly Index-Pair Encoding	Deep Neural Network (DNN) inference consumes significant computing resources and development efforts due to the growing model size. Quantization is a promising technique to reduce the computation and memory cost of DNNs. Most existing quantization methods rely on fixed-point integers or floating-point types, which require more bits to maintain model accuracy. In contrast, variable-length quantization, which combines high precision for values with significant magnitudes (i.e., outliers) and low precision for normal values, offers algorithmic advantages but introduces significant hardware overhead due to variable-length encoding and decoding. Also, existing quantization methods are less effective for both (dynamic) activations and (static) weights due to the presence of outliers. In this work, we propose INSPIRE, an algorithm/architecture co-designed solution that employs an Index-Pair (INP) quantization and handles outliers globally with low hardware overheads and high performance gains. The key insight of INSPIRE lies in identifying typical features associated with important values, encoding them as indexes, and precomputing corresponding results for efficient storage in lookup table. During inference, the results of inputs with paired index can be directly retrieved from the table, which eliminates the need for any computational overhead. Furthermore, we design a unified processing element architecture for INSPIRE and highlight its seamless integration with existing DNN accelerators. As a result, INSPIRE-based accelerator surpasses the state-of-the-art quantization accelerators with a remarkable $9.31\times$ speedup and 81.3% energy reduction, respectively, while maintaining superior model accuracy.	Fangxin Liu (Shanghai Jiaotong University); Ning Yang (Shanghai Jiao Tong University); Zhiyan Song (Shanghai Qizhi Institute (SQI)); Zongwu Wang (Shanghai Jiaotong University); Haomin Li (Shanghai Jiao Tong University); Shiyuan Huang (Shanghai Jiao Tong University); Zhuoran Song (Shanghai Jiao Tong University); Songwen Pei (University of Shanghai for Science and Technology); Li Jiang (Shanghai Jiao Tong University)
63	Ink: Efficient Incremental k-Critical Path Generation	Critical Path Generation (CPG) is crucial for static timing analysis applications to validate timing constraints. Recent years have witnessed CPG algorithms that rank critical paths efficiently and accurately. However, they all lack incrementality, which is the ability to quickly update critical paths after the circuit is incrementally modified. To solve this, we introduce Ink, an efficient incremental CPG algorithm. Ink identifies reusable paths for the next query and effectively prunes the path search space. Ink is up to 22.4× faster and consumes up to 31% less memory than a state-of-the-art timer when generating one million paths on a large design.	Che Chang (University of Wisconsin, Madison); Tsung-Wei Huang (University of Wisconsin at Madison); Dian-Lun Lin (University of Wisconsin-Madison); Guannan Guo (UIUC); Shiju Lin (The Chinese University of Hong Kong)
82	EPIM: Efficient Processing-In-Memory Accelerators based on Epitome	The exploration of Processing-In-Memory (PIM) accelerators has garnered significant attention within the research community. However, the utilization of large-scale neural networks on Processing-In-Memory (PIM) accelerators encounters challenges due to constrained on-chip memory capacity. To tackle this issue, current works explore model compression algorithms to reduce the size of Convolutional Neural Networks (CNNs). Most of these algorithms either aim to represent neural operators with reduced-size parameters (e.g., quantization) or search for the best combinations of neural operators (e.g., neural architecture search). Designing neural operators to align with PIM accelerators' specifications is an area that warrants further study. In this paper, we introduce the Epitome, a lightweight neural operator offering convolution-like functionality, to craft memory-efficient CNN operators for PIM accelerators (EPIM). On the software side, we evaluate epitomes' latency and energy on PIM accelerators and introduce a PIM-aware layer-wise design method to enhance their hardware efficiency. We apply epitome-aware quantization to further reduce the size of epitomes. On the hardware side, we modify the datapath of current PIM accelerators to accommodate epitomes and implement a feature map reuse technique to reduce computation cost. Experimental results reveal that our 3-bit quantized EPIM-ResNet50 attains 71.59% top-1 accuracy on ImageNet, reducing crossbar areas by 30.65×. EPIM surpasses the state-of-the-art pruning methods on PIM	Chenyu Wang (Princeton University); Zhen Dong (UC Berkeley); Daquan Zhou (Bytedance); Zhenhua Zhu (Tsinghua University); Yu Wang (Tsinghua University); Jiashi Feng (ByteDance); Kurt Keutzer (University of California, Berkeley)
89	G2PM: Performance Modeling for ACAP Architecture with Dual-Tiered Graph Representation Learning	Performance estimation is a crucial component in the optimization processes of accelerator development on the Versal ACAP architecture. However, existing approaches present limitations - they are either too slow to facilitate efficient iterations, or they lack the necessary accuracy due to the specific AIE array architecture and two-level programming model of Versal ACAP. To tackle this challenge, we propose G^{2}PM, a performance modeling technique based on a hierarchical graph representation centered on the AIE array. More specifically, we employ a hierarchical graph neural network to identify features of both kernel programs and dataflow programs, taking into account the hardware and software characteristics of the Versal ACAP architecture. In our evaluations, our method demonstrates significant improvements, achieving a mean error rate of less than 1.6\% and providing a speed-up factor of 4165$\times$ compared to the simulation-based method.	Tuo Dai (Peking University); Bizhao Shi (Peking University); Guojie Luo (Peking University)

Submissio	Title	Abstract	Authors with Affiliations
90	PT-Map: Efficient Program Transformation Optimization for CGRA Mapping	Coarse-Grained Reconfigurable Array (CGRA) is a parallel architecture providing high energy efficiency and spatial-temporal reconfigurability. Beyond loop scheduling for throughput optimization, program transformation is also crucial in CGRA mapping to optimize overall performance and efficiency. However, existing studies on program transformation optimization face challenges in exploring the transformation space systematically and evaluating candidates efficiently, leading to sub-optimal results. To tackle these challenges, this paper introduces PT-Map, an efficient program transformation optimization framework for CGRA mapping. PT-Map defines a comprehensive transformation space and employs a CGRA-specialized top-down exploration approach. It also incorporates a bottom-up evaluation scheme using architectural parameters and a graph neural network-based predictive model. Experiments demonstrate that PT-Map achieves up to 2.95x/1.80x speedups and 59.0%/23.2% energy-delay-product (EDP) reductions over the state-of-the-art approaches MapZero and PBP, respectively.	Bizhao Shi (Peking University); Tuo Dai (Peking University); Jiaxi Zhang (Peking University); Xuechao Wei (Peking University); Guojie Luo (Peking University)
91	HiLight: A Comprehensive Framework For High Performance And Light-Weight Scalability In Surface Code Communication	In this paper, we introduce HiLight, an optimization framework designed for enhancing SC communication. HiLight integrates qubit-mapping strategies with program- and hardware-level optimizations, providing high-performance and lightweight scalable solutions. Featuring SWAP-less initial placement, HiLight utilizes qubit-proximity and pattern matching to minimize path congestion. In its routing strategy, HiLight employs fast gate-ordering and braiding path-finding to maximize gate parallelism and expedite optimal path selection. The combined optimizations improve latency and resource utilization. Compared with the state-of-the-art approach, HiLight achieves a remarkable reduction in latency and runtime by 43.5% and 91.9%, respectively, signifying its potential to advance the FTQC era.	Sunghye Park (Pohang University of Science and Technology); Dohun Kim (Pohang University of Science and Technology); Seokhyeong Kang (Pohang University of Science and Technology)
110	CFTCG: Test Case Generation for Simulink Model through Code Based Fuzzing	Simulink is extensively utilized in system design for its ability to facilitate modeling and synthesis of embedded controllers. It provides automatic test case generation to assist testers in inspecting the model. However, with the continuous increase in the model's scale, the control logic and internal states of the model are becoming more and more complex. Mainstream test case generation methods based on constraint solving and model simulation face challenges in achieving high coverage metrics. In this paper, we propose CFTCG, a fuzzing based test case generation method for Simulink models. First, CFTCG generates the fuzzing code, which includes the fuzz driver based on the model's input information and the fuzz code with model-level branch instrumentation. These codes are then compiled together to execute the model oriented fuzzing loop. During this fuzzing loop, we make use of the field information of the model inports and the coverage difference between iterative executions, allowing for more targeted input mutation. We evaluated CFTCG on several benchmark Simulink models. In comparison to the built-in Simulink Design Verifier and the state-of-the-art academic work SimCoTest, CFTCG demonstrates an average improvement of 47.2% and 100.8% on Decision Coverage, 38.3% and 44.6% on Condition Coverage, and 144.5% and 232.4% on Modified Condition Decision Coverage, respectively.	Zhuo Su (Tsinghua University); Zehong Yu (Tsinghua University); Dongyan Wang (Renmin University of China); Rui Wang (Capital Normal University); Yang Tao (HUAWEI Technologies, Co. LTD); Yu Jiang (Tsinghua University)
111	Top-Level Routing for Multiply-Instantiated Blocks with Topology Hashing	Modern System-on-Chip (SoC) design is divided into hierarchical instances using the multiply-instantiated block (MIB) technique to simplify the design process. Top-level routing aims at providing routing prototyping between those instances. It requires consideration of replicated routing paths that can either be utilized for routing or remain as floating segments. Conventional path-searching based algorithm often fails to find a legal solution under such a scenario. To address this, we propose an effective and efficient top-level routing framework for MIBs by hashing the topology of each net and using a group maze routing scheme. Experimental results demonstrate promising performance compared to the winners of the MIB-aware top-level router contest 2022 organized by Synopsys.	Jiarui Wang (Peking University); Xun Jiang (Peking University); Yibo Lin (Peking University)
119	Boolean Matching Reversible Circuits: Algorithm and Complexity	Boolean matching is an important problem in logic synthesis and verification. Despite being well-studied for conventional Boolean circuits, its treatment for reversible logic circuits remains largely, if not completely, missing. This work provides the first such study. Given two (black-box) reversible logic circuits that are promised to be matchable, we check their equivalences under various input/output negation and permutation conditions subject to the availability/unavailability of their inverse circuits. Notably, among other results, we show that the equivalence up to input negation and permutation is solvable in quantum polynomial time, while the classical complexity is exponential. This result is arguably the first demonstration of quantum exponential speedup in solving design automation problems. Also, as a negative result, we show that the equivalence up to both input and output negations is not solvable in quantum polynomial time unless UNIQUE-SAT is, which is unlikely. This work paves the theoretical foundation of Boolean matching reversible circuits for potential applications, e.g., in quantum circuit synthesis.	Tian-Fu Chen (Graduate School of Advanced Technology, National Taiwan University); Jie-Hong Roland Jiang (National Taiwan University)
126	Voronoi Diagram-based Multiple Power Plane Generation on Redistribution Layers in 3D ICs	In three-dimensional integrated circuits, the interconnection design among chiplets on redistribution layers (RDLs) is crucial for achieving high-performance computing systems. To optimize the inter-chip connections, most of the previous works focused on automatic signal net routing and pin assignment. The power net routing, or the power plane generation, is still a manual and time-consuming task, especially when generating the power planes of more than ten power supplies on a limited number of RDLs. This paper proposes a novel Voronoi diagram-based multiple power plane generation methodology which simultaneously optimizes the power planes of all power nets by utilizing the white space of given RDLs, while considering the signal routing blockages, power integrity, and complex design rules. Experimental results show that the proposed approach can achieve not only optimal area utilization but also the best power integrity in terms of the total number of redundant vias.	Chia-Wei Lin (National Yang Ming Chiao Tung University); Jing-Yao Weng (National Yang Ming Chiao Tung University); I-Te Lin (Synopsys, Inc.); Ho-Chieh Hsu (Synopsys, Inc.); Chia-Ming Liu (Synopsys, Inc.); Mark Po-Hung Lin (National Yang Ming Chiao Tung University)
131	Efficient Code Generation for Data-Intensive Simulink Models via Redundancy Elimination	Simulink has emerged as the fundamental infrastructure that supports modeling, simulation, verification, and code generation for embedded software development. To improve the performance of the code generated from Simulink models, state-of-the-art code generators employ various optimization techniques, such as expression folding, variable reuse, and parallelism. However, they overlook the presence of redundant calculations within data-intensive models widely used to perform substantial data processing in embedded scenarios, which can significantly undermine the efficiency and performance of the generated code. This paper proposes Frodo, an efficient code generator for data-intensive Simulink models via redundancy elimination. Frodo first conducts model analysis to construct the dataflow graph and derive the I/O mapping of each block. Then, for each block within the dataflow graph, Frodo recursively determines its calculation range by leveraging the I/O mapping of its subsequent blocks. After that, Frodo generates concise code for optimizable blocks in accordance with the precise calculation range. We implemented and evaluated Frodo on benchmark Simulink models. Compared with the state-of-the-art code generators Simulink Embedded Coder, DFSynth, and HCG, the code generated by Frodo is 1.17x - 8.55x faster in terms of execution duration across different compilers and architectures, without incurring additional overhead of memory usage.	Zehong Yu (Tsinghua University); Zhuo Su (Tsinghua University); Yu Jiang (Tsinghua University); Aiguo Cui (Huawei); Rui Wang (Capital Normal University)

Submissio	Title	Abstract	Authors with Affiliations
136	AccMoS: Accelerating Model Simulation for Simulink via Code Generation	Simulink has been widely used in embedded software development, which supports simulation to validate the correctness of the constructed models. However, as the scale and complexity of models in industrial applications grows, it is time-consuming for the simulation engine of Simulink to achieve high coverage and detect potential errors, especially accumulative errors. In this paper, we propose AccMoS, an accelerating model simulation method for Simulink models via code generation. AccMoS generates simulation functionality code for Simulink models through simulation oriented instrumentation, including runtime actor information collection, coverage collection, and calculation diagnosis. The final simulation code is constructed by composing all the instrumentation code with actor code generated from a predefined template library and integrating test data import. After compiling and executing the code, AccMoS generates simulation results that include coverage and diagnostic information. We implemented AccMoS and evaluated it on several benchmark Simulink models. Compared to Simulink's simulation engine, AccMoS shows a 215.3× improvement in simulation efficiency, significantly reduces the time required for detecting errors. AccMoS also achieved greater coverage within equivalent time.	Yifan Cheng (University of Electronic Science and Technology of China); Zehong Yu (Tsinghua University); Zhuo Su (Tsinghua University); Ting Chen (University of Electronic Science and Technology of China); Xiaosong Zhang (University of Electronic Science and Technology of China); Yu Jiang (Tsinghua University)
140	CAP: A General Purpose Computation-in-memory with Content Addressable Processing Paradigm	Demands for efficient computing under memory wall have led to computation-in-memory (CIM) accelerators that leverages memory structure to perform in-situ computing. The content addressable memory (CAM) processing is a CIM paradigm that accomplishes general purpose functions, via sequences of search and update op- eration. However, the conventional CAM-based CIM is customized for inter-vector operation and require long search-update iterations for computing. To mitigate the drawback of prior works, this work proposes a content addressable processor (CAP), improving both of the func- tionality and performance. CAP supports general purpose inter- vector and intra-vector operation. Respectively, CAP shortens the search and update step latency. The sequence order of search-update pair is released by CAP to achieve parallel search-update. CAP is implemented in 22nm CMOS technology with 0.6 mm2 area. By integrating all the techniques, CAP achieves 2.68 x performance improvement over the baseline, also realizes 11.37 TOPs/W energy efficiency and 1.376 TOP/mm2 area efficiency.	Zhiheng Yue (Tsinghua University); Shaojun Wei (Tsinghua University); Yang Hu (Tsinghua University); Shouyi Yin (Tsinghua University)
150	Zeroth-Order Optimization of Optical Neural Networks with Linear Combination Natural Gradient and Calibrated Model	Optical neural networks (ONNs) have attracted great attention due to their low energy consumption and high-speed processing. The usual neural network training scheme leads to poor performance for ONNs because of their special parameterization and fabrication variations. This paper contributes to extend zeroth-order (ZO) optimization, which can be used to train such ONNs, in two ways. The first is to propose linear combination natural gradient, which mitigates the optimization difficulty caused by the special parameterization of an ONN. The second is to generate a guided direction vector by calibration for better guessing than random vectors generated in ZO optimization. Experimental results show that the two extensions significantly outperformed the existing ZO optimization and related methods with little computational overhead.	Hiroshi Sawada (NTT Corporation); Kazuo Aoyama (NTT Corporation); Kohei Ikeda (NTT Corporation)
151	DEFA: Efficient Deformable Attention Acceleration via Pruning-Assisted Grid-Sampling and Multi-Scale Parallel Processing	Multi-scale deformable attention (MSDeformAttn) has emerged as a key mechanism in various vision tasks, demonstrating explicit superiority attributed to multi-scale grid-sampling. However, this newly introduced operator incurs irregular data access and enormous memory requirement, leading to severe PE under-utilization. Meanwhile, existing approaches for attention acceleration cannot be directly applied to MSDeformAttn due to lack of support for this distinct procedure. Therefore, we propose a dedicated algorithm-architecture co-design dubbed DEFA, the first-of-its-kind method for MSDeformAttn acceleration. At the algorithm level, DEFA adopts frequency-weighted pruning and probability-aware pruning for feature maps and sampling points respectively, alleviating the memory footprint by over 80%. At the architecture level, it explores the multi-scale parallelism to boost the throughput significantly and further reduces the memory access via fine-grained layer fusion and feature map reusing. Extensively evaluated on representative benchmarks, DEFA achieves 10.1-31.9× speedup and 20.3-37.7× energy efficiency boost compared to powerful GPU platforms. It also rivals the related accelerators by 2.2-3.7× energy efficiency improvement while providing pioneering support of MSDeformAttn.	Yansong Xu (Shanghai Jiao Tong University); Dongxu Lyu (Shanghai Jiao Tong University); Zhenyu Li (Shanghai Jiao Tong University); Yuzhou Chen (Shanghai Jiao Tong University); Zilong Wang (Shanghai Jiao Tong University); Gang Wang (Shanghai Jiao Tong University); Zhican Wang (Shanghai Jiao Tong University); Haomin Li (Shanghai Jiao Tong University); Guanghui He (Shanghai Jiao Tong University)
152	Deep Reorganization: Retaining Residuals in TinyML	Designing intelligent, tiny devices with limited memory is immensely challenging, exacerbated by the additional memory requirement of residual connections in deep neural networks. In contrast to existing approaches that eliminate residuals to reduce peak memory usage at the cost of significant accuracy degradation, this paper presents DERO, which reorganizes residual connections by leveraging insights into the types and interdependencies of operations across residual connections. Evaluations were conducted across diverse model architectures designed for common computer vision applications. DERO consistently achieves peak memory usage comparable to plain-style models without residuals, while maintaining the accuracy of the original models with residuals.	Hashan Roshantha Mendis (Academia Sinica); Chih-Kai Kang (Academia Sinica); Chun-Han Lin (National Taiwan Normal University); Ming-Syan Chen (National Taiwan University); Pi-Cheng Hsiu (Academia Sinica)
157	Alchemist: A Unified Accelerator Architecture for Cross-Scheme Fully Homomorphic Encryption	The use of cross-scheme fully homomorphic encryption (FHE) in privacy-preserving applications challenges hardware accelerator design. Existing accelerator architectures fail to efficiently handle hybrid FHE schemes due to the mismatch between computational demands and hardware resources. We propose a novel architecture using a hardware-friendly, versatile low-level operator, i.e., Meta-OP. Our slot-based data management efficiently handles memory access patterns of the meta-op for diverse operations. Alchemist accelerates both arithmetic and logic FHE with high hardware utilization rates. Compared to existing ASIC accelerators, Alchemist outperforms with a 29.4× performance per area improvement for arithmetic FHE and a 7.0× overall speedup for logic FHE.	Jianan Mu (ICT, CAS); Husheng Han (ICT, CAS); Shangyi Shi (ICT, CAS); Jing Ye (ICT, CAS); Zizhen Liu (ICT, CAS); Shengwen Liang (ICT, CAS); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University); Mingzhe Zhang (IIE, CAS); Song Bian (BUAA); Xing Hu (ICT, CAS); Huaiwei Li (ICT, CAS); Xiaowei Li (ICT, CAS)
158	STAGGER: Enabling All-in-One Subarray Sensing for Efficient Module-level Processing in Open-Bitline ReRAM	Emerging resistive RAM (ReRAM) devices can in-situ execute vector-matrix-multiplication (VMM) for scientific computing. However, the peripheral separated S&Hs and ADCs for row buffering and sensing in conventional designs are the system bottleneck. We propose an ADC-less all-in-one subarray-VMM-sensing design that enables the precharge once, readout multiple-bits functionality. We propose a cascaded-feedback bitline sensing architecture and a buffering-and-sensing-collocated sense amplifier design with bitline and storage node fully decoupled for enabling conflict-free column accesses. We further propose cross-level interleaving for successive VMM accesses. Experimental results show that our design achieves 297% performance improvement and 85.8% energy reduction, compared with an aggressive baseline.	Chengning Wang (Huazhong University of Science and Technology); Dan Feng (Huazhong University of Science and Technology); Yuchong Hu (Huazhong University of Science and Technology); Wei Tong (Huazhong University of Science and Technology); Jingning Liu (Huazhong University of Science and Technology)
164	ThermalScope: A Practical Interrupt Side Channel Attack Based on Thermal Event Interrupts	While interrupts play a critical role in modern OSes, they have been exploited as a wide range of side channel attacks to break system confidentiality, such as keystroke interrupts, graphic interrupts and network interrupts. In this paper, we propose ThermalScope, a new side channel that exploits thermal event interrupts, which is adaptable for both native and browser scenarios and incorporates two heat amplifying techniques. The exploited thermal event interrupts are activated only when the CPU package temperature reaches a fixed threshold that is determined by manufacturers. Our key observation is that workloads running on CPUs inevitably generates their distinct heat, which can be correlated with the thermal event interrupts. To demonstrate the viability of ThermalScope, we conduct a comprehensive evaluation on multiple Ubuntu OSes with different Intel-based CPUs. First, we show that the activation of thermal event interrupts correlates with the level of CPU temperature. We then apply ThermalScope to mount different side channel attacks, i.e., building covert channels with a transmission rate of 0.1 b/s, fingerprinting DNN model architectures with an accuracy of over 90% and breaking KASLR within 8.2 hours.	Xin Zhang (Peking University); Zhi Zhang (University of Western Australia); Qingni Shen (Peking University); Wenhao Wang (Institute of Information Engineering, CAS); Yansong Gao (Data61, CSIRO); Zhuoxi Yang (Peking University); Zhonghai Wu (Peking University)

Submissio	Title	Abstract	Authors with Affiliations
175	Enabling On-Device Self-Supervised LLM Personalization with Selective Synthetic Data	After a large language model (LLM) is deployed on edge devices, it is desirable for these devices to learn from user-generated conversation data to generate user-specific and personalized responses in real-time. However, user-generated data usually contains sensitive and private information, and uploading such data to the cloud for annotation is not preferred if not prohibited. While it is possible to obtain annotation locally by directly asking users to provide preferred responses, such annotations have to be sparse to not affect user experience. In addition, the storage of edge devices is usually too limited to enable large-scale fine-tuning with full user-generated data. It remains an open question how to enable on-device LLM personalization, considering sparse annotation and limited on-device storage. In this paper, we propose a novel framework to select and store the most representative data online in a self-supervised way. Such data has a small memory footprint and allows infrequent requests of user annotations for further fine-tuning. To enhance fine-tuning quality, multiple semantically similar pairs of question texts and expected responses are generated using the LLM. Our experiments show that the proposed framework achieves the best user-specific content-generating capability (accuracy) and fine-tuning speed (performance) compared with vanilla baselines. To the best of our knowledge, this is the very first on-device LLM personalization framework.	Ruiyang Qin (University of Notre Dame); Jun Xia (University of Notre Dame); Zhenge Jia (University of Notre Dame); Meng Jiang (University of Notre Dame); Ahmed Abbasi (University of Notre Dame); Peipei Zhou (University of Pittsburgh); Jingtong Hu (University of Pittsburgh); Yiyu Shi (University of Notre Dame)
182	Synthesis of Compact Flow-based Computing Circuits from Boolean Expressions	Processing in-memory has the potential to accelerate high-data-rate applications beyond the limits of modern hardware. Flow-based computing is a computing paradigm for executing Boolean logic within nanoscale memory arrays by leveraging the natural flow of electric current. Previous approaches of mapping Boolean logic onto flow-based computing circuits have been constrained by their reliance on binary decision diagrams (BDDs), which translates into high area overhead. In this paper, we introduce a novel framework called FACTOR for mapping logic functions into dense flow-based computing circuits. The proposed methodology introduces Boolean connectivity graphs (BCGs) as a more versatile representation, capable of producing smaller crossbar circuits. The framework constructs concise BCGs using factorization and expression trees. Next, the BCGs are modified to be amenable for mapping to crossbar hardware. We also propose a time multiplexing strategy for sharing hardware between different Boolean functions. Compared with the state-of-the-art approach, the experimental evaluation using 14 circuits demonstrates that FACTOR reduces area, speed, and energy with 80%, 2%, and 12%, respectively, compared with the state-of-the-art synthesis method for flow-based computing.	Sven Thijssen (University of Central Florida); Muhammad Rashedul Haq Rashed (University of Central Florida); Sumit K. Jha (Florida International University); Rickard Ewetz (University of Central Florida)
184	Oltron: Algorithm-Hardware Co-design for Outlier-Aware Quantization of LLMs with Inter-/Intra-Layer Adaptation	The recent breakthroughs in the field of large language models (LLMs) owe much of their accomplishments to the exponential growth in model size (240×every two years), creating a significant challenge in computation and memory complexity for today's hardware. Quantization has emerged as a critical technique for reducing these complexities. However, existing approaches mainly employ a fixed quantization schemes, which is in-efficient in terms of requiring more bits to maintain model accuracy. In this work, we delve into the dynamics and heterogeneity present in both inter- and intra-layer distributions, particularly focusing on the highly dynamic range and compositions of the extremely large values, commonly referred to as outliers. We propose Oltron, an algorithm/hardware co-design solution for outlier-aware quantization of LLMs with inter-/intra-layer adaptation. Oltron employs a holistic quantization framework with three key innovations. First, we propose a novel quantization algorithm capable of determining the optimal composition ratio of outliers among different layers and various channel groups within a layer. Second, we propose a reconfigurable architecture that can adjust computation fabric based on inter- and intra-layer distributions. Third, we propose a tile-based dataflow optimizer to meticulously plan the complicated computation and memory access schedule for the mix-precision tensors. Oltron is demonstrated to surpass existing outlier-aware accelerator, OliVe, by 1.9× performance improvement and 1.6× energy efficiency improvement, with a superior model accuracy.	Chenhao Xue (School of Integrated Circuits, Peking University); Chen Zhang (Shanghai Jiao Tong University); Xun Jiang (Peking University); Gao ZhuTianya (SHANGHAI JIAO TONG UNIVERSITY); Yibo Lin (Peking University); Guangyu Sun (Peking University)
203	SymPhase: Phase Symbolization for Fast Simulation of Stabilizer Circuits	This paper proposes an efficient stabilizer circuit simulation algorithm that only traverses the circuit forward once. We introduce phase symbolization into stabilizer generators, which allows possible Pauli faults in the circuit to be accumulated explicitly as symbolic expressions in the phases of stabilizer generators. This way, the measurement outcomes are also symbolic expressions, and we can sample them by substituting the symbolic variables with concrete values, without traversing the circuit repeatedly. We show how to integrate symbolic phases into the stabilizer tableau and maintain them efficiently using bit-vector encoding. A new data layout of the stabilizer tableau in memory is proposed, which improves the performance of our algorithm (and other stabilizer simulation algorithms based on the stabilizer tableau). We implement our algorithm and data layout in a Julia package named SymPhase.jl, and compare it with Stim, the state-of-the-art simulator, on several benchmarks. We show that SymPhase.jl has superior performance in terms of sampling time, which is crucial for generating a large number of samples for further analysis.	Wang Fang (Institute of Software, Chinese Academy of Sciences); Mingsheng Ying (Institute of Software, Chinese Academy of Sciences)
215	Execution Sequence Optimization for Processing In-Memory using Parallel Data Preparation	Computation using Processing in-memory (PIM) is performed by breaking down computationally expensive operations into in-memory kernels that can be efficiently executed using non-volatile memory. Logic styles such as MAGIC requires that each output memory cell is prepared for evaluation before executing the functional logic operation. State-of-the-art synthesis algorithms perform the preparation immediately after memory cells have expired. Unfortunately, this results in that columns of cells are prepared one-by-one, instead of leveraging efficient parallel data preparation instructions. In this paper, we propose the PREP framework that maximizes the opportunities for parallel column preparation using execution sequence optimization.	Muhammad Rashedul Haq Rashed (University of Central Florida); Sven Thijssen (University of Central Florida); Dominic B. Simon (University of Central Florida); Sumit K. Jha (Florida International University); Rickard Ewetz (University of Central Florida)
220	Compact and Efficient CAM Architecture through Combinatorial Encoding and Self-Terminating Searching for In-Memory-Searching Accelerator	Content addressable memory (CAM) has triggered a lot of attention for data-intensive applications due to highly parallel pattern searching capability. Most state-of-the-art works focus on reducing hardware cost of CAM by exploiting various emerging non-volatile memory (NVM) technologies. However, existing CAM designs still mainly follow the conventional encoding scheme which requires two complementary storage nodes and search signals for each bit of entry and query respectively, along with separate precharging and evaluation phases for bit-vector searching, limiting the further improvement of area- and energy-efficiency. In this work, a compact and efficient CAM architecture is proposed through two techniques: (1) a combinatorial encoding scheme for CAM by encoding entry/query states with permutations and combinations of multiple storage nodes as a group, which can significantly improve the encoding efficiency and thus greatly reduce the hardware implementation cost of CAM compared with conventional encoding scheme; (2) an one-step self-terminating searching scheme for CAM by detecting matching condition during precharging phase and terminating precharging once a match is detected, which can further reduce the search delay and energy. The experiments and evaluations of the proposed CAM architecture with co-optimization of combinatorial encoding and self-terminating searching are carried out based on ferroelectric FET (FeFET), which can reduce the area-energy-delay product (AEDP) by 1182× over the conventional CMOS-based CAM in data searching tasks, showing its great potential for area- and energy-efficient in-memory-searching accelerator.	Weikai Xu (Peking university); Jin Luo (Peking university); Qianqian Huang (Peking university); Ru Huang (Peking university)

Submissio	Title	Abstract	Authors with Affiliations
223	Dyn-Bitpool: A Two-sided Sparse CIM Accelerator Featuring a Balanced Workload Scheme and High CIM Macro Utilization	Computing-in-memory has demonstrated great energy-efficiency by integrating computing units into memory. However, previous research on CIM has rarely utilized sparsity in activation and weight concurrently. Thus, we implemented an accelerator called Dyn-Bitpool which innovates on two fronts: 1) a balanced working scheme called "pool first and cross lane sharing" to maximize the performance benefiting from bit-level sparsity in activation; 2) dynamic topology of CIM arrays to effectively handle low hardware utilization issue stemming from value-level sparsity in weight. All the contributions collaborate to speed up Dyn-Bitpool by 1.89x and 2.64x on average compared with two state-of-the-art accelerators featuring CIM.	Xujiang Xiang (Tsinghua University); Zhiheng Yue (Tsinghua University); Yuxuan Li (Tsinghua University); Liuxin Lv (Tsinghua University); Shaojun Wei (Tsinghua University); Yang Hu (Tsinghua University); Shouyi Yin (Tsinghua University)
227	PowPrediCT: Cross-Stage Power Prediction with Circuit-Transformation-Aware Learning	Accurate and efficient power analysis at early VLSI design stages is critical for effective power optimization. It is a promising yet challenging task, especially during placement stage with the clock tree and final signal routing unavailable. Additionally, optimization-induced circuit transformations like circuit restructuring and gate sizing can invalidate fine-grained power supervision. Addressing these, we introduce the first generalizable circuit-transformation-aware power prediction model at placement stage. Compared to the cutting-edge commercial IC engine Innovus, we have significantly reduced the cross-stage power analysis error between placement and detailed routing.	Yufan Du (Peking University); Zizheng Guo (Peking University); Xun Jiang (Peking University); Zhuomin Chai (Wuhan University); Yuxiang Zhao (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)
238	A RRAM-based High Energy-efficient Accelerator Supporting Multimodal Tasks for Virtual Reality Wearable Devices	Virtual reality (VR) wearable devices can achieve immersive entertainment by fusing multi-modal tasks from various senses. However, constrained by the short battery life and limited hardware resources of VR devices, it is difficult to run multiple tasks simultaneously with different modals. Based on the above issues, we propose an energy-efficient accelerator that supports Multi-modal Tasks for VR devices, namely MTVR. We present a multi-task computing solution based on the flexible multi-task computing core design and efficient computing unit allocation strategy, which simultaneously achieves efficient work of multi-modal tasks. We have designed an early exit detector to skip invalid calculations, which saves energy greatly. In addition, a fine-grained tiny value skip method at multiplier and adder levels is proposed to save energy further. We provide a hybrid RRAM and SRAM memory access scheme, reducing the external memory access (EMA). Through experimental evaluation, the multi-task computing core achieves an average computational utilization of 95%. When the invalid input ratio is 90%, energy saving brought by the early exit detector can reach 88%. The tiny value skip method further achieved 13% energy saving. A hybrid memory access scheme obtains a 98.9% EMA reduction. We deployed the MTVR accelerator in FPGA and self-designed RRAM, achieving energy efficiency of 3.6 TOPS/W, higher than other single-task accelerators.	Xin ZHAO (University of Electronic Science and Technology of China); Zhicheng Hu (University of Electronic Science and Technology of China); Zilong Guo (University of Electronic Science and Technology of China); Haodong Fan (University of Electronic Science and Technology of China); Xi Yang (University of Electronic Science and Technology of China); Jing Zhou (University of Electronic Science and Technology of China); Liang Chang (University of Electronic Science and Technology of China)
243	Nona: Accurate Power Prediction Model Using Neural Networks	This paper proposes a neural-network-based power model, Nona, that accurately predicts the power consumption of heterogeneous CPUs on a commercial mobile device. With aggressive on-device power management in action, it becomes increasingly challenging to make accurate power predictions for diverse applications. To overcome the limitations of the existing power models based on linear regression, Nona uses a lightweight neural network with a small number of performance monitoring counters (PMCs) chosen from a system analysis and a loss function designed for power prediction. Experiments on Google Pixel 6 show that Nona has a 3.4% average prediction error, improving on prior work by 2.6x.	HoSun Choi (Yonsei university); Chanho Park (Yonsei university); Euijun Kim (Yonsei university); William Song (Yonsei University)
249	Artisan: Automated Operational Amplifier Design via Domain-specific Large Language Model	This paper presents Artisan, an automated operational amplifier design framework using large language models. We develop a bidirectional representation to align abstract circuit topologies with their structural and functional semantics. We further employ Tree-of-Thoughts and Chain-of-Thoughts approaches to model the design process as a hierarchical question-answer sequence, implemented by a mechanism of multi-agent interaction. A high-quality opamp dataset is developed to enhance the design proficiency of Artisan. Experimental results demonstrate that Artisan outperforms state-of-the-art optimization-based methods and benchmark LLMs, in success rate, circuit performance metrics, and interpretability, while accelerating the design process by up to 50.1x. Artisan will be released for public access.	Zihao Chen (Fudan University); Jiangli Huang (Fudan University); Yiting Liu (Fudan University); Fan Yang (Fudan University); Li Shang (fudan university); Dian Zhou (The University of Texas at Dallas); Xuan Zeng (Fudan University)
250	Advanced gate-level glitch modeling using ANNs	Multiple Input Switching (MIS) effects commonly induce undesired glitch pulses at the output of CMOS gates, potentially leading to circuit malfunction and significant power consumption. Thus, accurate and efficient glitch modeling is crucial for the design of high-performance, low-power, and reliable ICs. In this work, we present a new gate-level approach for modeling glitch effects under MIS. Unlike previous studies, we leverage efficient Machine Learning (ML) techniques to accurately estimate the glitch shape characteristics, propagation delay, and power consumption. To this end, we evaluate various ML engines and explore different Artificial Neural Network (ANN) architectures. Moreover, we introduce a seamless workflow to integrate our ANNs into existing standard cell libraries, striking an optimal balance between model size and accuracy in gate-level glitch modeling. Experimental evaluation on gates implemented in 7 nm FinFET technology demonstrates that the proposed models achieve an average error of 2.19% against SPICE simulation while maintaining a minimal memory footprint.	Anastasis Vagenas (University of Thessaly); Dimitrios Garyfallou (University of Thessaly); Nestor Evmorfopoulos (University of Thessaly); George Stamoulis (University of Thessaly)
252	PONO: Power Optimization with Near Optimal SMT-based Sub-circuit Generation	Generating high-quality sub-circuit for local substitution is an effective optimization technique in logic synthesis. There have been abundant works on generating area and delay optimal sub-circuits, greatly enhancing the logic optimization quality. However, power- oriented sub-circuit generation is rarely discussed, while optimizing power consumption in this sub-15 nm era is of paramount interest. We propose PONO, an SMT-based near optimal sub-circuit generation flow for power optimization. PONO enables power-oriented circuit library building and fills the gap in generating circuits near the Pareto frontier in PPA (Power, Performance, and Area). It manifests superiority in power reduction over traditional one in rewrite, a key logic optimization algorithm. We test PONO on EFPL benchmarks, and it shows 8.7% less power consumption with comparable performance and area after placement and routing.	Sunan Zou (School of Computer Science, Peking University); Guojie Luo (Peking University)
255	MoC: A Morton-Code-Based Fine-Grained Quantization for Accelerating Point Cloud Neural Networks	Point Cloud Neural Network (PCNN) plays an essential role in various 3D applications, with some of them even being time-sensitive and safety-critical. However, the large scale of unordered points with lengthy features results in heavy computational workloads, making them far from real-time processing. To address this challenge, we propose MoC, a Morton-code-based fine-grained quantization for accelerating PCNNs. Specifically, we utilize Morton code to capture the spatial locality among points. Then, we gather nearby points with similar features into a region. Considering the similarity in features of nearby points, we propose to decompose features into base and offsets, where the offsets fall within a narrow range. Building upon this, we introduce a two-level mixed-precision quantization. In the first level, we quantize offsets with low precision, while keeping the base in high precision to ensure accuracy. For the second level, noticing the different data distribution of offsets across various regions, we employ two types of low precision at the region level, which provides opportunities to further accelerate feature computations. To support our algorithm, we design a hardware architecture that parallelizes the Morton code path with the critical path. In our extensive experiments on various datasets, our algorithm-architecture co-designed method demonstrates 12x, 6.3x, 4.7x, 3.8x, 3.4x and 2.8x speedup and 19.3x, 9.7x, 6.0x, 5.2x, 4.6x and 4.1x energy savings over CPU, Server and Edge GPUs, state-of-the-art ASICs (incl. PointAcc, MARS, PRADA) with negligible accuracy loss.	Xueyuan Liu (Shanghai Jiao Tong University); Zhuoran Song (Shanghai Jiao Tong University); Hao Chen (Shanghai Jiao Tong University); Xing Li (Shanghai Jiao Tong University); Xiaoyao Liang (Shanghai Jiao Tong University)

Submissio	Title	Abstract	Authors with Affiliations
291	Co-Via: A Video Frame Interpolation Accelerator Exploiting Codec Information Reuse	Video Frame Interpolation(VFI) aims to generate intermediate frames between consecutive frames. Recent DNN-based VFI offers superior quality but suffers from performance issues. However, very few studies have focused on VFI hardware acceleration and existing work overlooks temporal information from compressed video bitstreams. In this paper, we propose a novel compressed VFI workflow and an accelerator, Co-Via. Co-Via exploits codec information reuse to reduce complex DNN computations and alleviate hardware pressure. FPGA-based Co-Via outperforms an RTX 4090 GPU 10.31x, offering a 43.08x energy efficiency boost. Its ASIC version achieves 2.4x higher throughput and 3.6x energy efficiency than the state-of-the-art solution.	Haishuang Fan (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences;University of Chinese Academy of Sciences); Qichu Sun (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences); Jingya Wu (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; YUSUR Technology Co., Ltd.); Wenyan Lu (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; YUSUR Technology Co., Ltd.); Xiaowei Li (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences); Guihai Yan (State Key Laboratory of Processors, Institute of Computing Technology, Chinese Academy of Sciences; YUSUR Technology Co., Ltd.)
293	DNN-Defender: A Victim-Focused In-DRAM Defense Mechanism for Taming Adversarial Weight Attack on DNNs	With deep learning deployed in many security-sensitive areas, machine learning security is becoming progressively important. Recent studies demonstrate attackers can exploit system-level techniques exploiting the RowHammer vulnerability of DRAM to deterministically and precisely flip bits in Deep Neural Networks (DNN) model weights to affect inference accuracy. The existing defense mechanisms are software-based, such as weight reconstruction requiring expensive training overhead or performance degradation. On the other hand, generic hardware-based victim-/aggressor-focused mechanisms impose expensive hardware overheads and preserve the spatial connection between victim and aggressor rows. In this paper, we present the first DRAM-based victim-focused defense mechanism tailored for quantized DNNs, named DNN-Defender that leverages the potential of in-DRAM swapping to withstand the targeted bit-flip attacks with a priority protection mechanism. Our results indicate that DNN-Defender can deliver a high level of protection downgrading the performance of targeted RowHammer attacks to a random attack level. In addition, the proposed defense has no accuracy drop on CIFAR-10 and ImageNet datasets without requiring any software training or incurring hardware overhead.	Ranyang Zhou (New Jersey Institute of Technology); Sabbir Ahmed (State University of New York at Binghamton); Adnan Siraj Rakin (Binghamton University); Shaahin Angizi (New Jersey Institute of Technology)
294	SMILE: LLC-based Shared Memory Expansion to Improve GPU Thread Level Parallelism	As the de facto high-throughput accelerators targeting at a wide spectrum of applications, graphics processing units (GPUs) keep adding computing and memory resources to meet the increasing demands. However, while designed for massive parallelism, GPUs are frequently suffering from low thread occupancy and limited data throughput, which are typically attributed to constrained on-chip resources, such as shared memory and register file. To alleviate the pressure, last-level cache (LLC) is being substantially enlarged to support continuously growing computation and to shrink the off-chip data traffic. Nevertheless, the frequent low usage of LLC leaves the space waste, impeding LLC from fully unleashing potentials. Towards the issue, we propose to manage partial LLC in a software way instead to expand precious shared memory, named as SMILE, helping to alleviate the low occupancy. SMILE splits the monolithic LLC into normal data cache and new software region, with the latter being to extend the limited SMEM. For adapting to diverse application characteristics, SMILE enables multiple splitting grades and meanwhile determines the appropriate partition through online profiling among streaming multiprocessors. Experimental results show that SMILE achieves average performance improvements of 14.7% and 8.4% respectively, compared to the default baseline and prior state-of-the-art.	Tianyu Guo (Sun Yat-sen University); Xuanteng Huang (Sun Yat-sen University); Kan Wu (Sun Yat-sen University); Xianwei Zhang (Sun Yat-sen University); Nong Xiao (Sun Yat-sen University)
303	LVF2: A Statistical Timing Model based on Gaussian Mixture for Yield Estimation and Speed Binning	As transistor size continues to scale down, process variation has become an essential factor determining semiconductor yield and economic return. The Liberty Variation Format (LVF) is the current industrial standard that expresses statistical timing behaviors based on single Gaussian model. However, it loses accuracy when the timing distribution is non-Gaussian due to growing process variations. This paper proposes a novel LVF2 distribution model to better capture the multi-Gaussian timing distribution while maintaining backward compatibility with LVF. The experiment using TSMC 22nm technology shows that compare to LVF, LVF2 reduces binning error of 7.74× in delay and 9.56× in transition, and reduces 3 -yield error of 4.79× in delay and 7.18× in transition. The error reduction is reduced for path delay due to Central Limit Theorem (CLT). But it is still 2× for a typical circuit path with 8 tiσmes Fanout-of-4 (FO4) inverter delays.	Junzhuo Zhou (UCLA); Li Huang (University of Nottingham Ningbo China); Haoxuan Xia (University of Nottingham Ningbo China); Yihui Cai (Southeast University); Leilei Jin (Southeast University); Xiao Shi (Southeast University); Wei W. Xing (The University of Sheffield); Ting-Jung Lin (Chiplet CAD and Manufacturing Research Center of Zhejiang Province, Ningbo Institute of Digital Twin, Eastern Institute of Technology); Lei He (Eastern Institute of Technology / UCLA)
306	PVTSizing: A TuRBO-RL-Based Batch-Sampling Optimization Framework for PVT-Robust Analog Circuit Synthesis	With the CMOS technology advancing and the complexity of circuits growing, the demand for analog/mixed-signal design automation tools is increasing quickly. Although some tools have been developed to tackle this challenge, the performance degradation caused by process, voltage, and temperature (PVT) variations has been less considered. This paper presents PVTSizing, an optimization framework for PVT-robust analog circuit synthesis. PVTSizing adopts trust region Bayesian optimization (TuRBO) for high-quality initial datasets and reference points. Multi-task reinforcement learning (RL) is utilized for PVT optimization. Both TuRBO and RL are batch-friendly, allowing parallel sampling of design solutions. Meanwhile, critic-assisted pruning and zoom target metrics are proposed to improve sample efficiency and reduce runtime. In addition, this framework naturally supports sizing over random mismatch. On 4 real-world circuits with TSMC 28/180nm process, PVTSizing achieves 1.9x-8.8x sample efficiency and 1.6x-9.8x time efficiency improvements compared to prior sizing tools from both industry and academia.	Zichen Kong (Peking University); Xiyuan Tang (Peking University); Wei Shi (University of Texas at Austin); Yiheng Du (Peking University); Yibo Lin (Peking University); Yuan Wang (Peking University)
310	Older and Wiser: The Marriage of Device Aging and Intellectual Property Protection of DNNs	Deep Neural Networks (DNNs), such as the widely-used ChatGPT model containing billions of parameters, are often kept secret due to the high training costs and privacy concerns surrounding the data used to train them. Previous approaches to securing DNNs typically require expensive circuit redesign, resulting in additional overheads such as increased area, energy consumption, and latency. To address these issues, we propose a novel hardware-software co-design for DNN protection that leverages the inherent aging characteristics of circuits to provide effect protection. Hardware-side, we employ random aging to produce authorized chips. This process circumvents the need for chip redesign, thereby eliminating any additional energy and area overhead. Moreover, the authorized chips demonstrate a considerable disparity in DNN inference performance when compared to unauthorized third-party chips. Software-side, we propose a novel Differential Orientation Fine-tuning method, which allows pre-trained DNNs to maintain its original accuracy on authorized chips with minimal fine-tuning, while the model's performance on unauthorized chips is reduced to random guessing. Comprehensive experiments on MLP, VGG, ResNet, Mixer and SwinTransformer validate the efficacy of our method.	Ning Lin (The University of Hong Kong); Shaocong Wang (The University of Hong Kong); Yue Zhang (The University of Hong Kong); Yangu He (The University of Hong Kong); Kwunhang Wong (The University of Hong Kong); Arindam Basu (City University of Hong Kong); Dashan Shang (Institute of Microelectronics, Chinese Academy of Sciences); Xiaoming Chen (Institute of Computing Technology, Chinese Academy of Sciences); Zhongrui Wang (The University of Hong Kong)

Submissio	Title	Abstract	Authors with Affiliations
311	Revisiting Automatic Pipelining: Gate-level Forwarding and Speculation	The key to pipeline throughput optimization is to resolve data hazards caused by read-after-write (RAW) dependencies, which are traditionally tackled by forwarding and speculation to avoid pipeline stalls. However, existing approaches are conducted based on high-level dataflow analysis, with potential loss of optimization opportunities for lack of analysis of the netlist structures. We propose an efficient method to resolve RAW dependencies with low-level netlist analysis by gate-level forwarding and speculation. With a greedy search method to detect and resolve short-delay gate-level signal paths for forwarding and an approximate circuit synthesis method with formal verification for gate-level speculation, the method efficiently utilizes the gate-level information to further improve pipeline throughput. We conduct experiments on the widely-used ISCAS/EPFL benchmark circuits and a large-scale RISC-V CPU. Experimental results show that our approach can increase the pipeline throughput. More importantly, our approach can find better designs than human experts.	Shuyao Cheng (Institute of Computing Technology, Chinese Academy of Sciences); Chongxiao Li (Institute of Computing Technology, Chinese Academy of Sciences); Zidong Du (Institute of Computing Technology, Chinese Academy of Sciences); Rui Zhang (ICT-CAS); Xing Hu (Institute of Computing Technology, Chinese Academy of Sciences); Xiaqing Li (Institute of Computing Technology, Chinese Academy of Sciences); Guanglin Xu (SKL of Processors, Institute of Computing Technology, CAS); Yuanbo Wen (Institute of Computing Technology, Chinese Academy of Sciences); Qi Guo (ICT/CAS)
313	Mixed-Dimensional Qudit State Preparation Using Edge-Weighted Decision Diagrams	Quantum computers have the potential to solve important problems which are fundamentally intractable on a classical computer. The underlying physics of quantum computing platforms supports using multi-valued logic, which promises a boost in performance over the prevailing two-level logic. One key element to exploiting this potential is the capability to efficiently prepare quantum states for multi-valued, or qudit, systems. Due to the time sensitivity of quantum computers, the circuits to prepare the required states have to be as short as possible. In this paper, we investigate quantum state preparation with a focus on mixed-dimensional systems, where the individual qudits may have different dimensionalities. The proposed approach automatically realizes quantum circuits constructing a corresponding mixed-dimensional quantum state. To this end, decision diagrams are used as a compact representation of the quantum state to be realized. We further incorporate the ability to approximate the quantum state to enable a finely controlled trade-off between accuracy, memory complexity, and number of operations in the circuit. Empirical evaluations demonstrate the effectiveness of the proposed approach in facilitating fast and scalable quantum state preparation, with performance directly linked to the size of the decision diagram. The implementation is freely available under the MIT license at redacted for double-blind submission.	Kevin Mato (Technische Universität München); Stefan Hillmich (Software Center Hagenberg (SCCH) GmbH); Robert Wille (Technical University of Munich)
328	Efficient Equivalence Checking of Nonlinear Analog Circuits using Gradient Ascent	In this paper, we present an optimized methodology for performing state-space-based equivalence checking of nonlinear analog circuits by using a gradient-ascent-based search algorithm to efficiently traverse a common state space. Essentially, the method searches for critical regions where the functional behaviors of two circuit designs show the greatest divergence. The key challenges in this approach are the mapping of both designs onto a common canonical state space, the computation of the gradient, and the exclusion of unreachable regions within the state space. To address the first challenge, we use locally linearized systems and leverage the Kronecker Canonical Form (KCF). To facilitate the computation of the gradient, we employ a purpose-built target function, and to exclude unreachable regions, we utilize vector projection techniques. Through experiments with nonlinear analog circuits and a scalability analysis, we demonstrate the successful and efficient computation performed with the proposed methodology, achieving speedups of up to 468 times.	Kemal Çağlar Coşkun (Institute of Computer Science, University of Bremen); Muhammad Hassan (Institute of Computer Science, University of Bremen); Lars Hedrich (Institute for Computer Science, Goethe University Frankfurt); Rolf Drechsler (Institute of Computer Science, University of Bremen)
331	MERSIT: A Hardware-Efficient 8-bit Data Format with Enhanced Post-Training Quantization DNN Accuracy	Post-training quantization (PTQ) models utilizing conventional 8-bit Integer or floating-point formats still exhibit significant accuracy drops in modern deep neural networks (DNNs), rendering them unreliable. This paper presents MERSIT, a novel 8-bit PTQ data format designed for various DNNs. While leveraging the dynamic configuration of exponent and fraction bits derived from Posit data format, MERSIT demonstrates enhanced hardware efficiency through the proposed merged decoding scheme. Our evaluation indicates that MERSIT yields more reliable 8-bit PTQ models, exhibiting superior accuracy across various DNNs compared to conventional floating-point formats.	Nguyen-Dong Ho (Kyunghee University); Gyujun Jeong (Kyunghee University); Cheol-Min Kang (Kyunghee University); Seungkyu Choi (Kyung Hee University); Ik Joon Chang (Kyunghee University)
339	Automatically Fixing RTL Syntax Errors with Large Language Model	This paper presents RTLFixer, a novel framework enabling automatic syntax errors fixing for Verilog code with Large Language Models (LLMs). Despite LLM's promising capabilities, our analysis indicates that approximately 55\% of errors in LLM-generated Verilog are syntax-related, leading to compilation failures. To tackle this issue, we introduce a novel debugging framework that employs Retrieval-Augmented Generation (RAG) and ReAct prompting, enabling LLMs to act as autonomous agents in interactively debugging the code with feedback. This framework demonstrates exceptional proficiency in resolving syntax errors, successfully correcting about 98.5\% of compilation errors in our debugging dataset, comprising 212 erroneous implementations derived from the VerilogEval benchmark. Our method leads to 32.3\% and 8.6\% increase in pass@1 success rates in the VerilogEval-Machine and VerilogEval-Human benchmarks, respectively.	YunDa Tsai (NVIDIA); Mingjie Liu (NVIDIA Corporation); Haoxing Ren (NVIDIA Corporation)
355	TITAN: A Fast and Distributed Large-Scale Trapped-Ion NISQ Computer	Trapped-Ion (TI) technology offers potential breakthroughs for Noisy Intermediate Scale Quantum (NISQ) computing. TI qubits provide advantages like extended coherence times and high gate fidelity, making them appealing for large-scale quantum systems. Constructing such systems demands a distributed architecture connecting Quantum Charge Coupled Devices (QCCDs) via quantum matter-links and photonic switches. However, current distributed TI NISQ computers face hardware and system challenges. Entangling qubits across a photonic switch introduces significant latency, impacting performance, while existing compilers generate suboptimal mappings and schedules. In response, we introduce TITAN, a large-scale distributed TI NISQ computer. TITAN employs an innovative photonic interconnection design to reduce entanglement latency and an advanced partitioning and mapping algorithm to optimize quantum matter-link communications. Our evaluations show that TITAN significantly enhances quantum application performance and fidelity compared to existing systems.	Cheng Chu (Indiana University Bloomington); Zhenxiao Fu (Indiana University Bloomington); Yilun Xu (Accelerator Technology and Applied Physics Division, Lawrence Berkeley National Laboratory); Gang Huang (Accelerator Technology and Applied Physics Division, Lawrence Berkeley National Laboratory); Hausi Muller (Department of Computer Science, University of Victoria); Fan Chen (Indiana University Bloomington); Lei Jiang (Indiana University Bloomington)
363	Duet: A Collaborative User Driven Recommendation System for Edge Devices	Recommendation systems are the backbone for numerous user applications on edge devices. However, the compute and memory-intensive nature of recommendation models renders them unsuitable for edge devices. Nevertheless, by decoupling the model fraction related to user history (e.g., past visited pages, liked posts) and user attributes (such as age, gender), we can offload partial recommendation models onto local edge devices. Hence, we present Duet, a novel collaborative edge-cloud recommendation system that intelligently decomposes the recommendation model into two smaller models – user and item models -- that execute simultaneously on the edge device and cloud before coming together to deliver final recommendations. Further, we propose a lightweight Duet architecture to support user models on resource-constrained edge devices. Overall, Duet reduces the average latency by 6.4x and improves energy efficiency by 4.6x across five recommendation models.	Vidushi Goyal (University of Michigan); Valeria Bertacco (University of Michigan); Reetuparna Das (University of Michigan)

Submissio	Title	Abstract	Authors with Affiliations
365	SpectraFlux: Harnessing the Flow of Multi-FPGA in Mass Spectrometry Clustering	The identification and quantification of proteins through mass spectrometry (MS) are foundational to proteomics, offering insights into biological systems and disease states. However, current clustering tools struggle to process large-scale datasets. We propose SpectraFlux, a multiple FPGA-based architecture for accelerated mass spectrum clustering that outperforms existing CPU, GPU, and FPGA designs. It employs heterogeneous clustering kernels for adaptive bucket size management and optimizes memory usage by distinguishing between on-chip and high-bandwidth memory (HBM) storage solutions. SpectraFlux is built upon the TAPA-CS framework, which automatically compiles and partitions a large dataflow design across multiple chips with RDMA-based inter-FGPA communication. Our solution shows 2.7× speed up on a quad-FPGA platform compared to a single FPGA. Additionally, we introduce a refined cost model for frame-based inter-FPGA communication to better accommodate the variable data rates inherent in proteomic data processing, which reduces the inter-FPGA data movement by up to 73%. Finally, SpectraFlux achieves speedups of up to 11× and 17× over SOTA FPGA and GPU accelerators, respectively。	Tianqi Zhang (UCSD); Neha Prakriya (University of California, Los Angeles); Sumukh Pinge (University of California, San Diego); Jason Cong (UCLA); Tajana Rosing (UCSD)
366	IG-CRM: Area/Energy-Efficient IGZO-Based Circuits and Architecture Design for Reconfigurable CIM/CAM Applications	Artificial intelligence is evolving with various algorithms such as deep neural network (DNN), Transformer, recommendation system (RecSys) and graph convolutional network (GCN). Correspondingly, multiply-accumulate (MAC) and content search are two main operations, which can be efficiently executed on the emerging computing-in-memory (CIM) and content-addressable-memory (CAM) paradigms. Recently, the emerging Indium-Gallium-Zine-Oxide (IGZO) transistor becomes a promising candidate for both CIM/CAM circuits, featuring ultra-low leakage with >300s data retention time and high-density BEOL fabrication. This paper proposes IG-CRM, the first IGZO-based circuits and architecture design for CIM/CAM applications. The main contributions include: 1) at cell level, propose IGZO-based 3T0C/4T0C cell design that enables both CIM and CAM functionalities while matching IGZO/CMOS voltage; 2) at circuit level, utilize the BEOL IGZO transistor to reduce digital adder tree area in CIM circuits; 3) at architecture level, propose a reconfigurable CIM/CAM architecture with four macro structures based on 3T0C/4T0C cells. The proposed IG-CRM architecture shows high area/energy efficiency on various applications including DNN, Transformer, RecSys and GCN. Experiment results show that IG-CRM achieves 8.09x area saving compared with the SRAM-based non-reconfigurable CIM/CAM baseline, and 1.53E3/51.9 times speedup and 1.63E4/7.62E3 times energy efficiency improvement compared with CPU and GPU on average.	Zeyu Guo (Institute of Microelectronics of the Chinese Academy of Sciences); Jinshan Yue (Institute of Microelectronics of the Chinese Academy of Sciences); Shengzhe Yan (Institute of Microelectronics of the Chinese Academy of Sciences); Zhuoyu Dai (Institute of Microelectronics of the Chinese Academy of Sciences); Xiangqu Fu (Institute of Microelectronics of the Chinese Academy of Sciences); Zhaori Cong (Institute of Microelectronics of the Chinese Academy of Sciences); Zening Niu (Institute of Microelectronics of the Chinese Academy of Sciences); Ke Hu (Institute of Microelectronics of the Chinese Academy of Sciences); Lihua Xu (Institute of Microelectronics of the Chinese Academy of Sciences); Jiawei Wang (Institute of Microelectronics of the Chinese Academy of Sciences); Lingfei Wang (Institute of Microelectronics of the Chinese Academy of Sciences); Guanhua Yang (Institute of Microelectronics of the Chinese Academy of Sciences); Di Geng (Institute of Microelectronics of the Chinese Academy of Sciences); Ling Li (Institute of Microelectronics of the Chinese Academy of Sciences)
374	EOS: An Energy-Oriented Attack Framework for Spiking Neural Networks	Spiking neural networks (SNNs) are emerging as energy-efficient alternatives to conventional artificial neural networks (ANNs). Their event-driven information processing significantly reduces computational demands while maintaining competitive performance. However, as SNNs are increasingly deployed in edge devices, various security concerns have emerged. While significant research efforts have been dedicated to addressing the security vulnerabilities stemming from malicious input, often referred to as adversarial examples, the security of SNN parameters remains relatively unexplored. This work introduces a novel attack methodology for SNNs known as Energy-Oriented SNN attack (EOS). EOS is designed to increase the energy consumption of SNNs through the malicious manipulation of binary bits within their memory systems (i.e., DRAM), where neuronal information is stored. The key insight of EOS lies in the observation that energy consumption in SNN implementations is intricately linked to spiking activity. The bit-flip operation, the well-known Row Hammer technique, is employed in EOS. It achieves this by identifying the most robust neurons in the SNN based on the spiking activity, particularly those related to the firing threshold, which is stored as binary bits in memory. EOS employs a combination of spiking activity analysis and a progressive search strategy to pinpoint the target neurons for bit-flip attacks. The primary objective is to incrementally increase the energy consumption of the SNN while ensuring that accuracy remains intact. With the implementation of EOS, successful attacks on SNNs can lead to an average of 43% energy increase with no drop in accuracy.	Ning Yang (Shanghai Jiao Tong University); Fangxin Liu (Shanghai Jiaotong University); Zongwu Wang (Shanghai Jiaotong University); Haomin Li (Shanghai Jiao Tong University); Zhuoran Song (Shanghai Jiao Tong University); Songwen Pei (University of Shanghai for Science and Technology); Li Jiang (Shanghai Jiao Tong University)
379	Maintaining Sanity: Algorithm-based Comprehensive Fault Tolerance for CNNs	As the deployment of neural networks in safety-critical applications proliferates, it becomes imperative that they exhibit consistent and dependable performance amidst hardware malfunctions. Several protection schemes have been proposed to protect neural networks, but they suffer from huge overheads or insufficient fault coverage. This paper presents Maintaining Sanity, a comprehensive and efficient protection technique for CNNs. Maintaining Sanity extends the state-of-the-art algorithm-based fault tolerance for CNN, utilizing hamming codes and checkpointing to correct over 99.6% of critical faults with about 72% runtime overhead and minimal memory overhead compared to traditional triple modular redundancy (TMR) techniques.	Jinhyo Jung (Yonsei University); Hwisoo So (Arizona State University); Woobin Ko (Yonsei University); Sumedh Shridhar Joshi (Arizona State University); Yebon Kim (Yonsei University); Yohan Ko (Yonsei University); Aviral Shrivastava (Arizona State University); Kyoungwoo Lee (Yonsei University)
382	Data is all you need: Finetuning LLMs for Chip Design via an Automated design-data augmentation framework	Recent advances in large language models have demonstrated their potential for automated generation of Verilog code from high-level prompts. Researchers have utilized fine-tuning to enhance the ability of these large language models (LLMs) in the field of Chip Design. However, the lack of Verilog data hinders further improvement in the quality of Verilog generation by LLMs. Additionally, the absence of a Verilog and EDA script data augmentation framework significantly increases the time required to prepare the training dataset for LLM trainers. In this paper, we propose an automated design-data augmentation framework, which generates high quality natural language description of the Verilog/EDA script. To evaluate the effectiveness of our data augmentation method, we finetune Llama2-13B and Llama2-7B models. The results demonstrate a significant improvement in the Verilog generation task when compared to the general data augmentation method. Moreover, the accuracy of Verilog generation surpasses that of the current state-of-the-art open-source Verilog generation model, increasing from 58.8% to 70.6% with the same benchmark and outperforms GPT-3.5 in Verilog repair and EDA Script Generation with only 13B weights. .	Kaiyan Chang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing; University of Chinese Academy of Sciences); Kun Wang (Hangzhou Institute for Advanced Study; Institute of Computing Technology, Chinese Academy of Science); Nan Yang (Institute of Computing Technology, Chinese Academy of Science; University of Chinese Academy of Science); Ying Wang (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences); Dantong Jin (Zhejiang Lab); Wenlong Zhu (Institute of Computing Technology, Chinese Academy of Science; University of Chinese Academy of Science); Zhirong Chen (Zhejiang University; University of Illinois at Urbana Champaign); Cangyuan Li (ICT); Hao Yan (Shanghai Innovation Center for Processor Technologies; Shanghai University); Yunhao Zhou (Shanghai Innovation Center for Processor Technologies; Shanghai Jiao Tong University); Zhuoliang Zhao (Shanghai Innovation Center for Processor Technologies; FuDan University); Yuan Cheng (Shanghai Innovation Center for Processor Technologies; Nanjing University); Yudong Pan (Research Center for Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Science); Yiqi Liu (Institute of Computing Technology, Chinese Academy of Science; University of Chinese Academy of Science); 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); yinhe han (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)

Submissio	Title	Abstract	Authors with Affiliations
390	Leanor: A Learning-Based Accelerator for Efficient Approximate Nearest Neighbor Search via Reduced Memory Access	Approximate Nearest Neighbor Search (ANNS) is a classical problem in data science. ANNS is both computationally-intensive and memory-intensive. As a typical implementation of ANNS, Inverted File with Product Quantization (IVFPQ) has the properties of high precision and rapid processing. However, the traversal of non-nearest neighbor vectors in IVFPQ leads to redundant memory accesses. This significantly impacts retrieval efficiency. A promising approach involves the utilization of learned indexes, leveraging insights from data distribution to optimize search efficiency. Existing learned indexes are primarily customized for low-dimensional data. How to tackle ANNS in high-dimensional vectors is a challenging issue. This paper introduces Leanor, a learned index-based accelerator for the filtering of non-nearest neighbor vectors within the IVFPQ framework. Leanor minimizes redundant memory accesses, thereby enhancing retrieval efficiency. Leanor incorporates a dimension reduction component, mapping vectors to one-dimensional keys and organizing them in a specific order. Subsequently, the learned index leverages this ordered representation for rapid predictions. To enhance result accuracy, we conduct a thorough analysis of model errors and introduce a specialized index structure named learned index forest (LIF). The experimental results show that, compared to representative approaches, Leanor can effectively filter out non-neighboring vectors within IVFPQ, leading to a substantial enhancement in retrieval efficiency.	Yi Wang (Shenzhen University); Huan Liu (Shenzhen University); Jianan Yuan (Shenzhen University); Jiaxian Chen (Shenzhen University); Tianyu Wang (Shenzhen University); Chenlin Ma (Shenzhen University); Rui Mao (Shenzhen University)
391	Annotating Slack Directly on Your Verilog: Fine-Grained RTL Timing Evaluation for Early Optimization	In digital IC design, the early register-transfer level (RTL) stage offers greater optimization flexibility than post-synthesis netlists or layouts. Some recent machine learning (ML) solutions propose to predict the overall timing of a design at the RTL stage, but the fine-grained timing information of individual registers remains unavailable. In this work, we introduce RTL-Timer, the first fine-grained general timing estimator applicable to any given design. RTL-Timer explores multiple promising RTL representations and customizes loss functions to capture the maximum arrival time at register endpoints. RTL-Timer's fine-grained predictions are further applied to guide optimization in a standard logic synthesis flow.	Wenji Fang (Hong Kong University of Science and Technology); Shang Liu (The Hong Kong University of Science and Technology); Hongce Zhang (The Hong Kong University of Science and Technology (Guangzhou)); Zhiyao Xie (Hong Kong University of Science and Technology)
395	RWriC: A Dynamic Writing Scheme for Variation Compensation for RRAM-based In-Memory Computing	RRAM-based compute-in-memory (CIM) suffers from programming variation issues, specifically device-to-device variation (DDV) and cycle-to-cycle variation (CCV), which can have a detrimental impact on inference accuracy. To address these variation issues, we propose RWriC, a dynamic Writing scheme for Variation Compensation for RRAM-based CIM. RWriC sequentially programs the weights, implemented by multiple RRAM cells, starting from the high significance cell (HSC) and moving towards the low significance cell (LSC). This approach leverages the knowledge of current cumulative errors and the programming targets (PTs) of other RRAM cells to dynamically adjust the PT of the RRAM currently under programming. By shifting the PT of HSC, RWriC enables the LSC to compensate for the programming errors of the HSC. Moreover, when the variation is substantial, RWriC allows the magnitude of LSC to be scaled up, providing an even wider compensation range. Through the combined application of the shifting and scaling techniques, experimental results show that the inference accuracy for ResNet50 on the CIFAR-10 dataset only drops by 0.9% under 18% device variation. In comparison to the conventional writing scheme, our RWriC approach achieves a 5-11x improvement in variation robustness for ResNet50 and Yolov8 across different tasks.	Yucong Huang (Hong Kong University of Science and Technology); Jingyu He (Hong Kong University of Science and Technology); Tim Cheng (HKUST); Chi Ying Tsui (HKUST); Terry Tao Ye (Southern University of Science and Technology)
399	Laser Shield: a Physical Defense with Polarizer against Laser Attack	Autonomous driving systems (ADS) are boosted with deep neural networks (DNN) to perceive environments, while their security is doubted by DNN's vulnerability to adversarial attacks. Among them, a diversity of laser attacks emerges to be a new threat due to its minimal requirements and high attack success rate in the physical world. Nevertheless, current defense methods exhibit either low defense success rate or high computation cost against laser attacks. To fill this gap, we propose Laser Shield which leverages a polarizer along with a min-energy rotation mechanism to eliminate adversarial lasers from ADS scenes. We also provide a physical world dataset, LAPA, to evaluate its performance. Through exhaustive experiments with three baselines, four metrics, and three settings, Laser Shield is proved to exhibit the SOTA performance.	Qingjie Zhang (Tsinghua University); Lijun Chi (Telecom Paris); Di Wang (Beijing University of Posts and Telecommunications); Mounira Msahli (Telecom Paris); Gerard Memmi (Telecom Paris); Tianwei Zhang (Nanyang Technological University); Chao Zhang (Tsinghua University); Han Qiu (Tsinghua University)
404	MSMAC: Accelerating Multi-Scalar Multiplication for Zero-Knowledge Proof	Multi-scalar multiplication (MSM) is the most computation-intensive part in proof generation of Zero-knowledge proof (ZKP). In this paper, we propose MSMAC, an FPGA accelerator for large-scale MSM. MSMAC adopts a specially designed Instruction Set Architecture (ISA) for MSM and optimizes pipelined Point Addition Unit (PAU) with hybrid Karatsuba multiplier. Moreover, a runtime system is proposed to split MSM tasks with the optimal sub-task size and orchestrate execution of Processing Elements (PEs). Experimental results show that MSMAC achieves up to 328X and 1.96X speedups compared to the state-of-the-art implementation on CPU (one core) and GPU, respectively, outperforming the state-of-the-art ASIC accelerator by 1.79X. On 4 FPGAs, MSMAC performs 1,261X faster than a single CPU core.	Pengcheng Qiu (Ant Group); Guiming Wu (Ant Group); Tingqiang Chu (Ant Group); Changzheng Wei (Ant Group); Runzhou Luo (Ant Group); Ying Yan (Ant Group); Wei Wang (Ant Group); Hui Zhang (Ant Group)
414	Reducing DRAM Latency via In-situ Temperature- and Process-Variation-Aware Timing Detection and Adaption	Long DRAM access latency has a significant impact on modern system performance. However, the improvement of access latency is limited as the DRAM vendors reserve considerable timing margins against seldom worst-case conditions. To mitigate such pessimistic timing margins, we propose a temperature- and process-variation-aware timing detection and adaption DRAM (TPDA-DRAM) architecture. It equips in-situ cross-coupled detectors to monitor the voltage difference between bitline pairs, enabling estimation of timing margins caused by process and temperature variations. Moreover, TPDA-DRAM incorporates two collaborative timing adaption schemes: 1) a process-variation-aware timing adaption scheme (PVA) that selectively accelerates the access to rare weak cells and 2) a temperature-variation-aware timing adaption scheme (TVA) that precisely adjust timing parameters by adopting temperature information. Compared to prior art, the proposed detector reduces detection deviation by 54.8% and area overhead by 88.1%. The system-level evaluation in an eight-core system shows that TPDA-DRAM improves the average performance and energy efficiency by 20.5% and 15.0%, respectively.	Yuxuan Qin (Shanghai Jiao Tong University); Chuxiong Lin (Shanghai Jiao Tong University); Mingche Lai (National University of Defense Technology); Zhang Luo (National University of Defense Technology); Shi Xu (National Innovation Institute of Defense Technology); Weifeng He (Shanghai Jiao Tong University)
415	TSAcc: An Efficient \underline{T}empo-\underline{S}patial Similarity Aware \underline{Acc}elerator for Attention Acceleration	Attention-based models provide significant accuracy improvement to Natural Language Processing (NLP) and computer vision (CV) fields at the cost of heavy computational and memory demands. Previous works seek to alleviate the performance bottleneck by removing useless relations for each position. However, their attempts only focus on intra-sentence optimization and overlook the opportunity in the temporal domain. In this paper, we accelerate attention by leveraging the tempo-spatial similarity across successive sentences, given the observation that successive sentences tend to bear high similarity. This is rational owing to many semantic similar words (namely tokens) in the attention-based models. We first propose an online-offline prediction algorithm to identify similar tokens/heads. We then design a recovery algorithm so that we can skip the computation on similar tokens/heads in succeeding sentences and recover their results by copying other tokens/heads features in preceding sentences to reserve accuracy. From the hardware aspect, we propose a specialized architecture TSAcc that includes a prediction engine and recovery engine to translate the computational saving in the algorithm to real speedup. Experiments show that TSAcc can achieve $8.5\times$, $2.7\times$, $14.1\times$, and $64.9\times$ speedup compared to SpAtten, Sanger, 1080TI GPU, and Xeon CPU, with negligible accuracy loss.	Zhuoran Song (Shanghai Jiao Tong University); Chunyu Qi (Shanghai Jiao Tong University); Yuanzheng Yao (Shanghai Jiao Tong University); Peng Zhou (Alibaba Cloud); Yanyi Zi (Alibaba Cloud); Nan Wang (Alibaba Cloud); Xiaoyao Liang (Shanghai Jiao Tong University)

Submissio	Title	Abstract	Authors with Affiliations
419	PathFuzz: Broadening Fuzzing Horizons with Footprint Memory for CPUs	Coverage metrics have been widely adopted to quantify the completeness of hardware verification. Recently, coverage-guided fuzzing has emerged as a popular method for automatically creating test inputs toward higher verification coverage reach. However, we observe that its effectiveness on CPUs is hindered by limited sources of seed corpus and efficiency of mutations. To broaden the fuzzing horizons, this paper proposes the PathFuzz framework incorporating an efficient input format for fuzzing CPUs, the footprint memory, with seed corpus from real-world large-scale programs. Experiments demonstrate that using PathFuzz reaches over 95% verification coverage with four long-standing bugs newly identified in two well-known open-source CPU designs.	Yinan Xu (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences); Sa Wang (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences); Dan Tang (Beijing Institute of Open Source Chip); Ninghui Sun (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences); Yungang Bao (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences)
426	Neural Barrier Certificates Synthesis of NN-Controlled Continuous Systems via Counterexample-Guided Learning	There is a pressing need to ensure the safety of closed-loop systems with NN controllers. To address this issue, we propose a novel approach for generating barrier certificates, which combines counterexample-guided learning with efficient SOS-based verification. Our proposed method offers an efficient verification procedure that solves three linear matrix inequality (LMI) constraint feasibility testing problems, instead of relying on an SMT solver to verify the barrier certificate conditions. We conduct comparison experiments on a set of benchmarks, demonstrating the advantages of our method in terms of efficiency and scalability, which enable effective verification of high-dimensional systems.	Hanrui Zhao (East China Normal University); Niuniu Qi (East China normal University); Mengxin Ren (East China Normal University); Xia Zeng (Southwest University); Zhenbing Zeng (Shanghai University); Zhengfeng Yang (East China Normal University)
430	GCS-Timer: GPU-Accelerated Current Source Model Based Static Timing Analysis	Composite Current Source (CCS) timing model plays an important role in modern static timing analysis (STA) because it precisely captures the timing behavior of a design at advanced nodes. However, CCS is extremely time-consuming due to its accurate but complicated timing models. To overcome this challenge, we introduce GCS-Timer, a GPU-accelerated CCS-based timing analysis algorithm. Unlike existing methods that perform model order reduction to trade accuracy for speed, GCS-Timer achieves high accuracy through a fast simulation-based analysis using GPU computing. Experimental results show that GCS-Timer can complete CCS analysis with better accuracy and achieve 3.2X faster runtime compared with a 16-threaded industrial standard timer.	Shiju Lin (The Chinese University of Hong Kong); Guannan Guo (UIUC); Tsung-Wei Huang (University of Wisconsin at Madison); Weihua Sheng (Huawei Hong Kong Research Center); Evangeline Young (The Chinese University of Hong Kong); Martin Wong (The Chinese University of Hong Kong)
432	SkyPlace: A New Mixed-size Placement Framework using Modularity-based Clustering and SDP Relaxation	Placement is one of the most essential problems of VLSI physical design. Recently, the electrostatics-based placement has made a great success and inspired many placement algorithms. However, the recent direction of improvement is missing two important problems for mixed-size placement – 1) how to initialize placement and 2) how to handle macros in the analytical placement. In this paper, we propose new mixed-size placer, SkyPlace which is enhanced by novel placement initialization using semidefinite programming relaxation and density weighting technique. Our experimental results show that SkyPlace clearly outperforms the leading-edge placer on the MMS benchmarks.	Jaekyung Im (Pohang University of Science and Technology); Seokhyeong Kang (Pohang University of Science and Technology)
438	Towards High-Performance Virtual Platforms: A Parallelization Strategy for SystemC TLM-2.0 CPU Models	SystemC TLM-2.0 is currently the industry standard for simulating full Systems-on-a-Chip (SoCs). Although SystemC is designed to simulate the behavior of complex, parallel systems, the simulation itself is by default single-threaded. We present a technique to overcome this performance limitation by parallelizing the CPU model of a SystemC-TLM-2.0-based system-level simulator, a so-called Virtual Platform (VP). Our solution is fully compliant with the SystemC standard. To further increase the performance, we developed algorithms for asynchronous DMI pointer caching and we introduced a new tunable parameter called async_rate. This parameter controls the frequency used to annotate timing information to SystemC. Evaluation results demonstrate a significant speedup compared to sequential execution, with a maximum of 7.8 x achieved for octacore VPs on fully parallelizable workloads. For the execution of the NPB suite on the SIM-V VP, an average speedup of 6.2 x is achieved. This approach is a promising solution for accelerating VPs while adhering to the SystemC standard.	Nils Bosbach (RWTH Aachen University); Niko Zurstraßen (RWTH Aachen Institute for Communication Technologies and Embedded Systems); Rebecca Pelke (RWTH Aachen University); Lukas Jünger (MachineWare GmbH); Jan Henrik Weinstock (MachineWare GmbH); Rainer Leupers (RWTH Aachen University)
440	Multi-Resonance Mesh-Based Wavelength-Routed Optical Networks-on-Chip	Wavelength-routed optical networks-on-chip (WRONoCs) are well-known for providing high-speed and collision-free communication in multi-core processors. Previous work was unable to simultaneously reduce the design complexity and total optical power consumption of WRONoC. Besides, in current designs, each microring resonator (MRR), which is the key component of WRONoCs, is configured to demultiplex to one specific wavelength. This significantly increases the MRR usage and the insertion loss. In this work, we adapt different types of ONoC routers into the mesh-based template. To reduce MRR usage, we take advantage of an important feature of MRR, multi-resonance, so that a single MRR can demultiplex signals on multiple wavelengths. To this end, we propose an efficient design method that synthesizes mesh-based WRONoCs using multi-resonance MRRs and existing optical routers to reduce total power consumption. The experimental results show that our method outperforms state-of-the-art design methods in significantly reducing MRR usage and optical power.	Zhidan Zheng (Technical University of Munich); Liaoyuan Cheng (Technical University of Munich); Kanta Arisawa (Ritsumeikan University); Qingyu Li (Technical University Munich); Alexandre Truppel (Technical University of Munich); Shigeru Yamashita (Ritsumeikan University); Tsun-Ming Tseng (Technical University of Munich); Ulf Schlichtmann (Technical University of Munich)
452	Obstacle-Aware Length-Matching Routing for Any-Direction Traces in Printed Circuit Board	Emerging applications in Printed Circuit Board (PCB) routing impose new challenges on automatic length matching, including adaptability for any-direction traces with their original routing preserved for interactiveness. The challenges can be addressed through two orthogonal stages: assign non-overlapping routing regions to each trace and meander the traces within their regions to reach the target length. In this paper, mainly focusing on the meandering stage, we propose an obstacle-aware detailed routing approach to optimize the utilization of available space and achieve length matching while maintaining the original routing of traces. Furthermore, our approach incorporating the proposed Multi-Scale Dynamic Time Warping (MSDTW) method can also handle differential pairs against common decoupled problems. Experimental results demonstrate that our approach has effective length-matching routing ability and compares favorably to previous approaches under more complicated constraints.	Weijie Fang (Fuzhou University); Longkun Guo (Fuzhou University & Chinese Academy of Sciences Shenzhen Advanced Technology Academe); Jiawei Lin (Fuzhou University); Silu Xiong (Hangzhou Huawei Enterprises Telecommunication Technologies Co., Ltd); Huan He (Hangzhou Huawei Enterprises Telecommunication Technologies Co., Ltd); Jiacen Xu (Shanghai LEDA Technology Co., Ltd); Jianli Chen (Fudan University)
458	EasyACIM: An End-to-End Automated Analog CIM with Synthesizable Architecture and Agile Design Space Exploration	Analog Computing-in-Memory (ACIM) is an emerging architecture to perform efficient AI edge computing. However, current ACIM designs usually have unscalable topology and still heavily rely on manual efforts. These drawbacks limit the ACIM application scenarios and lead to an undesired time-to-market. This work proposes an end-to-end automated ACIM based on a synthesizable architecture (EasyACIM). With a given array size and customized cell library, EasyACIM can generate layouts for ACIMs with various design specifications end-to-end automatically. Leveraging the multi-objective genetic algorithm (MOGA)-based design space explorer, EasyACIM can obtain high-quality ACIM solutions based on the proposed synthesizable architecture, targeting versatile application scenarios. The ACIM solutions given by EasyACIM have a wide design space and competitive performance compared to the state-of-the-art (SOTA) ACIMs.	Haoyi Zhang (Peking University); Jiahao Song (Peking University); Xiaohan Gao (Peking University); Xiyuan Tang (Peking University); Yibo Lin (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University)

Submissio	Title	Abstract	Authors with Affiliations
465	G-PASTA: GPU Accelerated Partitioning Algorithm for Static Timing Analysis	Static timing analysis (STA) is an important stage in the modern EDA design flow. But STA becomes timing-consuming with the growth of modern circuit size. Recent research has leveraged task dependency graph (TDG) parallelism to accelerate STA. Despite the speedup through TDG parallelism, the performance can be further enhanced by reducing the scheduling cost. A common solution for reducing scheduling cost is TDG partitioning. However, the runtime of existing TDG partitioning algorithms grows rapidly as the TDG size enlarges. Also, TDG partitioning is frequently invoked during STA process. This make TDG partitioning runtime adds up to a significant portion of the entire STA runtime. As a result, it is important to optimize the runtime performance of TDG partitioning. In this paper, we propose G-PASTA, a GPU-accelerated TDG partitioning algorithm by harnessing the computation power of modern GPU architectures. We evaluate the performance of G-PASTA on a set of TDGs from large designs. Compared to the state-of-the-art TDG partitioner, G-PASTA is up to 41.8× faster, while improving TDG runtime by 2×.	Boyang Zhang (University of Wisconsin, Madison); Dian-Lun Lin (University of Wisconsin-Madison); Che Chang (University of Wisconsin, Madison); Cheng-Hsiang Chiu (University of Wisconsin-Madison); Bojue Wang (Rutgers University); Wan Luan Lee (University of Wisconsin-Madison); Chih-Chun Chang (University of Wisconsin-Madison); Donghao Fang (Texas A&M University); Tsung-Wei Huang (University of Wisconsin at Madison)
471	Finding Bugs in RTL Descriptions: High-Level Synthesis to the Rescue	Most RTL designs originate from behavioral descriptions specified in C or C++. These are often written by SW designers. Hardware (HW) designers then manually build an efficient hardware implementation of that application using a Hardware Description Language (HDL) like Verilog or VHDL. Although it has been shown that High-Level Synthesis (HLS) provides a direct path to synthesizing these behavioral descriptions into RTL, the quality of the generated RTL is often still unacceptable, hence, requiring the manual RTL design. This is nevertheless time consuming and error prone. In particular, finding bugs introduced in the manual design is very tedious as HW designers typically rely on long simulations that generate large waveforms that have to be thoroughly scrutinized. To address this, in this work we present an automated method to accurately point to where in an RTL description a bug is located by using HLS. In particular we leverage the ability of HLS to generate a variety of different micro-architectures to automatically find a design architecturally `similar' to the manually optimized one in order to help locate the bug.	Baharealsadat Parchamdar (The University of Texas at Dallas); Benjamin Carrion Schaefer (The University of Texas at Dallas)
474	Digital CIM with Noisy SRAM Bit: A Compact Clustered Annealer for Large-Scale Combinatorial Optimization	Combinatorial optimization problems (COP) are NP-hard and intractable to solve using conventional computing. The Ising model-based annealer has gained increasing attention recently due to its efficiency and speed in finding approximate solutions. However, Ising solvers for travelling salesman problems (TSP) usually suffer from a scalability issue due to quadratically increasing number of spins. In this paper, we propose a digital computing-in-memory (CIM) based clustered annealer to solve tens of thousands of city-scale TSP with only a few mega-byte (MB) of static random access memory (SRAM), using hierarchical clustering to solve input sparsity and digital CIM flexibility to solve weight sparsity. The intrinsic process variations between SRAM devices are utilized to generate the noisy bit errors during pseudo-read under reduced supply voltage, realizing the annealing process. The design space of cluster size and programmability is explored to understand the trade-offs of solution quality and hardware cost, for TSP scale ranging from 3080 to 85900 cities. The proposed design speeds up the convergence by >10^9× with <25% solution quality overhead compared with the CPU baseline. The comparison with state-of-the-art scalable annealers shows a >10^13× improvement on functionally normalized area and power.	Anni Lu (Georgia Institute of Technology); Junmo Lee (Georgia Institute of Technology); Yuan-Chun Luo (Georgia Institute of Technology); Hai Li (Components Research, Intel Corporation); Ian Young (Components Research, Intel Corporation); Shimeng Yu (Georgia Institute of Technology)
481	RL-PTQ: RL-based Mixed Precision Quantization for Hybrid Vision Transformers	Existing quantization approaches incur significant accuracy loss when compressing hybrid transformers with low bit-width. This paper presents RL-PTQ, a novel post-training quantization (PTQ) framework utilizing reinforcement learning (RL). Our focus is on determining the most effective bit-width and observer for quantization configurations tailored for mixed-precision by grouping layers and addressing the challenges of quantization of hybrid transformers. We achieved the highest quantized accuracy for MobileViTs compared to the previous PTQ methods. Furthermore, our quantized model on PIM architecture exhibited an energy efficiency enhancement of 10.1× and 22.6× compared to the baseline model, on the state-of-the-art PIM accelerator and GPU, respectively.	Eunji Kwon (Postech); Minxuan Zhou (UCSD); Weihong Xu (University of California San Diego); Tajana Rosing (UCSD); Seokhyeong Kang (Pohang University of Science and Technology)
482	CDLS: Constraint Driven Generative AI Framework for Analog Layout Synthesis	In advanced process technology nodes, analog circuit performance is intrinsically linked to layout parasitics, and layout dependent effects (LDE). In contrast to digital designs, layout generation for analog mixed signal circuits remains predominantly a slow manual task, impeding rapid design convergence. To address this bottleneck, we introduce CDLS - a Constraint Driven Generative AI Framework for Analog Layout Synthesis. CDLS is fundamentally a constraint driven framework that enables analog circuit designers to auto-generate simulation-ready layout. Unlike traditional algorithmic approaches, CDLS uses generative AI and machine learning techniques to generate key design constraints that drive the quality of autogenerated placement and routing. Using CDLS on average we reduce layout iteration time by 2-3X on industrial designs. By reducing the turn-around-time on layout iterations we estimate a 30% reduction to overall design convergence cycle. We also demonstrate the quality of results achieved through CDLS is on par with manual drawn layout, on state-of-the-art analog designs developed on an Intel sub-10nm process technology node.	Prasanth Mangalagiri (Intel Corporation); Lynn Qian (Intel Corporation); Farrukh Zafar (Intel Corporation); Praveen Mosalikanti (Intel Corporation); Phoebe Chang (Intel Corporation); Arun Kurian (Intel Corporation); Saripalli Vinay (Intel Corporation)
489	Toward High-Accuracy, Programmable Extreme-Edge Intelligence for Neuromorphic Vision Sensors utilizing Magnetic Domain Wall Motion-based MTJ	The desire to empower resource-limited edge devices with computer vision (CV) must overcome the high energy consumption of collecting and processing vast sensory data. To address the challenge, this work proposes an energy-efficient non-von-Neumann in-pixel processing solution for neuromorphic vision sensors employing emerging (X) magnetic domain wall magnetic tunnel junction (MDWMTJ) for the first time, in conjunction with CMOS-based neuromorphic pixels. Our hybrid CMOS+X approach performs in-situ massively parallel asynchronous analog convolution, exhibiting low power consumption and high accuracy across various CV applications by leveraging the non-volatility and programmability of the MDWMTJ. Moreover, our developed device-circuit-algorithm co-design framework captures device constraints (low tunnel-magnetoresistance, low dynamic range) and circuit constraints (non-linearity, process variation, area consideration) based on monte-carlo simulations and device parameters utilizing GF22nm FD-SOI technology. Our experimental results suggest we can achieve an average of 45.3% reduction in backend-processor energy, maintaining similar front-end energy compared to the state-of-the-art and high accuracy of 79.17% and 95.99% on the DVS-CIFAR10 and IBM DVS128-Gesture datasets, respectively.	Md Abdullah-Al Kaiser (University of Southern California); Gourav Datta (University of Southern California); Peter A. Beerel (University of Southern California); Akhilesh R. Jaiswal (University of Wisconsin-Madison)
494	Low-Complexity Algorithmic Test Generation for Neuromorphic Chips	We propose an algorithmic test generation method for neuromorphic chips without Design-for-Testability. Fault activation differentiates a neuron's good output and faulty output. Fault propagation sensitizes fault effects to differentiate outputs of faulty chips and good chips. On an L-layer Spiking Neural Network (SNN) model, we achieve 100% fault coverage using O(L) test configurations and test patterns under negligible or no weight variation. Our results show that test effectiveness is maintained even with 4-bit weight quantization. We incur no test escape and overkill even under 10% weight variation. Our total test length is over 73K times shorter than previous works.	Hsu-Yu Huang (National Taiwan University); Chu-Yun Hsiao (National Taiwan University); Tsung-Te Liu (National Taiwan University); James Chien-Mo Li (National Taiwan University)

Submissio	Title	Abstract	Authors with Affiliations
497	SSRESF: Sensitivity-aware Single-particle Radiation Effects Simulation Framework in SoC Platforms based on SVM Algorithm	The ever-expanding scale of integrated circuits has brought about a significant rise in the design risks associated with radiation-resistant integrated circuit chips. Traditional single-particle experimental methods, with their iterative design approach, are increasingly ill-suited for the challenges posed by large-scale integrated circuits. In response, this article introduces a novel sensitivity-aware single-particle radiation effects simulation framework tailored for System-on-Chip platforms. Based on SVM algorithm we have implemented fast finding and classification of sensitive circuit nodes. Additionally, the methodology automates soft error analysis across the entire software stack. The study includes practical experiments focusing on RISC-V architecture, encompassing core components, buses, and memory systems. It culminates in the establishment of databases for Single Event Upsets (SEU) and Single Event Transients (SET), showcasing the practical efficacy of the proposed methodology in addressing radiation-induced challenges at the scale of contemporary integrated circuits. Experimental results have shown up to 12.78× speed-up on the basis of achieving 94.58\% accuracy.	MENG LIU (Beijing University of Technology); Shuai Li (Beijing University of Technology); Fei Xiao (Beijing University of Technology); Ruijie Wang (Beijing University of Technology); Chunxue Liu (Beijing Microelectronics Technology Institute); Liang Wang (Beijing Microelectronics Technology Institute)
501	AdaptiveFL: Adaptive Heterogeneous Federated Learning for Resource-Constrained AIoT Systems	Although Federated Learning (FL) is promising to enable collaborative learning among Artificial Intelligence of Things (AIoT) devices, it suffers from the problem of low classification performance due to various heterogeneity factors (e.g., computing capacity, memory size) of devices and uncertain operating environments. To address these issues, this paper introduces an effective FL approach named AdaptiveFL based on a novel fine-grained width-wise model pruning strategy, which can generate various heterogeneous local models for heterogeneous AIoT devices. By using our proposed reinforcement learning-based device selection mechanism, AdaptiveFL can adaptively dispatch suitable heterogeneous models to corresponding AIoT devices on the fly based on their available resources for local training. Experimental results show that, compared to state-of-the-art methods, AdaptiveFL can achieve up to 16.83% inference improvements for both IID and non-IID scenarios.	Chentao Jia (East China Normal University); Ming Hu (Nanyang Technological University); Zekai Chen (East China Normal University); Yanxin Yang (East China Normal University); Xiaofei Xie (Singapore Management Univerisity); Yang Liu (Nanyang Technological University); Mingsong Chen (East China Normal University)
502	Concurrent Detailed Routing with Pin Pattern Re-generation for Ultimate Pin Access Optimization	Pin access has become one of the most significant challenges in large-scale full-chip routing due to the continuous reduction in feature sizes and the increasing complexity of designs. The conventional standard cell layout synthesis approaches usually optimize pin accessibility by maximizing pin lengths and access points. However, these pre-determined pin patterns greatly occupy routing resources and may contrarily degrade routability. To address this problem, this paper proposes the first work of concurrent detailed routing with pin pattern re-generation to achieve ultimate pin access optimization. A pseudo-pin extraction and routing technique is proposed that can secure one access point for each input/output pin while allowing the remaining access points to be routable by other nets. The experimental results demonstrate that the proposed method can resolve 89% of local regions that are unroutable with original layout patterns without compromising power and timing performances.	Ying-Jie Jiang (National Taiwan University of Science and Technology); Shao-Yun Fang (National Taiwan University of Science and Technology)
504	Go Beyond Black-box Policies: Rethinking the Design of Learning Agent for Interpretable and Verifiable HVAC Control	Recent research has shown the potential of Model-based Reinforcement Learning (MBRL) to enhance energy efficiency of Heating, Ventilation, and Air Conditioning (HVAC) systems. However, existing methods rely on black-box thermal dynamics models and stochastic optimizers, lacking reliability guarantees and posing risks to occupant health. We address this by redesigning HVAC controllers using decision trees extracted from thermal models and historical data, providing deterministic, verifiable, and interpretable policies. Extensive experiments show that our method saves 68.4% more energy and increases human comfort gain by 14.8% compared to the state-of-the-art method, plus a 1127x reduction in computation overhead. Code: https://github.com/30363/Veri-HVAC.	Zhiyu An (University of California, Merced); Xianzhong Ding (University of California, Merced); Wan Du (University of California, Merced)
515	ChatPattern: Layout Pattern Customization via Natural Language	Existing works focus on fixed-size layout pattern generation, while the more pratical free-size pattern generation receives limited attention. In this paper, we propose ChatPattern, a novel Large-Language-Model (LLM) powered framework for flexible pattern customization. ChatPattern utilizes a two-part system featuring an expert LLM-agent and a highly controllable layout pattern generator. The LLM-agent can interpret natural language requirements and operate design tools to meet specified needs, while the generator excels in conditional layout generation, pattern modification, and memory-friendly patterns extension. Experiments on challenging pattern generation setting shows the ability of ChatPattern to synthesize high-quality large-scale patterns.	Zixiao WANG (The Chinese University of Hong Kong); Yunheng Shen (Tsinghua University); Xufeng Yao (Chinese University of HongKong); Wenqian Zhao (The Chinese University of Hong Kong); Yang BAI (CUHK); Farzan Farnia (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong)
519	EmMark: Robust Watermarks for IP Protection of Embedded Quantized Large Language Models	This paper introduces EmMark, a novel watermarking framework for protecting intellectual property (IP) of embedded large language models deployed on resource-constrained edge devices. To address the IP theft risks posed by malicious end-users, EmMark enables proprietors to authenticate ownership by querying the watermarked model weights and matching the inserted signatures. EmMark's novelty lies in its strategic watermark weight parameters selection, ensuring robustness and maintaining model quality. Extensive proof-of-concept evaluations of models from OPT and LLaMA-2 families demonstrate EmMark's fidelity, achieving 100% success in watermark extraction with model performance preservation. EmMark also showcased its resilience against watermark removal and forging attacks.	Ruisi Zhang (UC San Diego); Farinaz Koushanfar (University of California San Diego)
520	Accelerating Regular Path Queries over Graph Database with Processing-in-Memory	Regular path queries (RPQs) in graph databases are bottlenecked by the memory wall. Emerging processing-in-memory (PIM) technologies offer a promising solution to dispatch and execute path matching tasks in parallel within PIM modules. We present Moctopus, a PIM-based data management system for graph databases that supports efficient batch RPQs and graph updates. Moctopus employs a PIM-friendly dynamic graph partitioning algorithm, which tackles graph skewness and preserves graph locality with low overhead for RPQ processing. Moctopus enables efficient graph updates by amortizing the host CPU's update overhead to PIM modules. Evaluation of Moctopus demonstrates superiority over the state-of-the-art traditional graph database.	Ruoyan Ma (Shanghai Jiao Tong University); Shengan Zheng (Shanghai Jiao Tong University); Guifeng Wang (Shanghai Jiao Tong University); Jin Pu (Shanghai Jiao Tong University); Yifan Hua (Shanghai Jiao Tong University); Wentao Wang (Peking University); Linpeng Huang (Shanghai Jiao Tong University)
521	Hynify: A High-throughput and Unified Accelerator for Multi-Mode Nonparametric Statistics	Nonparametric statistics methods are a class of robust and potent statistics, which are widely used in various domains such as finance, medicine, and computer science. Such methods deliver an accurate estimation without an assumed data distribution. Moreover, they can handle discrete data with various data sources. Despite their desirable features, the calculation of large-scale nonparametric statistics is both compute- and memory-intensive, and the performance overhead hinders them from widespread usage. This paper identifies that the key performance bottleneck lies in the rank-based operations which are intensively involved in variants of nonparametric statistics methods. These rank-based operations can thereby be fully accelerated and structurally reused among diverse statistics. We then introduce Hynify, a high-throughput and unified accelerator that facilitates a rich set of nonparametric statistics. To ensure comprehensiveness, we capture three primary computational paradigms of nonparametric statistical methods, namely, aggregation, pair-wise rank, and concordance, with the right architecture designs. To improve throughput, Hynify exploits fine-grained computation and pipelining for increased performance. We implement Hynify in FPGA demonstration and representative experimental results demonstrate that Hynify delivers up to 160x/21x throughput improvement over GPU and 64-core CPU, respectively, while achieving up to 781x/62x energy efficiency improvement.	Kaihong Huang (Southeast University); Dian Shen (Southeast University); Zhaoyang Wang (Southeast University); Juntao Yang (Southeast University); Beilun Wang (Southeast University)

Submissio	Title	Abstract	Authors with Affiliations
526	SpREM: Exploiting Hamming Sparsity for Fast Quantum Readout Error Mitigation	The current Noisy Intermediate-Scale Quantum (NISQ) era suffers from high quantum readout error that severely reduces the measurement fidelity. Matrix-based error mitigation has been demonstrated as a promising software-level technique, which performs matrix-vector multiplication to calibrate the probability distribution with noise. However, this approach shows poor scalability and limited fidelity improvement as the matrix size exponentially increases with the number of qubits. In this paper, we propose SpREM to exploit the inherent sparsity in the mitigation matrix. Inspired by the interaction mechanism between qubits, we identify structured sparsity patterns using Hamming distance. With this insight, we propose the Hamming-Distance Sparse Row (HDSR) compression method and its format, which can achieve higher sparsity than threshold-based pruning meanwhile exhibiting great fidelity improvement. Finally, we propose the computational dataflow of the HDSR format and implement it on hardware. Experiments demonstrate that SpREM achieves 98.9% sparsity and a 27.3× reduction in fidelity loss on the real-world quantum device, compared to threshold pruning. It achieves an average 11.2× ~ 36.4× speedup compared to Xilinx Vitis SPARSE library and NVIDIA A100 GPU implementations.	Hanyu Zhang (Zhejiang University); Liqiang Lu (Zhejiang University); Siwei Tan (Zhejiang university); Size Zheng (Peking University); Jia Yu (Zhejiang University); Jianwei Yin (Zhejiang University)
530	MAFin: Maximizing Accuracy in FinFET based Approximated Real-Time Computing	We propose MAFin that exploits the unique temperature effect inversion (TEI) property of a FinFET based multicore platform, where processing speed increases with temperature, in the context of approximate real-time computing. With an objective to maximize the QoS for a FinFET based multicore system, MAFin, our proposed real-time scheduler, first derives a task-to-core allocation while respecting system-wide constraints and prepares a schedule. During execution, MAFin further increases the achieved QoS by exploiting TEI property of FinFET based processors while balancing the performance and temperature and respects the imposed constraints on-the-fly by incorporating a prudential temperature cognizant frequency management mechanism.	Shounak Chakraborty (Norwegian University of Science and Technology (NTNU)); Sangeet Saha (University of Essex); Magnus Själander (Norwegian University of Science and Technology); Klaus McDonald-Maier (University of Essex)
532	SWAT: Scalable and Efficient Window Attention-based Transformers Acceleration on FPGAs	Efficiently supporting long context length is crucial for Transformer models. The quadratic complexity of the self-attention computation plagues traditional Transformers. Sliding window-based static sparse attention mitigates the problem by limiting the attention scope of the input tokens, reducing the theoretical complexity from quadratic to linear. Although the sparsity induced by window attention is highly structured, it does not align perfectly with the microarchitecture of the conventional accelerators, leading to suboptimal implementation. In response, we propose a dataflow-aware FPGA-based accelerator design, SWAT, that efficiently leverages the sparsity to achieve scalable performance for long input. The proposed microarchitecture is based on a design that maximizes data reuse by using a combination of row-wise dataflow, kernel fusion optimization, and an input-stationary design considering the distributed memory and computation resources of FPGA. Consequently, it achieves up to 22x and 5.7x improvement in latency and energy efficiency compared to the baseline FPGA-based accelerator and 15x energy efficiency compared to GPU-based solution.	Zhenyu Bai (National University of Singapore); Pranav Dangi (National University of Singapore); Huize Li (National University of Singapore); Tulika Mitra (National University of Singapore)
541	Gypsophila: A Scalable and Bandwidth-Optimized Multi-Scalar Multiplication Architecture	Multi-Scalar Multiplication (MSM) is a fundamental cryptographic primitive, which plays a crucial role in Zero-knowledge proof systems. In this paper, we optimize the single MSM Process Element (PE) utilizing buckets with fewer conflicts, enhanced by Greedy-based scheduling, to achieve higher efficiency. The evaluation results show our optimized single MSM PE achieving a speedup of over two times on average, peaking at 3.63 times compared to previous works. Furthermore, we introduce Gypsophila, a scalable and bandwidth-optimized architecture for implementing multiple MSM PEs. Leveraging the characteristics of the bucket method, we optimize the data flow by balancing the throughput of bucket classification, bucket aggregation, and result aggregation in MSM. Simultaneously, multiple PEs with different data access patterns share a universal point input channel and post-processing unit, which improves the module utilization and mitigates the bandwidth pressure. Gypsophila with 16 PEs, accomplishes 16 MSM tasks in a mere 1.01% additional time, showcasing an approximate 7.8% reduction in area, with only about 1/16 of the bandwidth requirement, compared with 16 PEs without input channel and post-process unit sharing.	Changxu Liu (Fudan University); Hao Zhou (Fudan University); Lan Yang (Fudan University); Jiamin Xu (Fudan University); Patrick Dai (Semisand Chip Design Pte.Ltd); Fan Yang (Fudan University)
542	DH-TRNG: A Dynamic Hybrid TRNG with Ultra-High Throughput and Area-Energy Efficiency	As a vital security primitive, the true random number generator (TRNG) is a mandatory component to build roots of trust for any encryption system. However, existing TRNGs suffer from bottlenecks of low throughput and high area-energy consumption. In this work, we propose DH-TRNG, a dynamic hybrid TRNG circuitry architecture with ultra-high throughput and area-energy efficiency. Our DH-TRNG exhibits portability to distinct process FPGAs and passes both NIST and AIS-31 tests without any post-processing. The experiments show it incurs only 8 slices with the highest throughput of 670 Mbps and 620 Mbps on Xilinx Virtex-6 and Artix-7, respectively. Compared to the state-of-the-art TRNGs, our proposed design has the highest Throughput/(Slices·Power) with 2.63× increase.	Yuan Zhang (Hunan University); Kuncai Zhong (College of Integrated Circuits, Hunan University); Jiliang Zhang (College of Integrated Circuits, Hunan University)
549	DySpMM: From Fix to Dynamic for Sparse Matrix-Matrix Multiplication Accelerators	Sparse Matrix-Matrix Multiplication (SpMM) is one of the key operators in many fields, showing dynamic features in terms of sparsity, element distribution, and data dependency. Previous studies have proposed FPGA-based SpMM accelera- tors with fixed configurations, leaving three major challenges unsolved: 1) Partitioning matrices with the fixed sub-matrix size leads to performance loss, because the optimal feasible sub- matrix size to minimize memory access varies with dynamic sparsity. 2) The fixed row-base allocation scheme of streaming architecture leads to unbalanced workloads because of dynamic element distribution across sparse matrix rows. 3) Data conflict makes the elements in one row cannot be processed consecutively. Architectures with fixed execution order rely on time-consuming pre-processing to deal with dynamic data dependency. Motivated by the observation that fixed configurations leads to performance loss, we propose DySpMM by introducing the dynamic design methodology to SpMM architectures. The configurable data distribution data path is designed to enable dynamic sub-matrix size, achieving up to 3.43× speed-up. The element-wise allocation unit is introduced into hardware for dynamic workload balancing, improving utilization up to 3.74×. The interleaved reorder unit is proposed to automatically reorder the sparse elements and dynamically avoid conflicts, completely avoiding the pre-processing overhead. The evaluation of DySpMM on FPGA shows that DySpMM achieves 1.42× geomean throughput of the state-of-the-art accelerator Sextans and 1.78× energy efficiency compared with V100S GPU.	Hongyi Wang (Tsinghua University); Kai Zhong (Tsinghua University); Haoyu Zhang (Tsinghua University); Shulin Zeng (Tsinghua university); Zhenhua Zhu (Tsinghua University); Xinhao Yang (Tsinghua University); Shuang Wang (Tsinghua University); Guohao Dai (Shanghai Jiao Tong University); Huazhong Yang (Tsinghua University); Yu Wang (Tsinghua University)

Submissio	Title	Abstract	Authors with Affiliations
552	VVIP: Versatile Vertical Indexing Processor for Edge Computing	This paper presents a versatile vertical indexing processor (VVIP) based on a single-instruction multiple-data architecture for edge computing. In VVIP, the vertical source and destination indexing instructions are customized for area-efficient computations. The proposed indexing method reorders data within a processing module by using more registers and data-steering logic in the calculations. In particular, VVIP supports multibit-serial multiplication and sparse data operations by leveraging register files as lookup tables or accumulators. The VVIP, verified on a vector processor, has an area overhead of less than 2.8%. It exhibits an average computation rate that is 10.1 times faster than the 1-bit-serial multiplication in linear algebra benchmarks, and 1.2 times average performance improvement in unstructured sparse point-wise convolution tasks when compared to conventional control sequences.	Hyungjoon Bae (KAIST); Da Won Kim (Columbia University); Wanyeong Jung (KAIST)
553	DL2Fence: Integrating Deep Learning and Frame Fusion for Enhanced Detection and Localization of Refined Denial-of-Service in Large-Scale NoCs	This study introduces a refined Flooding Injection Rate-adjustable Denial-of-Service (DoS) model for Network-on-Chips (NoCs) and more importantly presents DL2Fence, a novel framework utilizing Deep Learning (DL) and Frame Fusion (2F) for DoS detection and localization. Two Convolutional Neural Networks models for classification and segmentation were developed to detect and localize DoS respectively. It achieves detection and localization accuracies of 95.8% and 91.7%, and precision rates of 98.5% and 99.3% in a 16x16 NoC. The framework's hardware overhead notably decreases by 76.3% when scaling from 8x8 to 16x16, and it requires 42.4% less hardware compared to state-of-the-arts. This advancement demonstrates DL2Fence's effectiveness in balancing outstanding detection performance in large-scale NoCs with extremely low hardware overhead.	Haoyu Wang (University of Southampton); Basel Halak (University of Southampton); Jianjie Ren (University of Southampton); Ahmad Atamli (University of Southampton)
558	InterArch: Video Transformer Acceleration via Inter-Feature Deduplication with Cube-based Dataflow	In the realm of video-oriented tasks, Video Transformer models (VidT), an evolution from vision Transformers (ViT), have demonstrated considerable success. However, their widespread application is constrained by substantial computational demands and high energy consumption. Addressing these limitations and thus improving VidT efficiency has become a hot topic. Current methodologies solve this challenge by dividing a video into several features and applying intra-feature sparsity. However, they neglect the crucial point of inter-feature redundancy and often entail prolonged latency in fine-tuning phases. In response, this paper introduces InterArch, a tailored framework designed to significantly enhance VidT efficiency. We first design a novel inter-feature sparsity algorithm consisting of hierarchical deduplication and recovery. The deduplication phase capitalizes on temporal similarities at both block and element levels, enabling the elimination of redundant computations across features in both coarse-grained and fine-grained manners. To prevent long-latency fine-tuning, we employ a lightweight recovery mechanism that constructs approximate features for the sparsified data. Furthermore, InterArch incorporates a regular dataflow strategy, which consolidates sparse features and effectively translates sparse computations into dense ones. Complementing this, we develop a spatial array architecture equipped with augmented processing elements (PEs), specifically optimized for our proposed dataflow. Extensive experiment results demonstrate that InterArch can achieve satisfactory performance speedups and energy saving.	Xuhang Wang (Shanghai Jiao Tong University); Zhuoran Song (Shanghai Jiao Tong University); Xiaoyao Liang (Shanghai Jiao Tong University)
562	QuGeo: An End-to-end Quantum Learning Framework for Geoscience --- A Case Study on Full-Waveform Inversion	The rapid advancement of quantum computing has generated considerable anticipation for its transformative potential. However, harnessing its full potential relies on identifying "killer applications". In this regard, QuGeo emerges as a groundbreaking quantum learning framework, poised to become a key application in geoscience, particularly for Full-Waveform Inversion (FWI). This framework integrates variational quantum circuits with geoscience, representing a novel fusion of quantum computing and geophysical analysis. This synergy unlocks quantum computing's potential within geoscience. It addresses the critical need for physics-guided data scaling, ensuring high-performance geoscientific analyses aligned with core physical principles. Furthermore, QuGeo's introduction of a quantum circuit custom-designed for FWI highlights the critical importance of application-specific circuit design for quantum computing. In the OpenFWI's FlatVelA dataset experiments, the variational quantum circuit from QuGeo, with only 576 parameters, achieved significant improvement in performance. It reached a Structural Similarity Image Metric (SSIM) score of 0.905 between the ground truth and the output velocity map. This is a notable enhancement from the baseline design's SSIM score of 0.800, which was achieved without the incorporation of physics knowledge.	Weiwen Jiang (George Mason University); Youzuo Lin (University of North Carolina at Chapel Hill)
565	CLUMAP: Clustered Mapper for CGRAs with Predication	Coarse-grained reconfigurable architectures (CGRAs) have gained popularity as accelerators for compute-intensive kernels. Complex CGRA architectures that support key features such as multi-context and predication are being developed to support a wider range of kernels. However, mapping applications on these complex architectures poses significant challenges. In this paper, we provide an architecture-agnostic clustered mapping technique and a new cost function tailored for simulated-annealing placement. The mapper simplifies placement and routing phases, demonstrating significant speedup for popular CGRA architectures: HyCUBE and ADRES. Additionally, our method demonstrates an increase in mapping success for the ADRES architecture.	Omar Hassan Ali Ragheb Ismail (University of Toronto); Jason H. Anderson (University of Toronto)
569	MoteNN: Memory Optimization via Fine-grained Scheduling for Deep Neural Networks on Tiny Devices	There has been a growing trend in deploying deep neural networks (DNNs) on tiny devices. However, deploying DNNs on such devices poses significant challenges due to the contradiction between DNNs' substantial memory requirements and the stringent memory constraints of tiny devices. Some prior works incur large latency overhead to save memory and target only simple CNNs, while others employ coarse-grained scheduling for complicated networks, leading to limited memory footprint reduction. This paper proposes MoteNN that performs fine-grained scheduling via operator partitioning on arbitrary DNNs to dramatically reduce peak memory usage with little latency overhead. MoteNN presents a graph representation named Axis Connecting Graph (ACG) to perform operator partition at graph-level efficiently. MoteNN further proposes an algorithm that finds the partition and schedule guided by memory bottlenecks. We evaluate MoteNN using various popular networks and show that MoteNN achieves up to 80% of peak memory usage reduction compared to state-of-art works with nearly no latency overhead on tiny devices.	Renze Chen (Peking University); Zijian Ding (University of California, Los Angeles); Size Zheng (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University); Yun (Eric) Liang (Peking University)
575	An IP-Agnostic Foundational Cell Array Offering Supply Chain Security	Growing IC manufacturing complexity and reliance on third-party fabrication create supply chain fragility, contributing to chip shortages and IP security risks. General-purpose ICs can mitigate manufacturing security risks but rely on rely on software-based configurations, which is not optimal for high-consequence applications. Our work proposes a novel IP-agnostic Foundational Cell Array (FC-Array) platform to overcome these challenges. Built on only verified standard cells and industry-standard EDA tools, this platform preserves many advantages of an ASIC. By incorporating 3D split manufacturing, we provide semantically secure IP protection and a base wafer that can be stockpiled. Our tests demonstrate both power-efficient (100 MHz) and high-performance (1 GHz) options. In a post-place-and-route simulated 28nm design, our FC-Array shows a worst-case 1.85x increase in power consumption and a 2.61x increase in area compared to standard cell ASICs for equivalent timing performance.	Christopher Michael Talbot (Carnegie Mellon University); Deepali Garg (Carnegie Mellon University); Lawrence Pileggi (Carnegie Mellon University); Kenneth Mai (Carnegie Mellon University)

Submissio	Title	Abstract	Authors with Affiliations
577	Fast Virtual Gate Extraction For Silicon Quantum Dot Devices	Silicon quantum dot devices stand as promising candidates for large-scale quantum computing due to their extended coherence times, compact size, and recent experimental demonstrations of sizable qubit arrays. Despite the great potential, controlling these arrays remains a significant challenge. This paper introduces a new virtual gate extraction method to quickly establish orthogonal control on the potentials for individual quantum dots. Leveraging insights from the device physics, the proposed approach significantly reduces the experimental overhead by focusing on crucial regions around charge state transition. Furthermore, by employing an efficient voltage sweeping method, we can efficiently pinpoint these charge state transition lines and filter out erroneous points. Experimental evaluation using real quantum dot chip datasets demonstrates a substantial 5.84x to 19.34x speedup over conventional methods, thereby showcasing promising prospects for accelerating the scaling of silicon spin qubit devices.	Shize Che (University of Pennsylvania); Seongwoo Oh (University of Pennsylvania); Haoyun Qin (University of Pennsylvania); Yuhao Liu (University of Pennsylvania); Anthony Sigillito (University of Pennsylvania); Gushu Li (University of Pennsylvania)
581	G-kway: Multilevel GPU-Accelerated k-way Graph Partitioner	Graph partitioning is fundamental for many CAD algorithms because it divides a large circuit into smaller pieces with manageable complexity. As the size of the circuit graph continues to grow, graph partitioning becomes increasingly time-consuming. Recent research has introduced parallel graph partitioners using either multi-core CPUs or GPUs. However, their performance is limited to a few CPU cores and available GPU memory. As a result, we propose G-kway, an efficient multilevel GPU-accelerated k-way graph partitioner. Experimental results have shown that G-kway outperforms both the state-of-the-art CPU-based and GPU-based parallel partitioners with an average speedup of 8.6x and 3.8x, respectively.	Wan Luan Lee (University of Wisconsin-Madison); Dian-Lun Lin (University of Wisconsin-Madison); Tsung-Wei Huang (University of Wisconsin at Madison); Shui Jiang (The Chinese University of Hong Kong); Tsung-Yi Ho (The Chinese University of Hong Kong); Yibo Lin (Peking University); Bei Yu (The Chinese University of Hong Kong)
584	ICGMM: CXL-enabled Memory Expansion with Intelligent Caching Using Gaussian Mixture Model	The memory wall is a growing issue in modern computing systems due to the disparity between device computing power and data communication speed. To alleviate memory wall, Compute Express Link (CXL) is proposed to create a shared and coherent memory space between the host and device, offering opportunities to use device DRAM as cache and device memory as primary storage for memory-intensive tasks. However, challenges arise when utilizing device DRAM as cache, including high cache miss penalties caused by data access granularity mismatches and inefficient hardware cache management mechanisms. To tackle these issues, we propose Smart DRAM-Caching, an efficient framework that employs Gaussian Mixture Model (GMM) for intelligent caching and eviction on hardware. Compared with traditional cache replacement strategies LRU, our on-board measurements reveal that a ?% increase in cache hit rate can result in a ?% reduction in average device memory access latency. Compared with learning-based methods like LSTM, our approach achieves ?× speedup with less hardware resource consumption.	Hanqiu Chen (Georgia Institute of Technology); Yitu Wang (Duke University); Luis Vitorio Cargnini (Samsung Semiconductor); Mohammadreza Soltaniyeh (Samsung Semiconductor); Dongyang Li (Samsung Semiconductor); Gongjin Sun (Samsung Semiconductor); Pradeep Subedi (Samsung Semiconductor); Andrew Chang (Samsung Semiconductor); Yiran Chen (Duke University); Cong Hao (Georgia Institute of Technology)
588	APTQ: Attention-aware Post-Training Mixed-Precision Quantization for Large Language Models	Large Language Models have greatly advanced the natural language processing paradigm. However, the high computational load and huge model sizes pose a grand challenge for deployment on edge devices. To this end, we propose APTQ (Attention-aware Post-Training Mixed-Precision Quantization) for LLMs, which considers not only the second-order information of each layer's weights, but also, for the first time, the nonlinear effect of attention outputs on the entire model. We leverage the Hessian trace as a sensitivity metric for mixed-precision quantization. Experiments show APTQ attains state-of-the-art zero-shot accuracy of 68.24% and 70.48% at 3.8 bitwidth in LLaMa-7B and LLaMa-13B, respectively.	Ziyi Guan (The University of Hong Kong); Hantao Huang (Southern University of Science and Technology); Yupeng Su (Southern University of Science and Technology); Hong Huang (Southern University of Science and Technology); Ngai Wong (The University of Hong Kong); Hao Yu (Southern University of Science and Technology)
593	Engineering an Efficient Preprocessor for Model Counting	Given a formula F, the problem of model counting is to compute the number of solutions (also known as models) of F. Over the past decade, model counting has emerged as key building block of quantitative reasoning in design automation and artificial intelligence. Given the wide ranging applications, scalability remains the major challenge in the development of model counters. Motivated by the observation that the formula simplification can dramatically impact the performance of the state of the art exact model counters, we design a new state of the art preprocessor, Puura, that relies on tight integration of techniques. The design of Puura is motivated from our observation that it is often beneficial to employ preprocessing techniques whose overhead may be prohibitive for the task of SAT solving but not for model counting: accordingly, we rely on a specifically tailored SAT solver design for redundancy detection, sampling-boosted backbone detection, as well as storing of redundancy information for the purposes of improving propagation within top-down model counters. Our detailed empirical evaluation demonstrates that Puura achieves significant performance improvements over prior model counting preprocessors in terms of instance-size reductions achieved as well as the runtime improvements of the downstream model counters.	Mate Soos (Ethereum Foundation); Kuldeep S Meel (University of Toronto)
607	CDA-GNN: A Chain-driven Accelerator for Efficient Asynchronous Graph Neural Network	Asynchronous Graph Neural Network (AGNN) has attracted much research attention because it enables faster convergence speed than the synchronous GNN. However, existing software/hardware solutions suffer from redundant computation overhead and excessive off-chip communications for AGNN due to irregular state propagations along the dependency chains between vertices. This paper proposes a chain-driven asynchronous accelerator, CDA-GNN, for efficient AGNN inference. Specifically, CDA-GNN proposes a chain-driven asynchronous execution approach into novel accelerator design to regularize the vertex state propagations for fewer redundant computations and off-chip communications, and also designs a chain-aware data caching method to improve data locality for AGNN. We have implemented and evaluated CDA-GNN on a Xilinx Alveo U280 FPGA card. Compared with the state-of-the-art software solutions (i.e., Dorylus and AMP) and hardware solutions (i.e., BlockGNN and FlowGNN), CDA-GNN improves the performance of AGNN inference by an average of 1,173x, 182.4x, 10.2x, and 7.9x and saves energy by 2,241x, 242.2x, 12.4x, and 8.9x, respectively.	Hui Yu (Huazhong University of Science and Technology); Yu Zhang (Huazhong University of Science and Technology); Ligang He (University of Warwick); Donghao He (Huazhong University of Science and Technology); Qikun Li (Huazhong University of Science and Technology); Jin Zhao (Huazhong University of Science & Technology/Zhejiang Lab); Xiaofei Liao (Huazhong University of Science and Technology); Hai Jin (Huazhong University of Science and Technology); Lin Gu (Huazhong University of Science and Technology); Haikun Liu (Huazhong University of Science and Technology)
614	Towards Cost-Effective High-Throughput End Station Design for Time-Sensitive Networking (TSN)	Time-Sensitive Networking (TSN) technology has been increasingly deployed in mission- and safety-critical industrial applications to achieve high throughput and deterministic communications. To provide stringent timing guarantee, TSN requires that network devices follow a predefined communication schedule for real-time end-to-end packet processing, involving both TSN bridges and end stations. Extensive efforts have been devoted on the TSN bridge design in the literature. Achieving TSN compatibility on the end stations (especially on Commercial Off-The-Shelf (COTS) hardware), however is challenging due to the inefficiencies of general CPU and unpredictable bus contention. To fill this gap, this work presents a software-based open-source approach that i) enables nanosecond-level packet transmission accuracy based on DPDK, and ii) employs a novel multi-core scheduling algorithm to boost the throughput of real-time TSN traffic. Our proposed solution leverages existing COTS hardware and thus is more generic and cost-effective compared to existing hardware-centric solutions. We validate our design by developing a prototype end station and incorporating it within an eight-bridge TSN network testbed. Our extensive experiments demonstrate the efficiency and effectiveness of our design at both device and system levels.	Chuanyu Xue (University of Connecticut); Tianyu Zhang (University of Connecticut); Song Han (University of Connecticut)

Submissio	Title	Abstract	Authors with Affiliations
615	RTGA: A Redundancy-free Accelerator for High-Performance Temporal Graph Neural Network Inference	Temporal Graph Neural Network (TGNN) has attracted much research attention because it can capture dynamic nature of complex networks. However, existing software/hardware solutions suffer from and redundant computation overhead and excessive off-chip communications for TGNN due to they need to recompute identical messages and unnecessarily updates the vertex memory of unaffected vertices. This paper proposes a redundancy-free accelerator, RTGA, for highperformance TGNN inference. Specifically, RTGA proposes a redundancy-aware execution approach with temporal tree into novel accelerator design to effectively eliminate unnecessary data processing for fewer redundant computations and off-chip communications, and also designs a temporal-aware data caching method to improve data locality for TGNN. We have implemented and evaluated RTGA on a Xilinx Alveo U280 FPGA card. Compared with the state-of-the-art software solutions (i.e., TGN and TGL) and hardware solutions (i.e., BlockGNN and FlowGNN), RTGA improves the performance of TGNN inference by an average of 473.2x, 87.4x, 8.2x, and 6.9x and saves energy by 542.8x, 102.2x, 9.4x, and 8.3x, respectively.	Hui Yu (Huazhong University of Science and Technology); Yu Zhang (Huazhong University of Science and Technology); Andong Tan (Huazhong University of Science and Technology); Chenze Lu (Huazhong University of Science and Technology); Jin Zhao (Huazhong University of Science and Technology); Xiaofei Liao (Huazhong University of Science and Technology); Hai Jin (Huazhong University of Science and Technology); Haikun Liu (Huazhong University of Science and Technology)
616	Advanced Reinforcement Learning Algorithms to Optimize Design Verification	Given the increasing complexity of integrated circuits, the utilization of machine learning in simulation-based hardware design verification (DV) has become crucial to ensure comprehensive coverage of hard-to-hit states. Our paper proposes a deep deterministic policy gradient (DDPG) algorithm combined with prioritized experience replay (PER) to determine the stimulus settings that result in the highest average FIFO depth in a modified exclusive shared invalid (MESI) cache controller architecture. This architecture includes four FIFOs, each corresponding to a distinct CPU. Through extensive experimentation, DDPG coupled with PER (DDPG-PER) proves to be more effective than DDPG with uniform experience replay in enhancing average FIFO depth and coverage within the DV process. Furthermore, our proposed DDPG-PER framework significantly increases the occurrence of higher FIFO depths, thereby addressing the challenges associated with reaching hard-to-hit states in DV. The proposed DDPG-PER and DDPG algorithms also demonstrate a larger average FIFO depth over four CPUs, requiring considerably less execution time than Bayesian Optimization (BO).	Zahra Aref (Rutgers, The State University of New Jersey); Rohit Suvarna (VerifAI Inc.); Bill Hughes (VerifAI Inc.); Sandeep Srinivasan (VerifAI Inc.); Narayan B. Mandayam (Rutgers, The State University of New Jersey)
618	Thermal Resistance Network Derivative (TREND) Model for Efficient Thermal Simulation and Design of ICs and Packages	In the thermal design of 3-D integrated circuits (ICs) and packages, numerical simulation is extensively employed to investigate the impact of model parameters on hotspot temperature. However, conventional simulation approaches usually require plenty of computational resource and thus lead to expensive time cost for thermal designs. In this paper, we present a novel technique to efficiently and accurately conduct thermal simulation of 3-D ICs and packages, potentially reducing thermal design timeline from weeks to minutes. The proposed thermal resistance network derivative (TREND) model facilitates to focus the solution domain on the crucial regions for thermal designs and accelerate simulation without sacrificing accuracy. Also, the TREND model protects the internal details of chips and packages, which is quite suitable for modular thermal designs. The flexibility, accuracy, and efficiency of the proposed method are demonstrated through several numerical examples. Compared with commercial software, a speed-up of 2695x is achieved in a typical thermal design case without the loss of accuracy.	Shunxiang Lan (Shanghai Jiao Tong University); Min Tang (Shanghai Jiao Tong University); Liang Chen (Shanghai University); Junfa Mao (Shanghai Jiao Tong University)
622	Geneva: A Dynamic Confluence of Speculative Execution and In-Order Commitment Windows	Modern out-of-order processors are increasingly expanding resources such as reorder buffer (ROB) and instruction queue (IQ) for memory-level parallelism (MLP). While this expansion effectively addresses the memory wall challenge, it also incurs notable cost and energy trade-offs. To tackle this, we propose Geneva, a microarchitecture that improves performance and energy efficiency. Geneva reallocates a portion of the ROB to serve as a dynamic queue (DQ), used as the ROB, IQ, or both depending on operational needs. Geneva saves energy by 15.6% and improves performance by 2.6% compared to the out-of-order core baseline.	Yanghee Lee (Yonsei University); Jiwon Lee (Yonsei University); Jaewon Kwon (Yonsei University); Yongju Lee (Yonsei University); Won Woo Ro (Yonsei University)
626	Lightator: An Optical Near-Sensor Accelerator with Compressive Acquisition Enabling Versatile Image Processing	This paper proposes a high-performance and energy-efficient optical near-sensor accelerator for vision applications, called Lightator. Harnessing the promising efficiency offered by photonic devices, Lightator features innovative compressive acquisition of input frames and fine-grained convolution operations for low-power and versatile image processing at the edge for the first time. This will substantially diminish the energy consumption and latency of conversion, transmission, and processing within the established cloud-centric architecture as well as recently designed edge accelerators. Our device-to-architecture simulation results show that with favorable accuracy, Lightator achieves 84.4 Kilo FPS/W and reduces power consumption by a factor of ~24x and 73x on average compared with existing photonic accelerators and GPU baseline	Mehrdad Morsali (New jersey Institute of Technology); Brendan C. Reidy (University of South Carolina); Deniz Najafi (New Jersey Institute of Technology); Sepehr Tabrizchi (University of Nebraska–Lincoln); Mohsen Imani (University of California Irvine); Mahdi Nikdast (Colorado State University); Arman Roohi (University of Nebraska - Lincoln); Ramtin Zand (University of South Carolina); Shaahin Angizi (New Jersey Institute of Technology)
637	Lesyn: Placement-aware Logic Resynthesis for Non-Integer Multiple-Cell-Height Designs	Non-integer multiple cell height (NIMCH) standard-cell libraries offer promising co-optimization for power, performance and area in advanced technology nodes. However, such non-uniform design introduces new layout constraints where any sub-region can only accommodate gates of the same cell height. The existing physical design flow for NIMCH circuits handles the constraint by clustering and relocating gates according to their cell heights, inevitably causing displacement that harms circuit performance. This paper proposes a placement-aware logic resynthesis procedure that explicitly adjusts cell heights after initial placement without changing cell location. Experiment results demonstrate that our approach can reduce the maximal delay by 26.1%.	Yuan Pu (The Chinese University of Hong Kong); Fangzhou Liu (The Chinese University of Hong Kong); Yu Zhang (The Chinese University of Hong Kong); Zhuolun He (The Chinese University of Hong Kong); Yibo Lin (Peking University); Kai-Yuan Chao (Siemens Digital Industries Software); Bei Yu (The Chinese University of Hong Kong)
638	Effectively Sanitizing Embedded Operating Systems	Embedded operating systems, considering their widespread use in security-critical applications, are not effectively tested with sanitizers to effectively root out bugs. Sanitizers provide a means to detect bugs that are not visible directly through exceptional or erroneous behaviors, thus uncovering more potent bugs during testing. In this paper, we propose EmbSan, an embedded systems sanitizer for a diverse range of embedded operating system firmware through the use of dynamic instrumentation of sanitizer facilities and de-coupled on-host runtime libraries. This allows us to perform sanitation for multiple embedded OSs during fuzzing, such as many Embedded Linux-based firmware, various FreeRTOS firmware, and detect actual bugs within them. We evaluated EmbSan's effectiveness on firmware images based on Embedded Linux, FreeRTOS, LiteOS, and VxWorks. Our results show that EmbSan can detect the same criteria of actual bugs found in the Embedded Linux kernel as reference implementations of KASAN, and exhibits a slowdown of 2.2× to 3.2× and 5.2× to 5.7× for KASAN and KCSAN, respectively, which is on par with established kernel sanitizers. EmbSan and embedded OS fuzzers also found a total of 41 new bugs in Embedded Linux, FreeRTOS, LiteOS and VxWorks.	Jianzhong Liu (Tsinghua University); Yuheng Shen (Tsinghua University); Yiru Xu (Tsinghua University); Hao Sun (ETH Zurich); Heyuan Shi (Central South University); Yu Jiang (Tsinghua University)

Submissio	Title	Abstract	Authors with Affiliations
639	SPFuzz: Stateful Path based Parallel Fuzzing for Protocols in Autonomous Vehicles	Protocols in autonomous vehicles are essential for efficient in-vehicle network communication. To ensure their security, many research efforts have been paid to the fuzz testing of their implementations. However, those fuzzing optimizations often struggle to manage the protocols' complex state, resulting in low efficiency in branch covering and vulnerability detection. This paper introduces SPFuzz, a stateful path based parallel fuzzing framework to improve the testing performance of protocols in autonomous vehicles. The basic idea is to accelerate fuzzing speed by dividing tasks to reduce conflicts and dispatching them on different fuzzing instances. SPFuzz first leverages protocol state and data models to generate stateful paths, then divides them into discrete tasks and dispatches them based on their complexity and diversity, ensuring a balanced workload distribution across all fuzzing instances. For evaluation, we implement SPFuzz on top of the state-of-the-art protocol fuzzer Peach and conduct experiments on four prominent vehicle protocols, including ZMTP, MQTT, DDS, and DoIP. The results show that, compared to the original parallel mode of Peach, SPFuzz achieves the same code coverage at a speed of 2.8X-473.2X, with 5.52% more branch coverage within 24 hours. SPFuzz uncovered six previously unknown vulnerabilities in those heavily tested protocol implementations, with four CVEs assigned in the national vulnerability database. Additionally, SPFuzz has been adapted to ECUs from several vendors, such as NISSAN, and triggered a total of four vulnerabilities that may cause system crashes.	Junze Yu (Tsinghua University); Zhengxiong Luo (Tsinghua University); Fangshangyuan Xia (Central South University); Yanyang Zhao (Tsinghua University); Heyuan Shi (Central South University); Yu Jiang (Tsinghua University)
641	PDRC: Package Design Rule Checking via GPU-Accelerated Geometric Intersection Algorithms for Non-Manhattan Geometry	With the emergence of chiplet technology, the scale of IC packaging design has been steadily increasing, making the utilization of traditional design rule checking (DRC) methods more time-consuming. In this paper, we propose PDRC, a package-level design rule checker for non-manhattan geometry with GPU acceleration. PDRC employs hierarchical interval lists within an iterative parallel sweepline framework to implement the geometric intersection algorithm, thereby finishing design rule checking tasks. Experimental results have demonstrated 30 - 50 times speedup achieved by PDRC compared with two CPU-based checkers.	Jiaxi Jiang (The Chinese University of Hong Kong); Lancheng Zou (The Chinese University of Hong Kong); Wenqian Zhao (The Chinese University of Hong Kong); Zhuolun He (The Chinese University of Hong Kong); Tinghuan Chen (The Chinese University of Hong Kong, Shenzhen); Bei Yu (The Chinese University of Hong Kong)
649	TIGA: Towards Efficient Near Data Processing in SmartNICs-based Disaggregated Memory Systems	Memory disaggregation, facilitated by SmartNICs, has emerged as a cost-effective approach for sharing memory resources in data centers. However, current SoC-based SmartNICs face several challenges for supporting near-data processing (NDP) in DM systems effectively. To address these challenges, we propose TIGA, an efficient NDP framework for SmartNICs-based DM systems. We propose an adaptive resource allocator to fully utilize SoC cores, and a SmartNIC-CPU cooperative mechanism to schedule NDP tasks. We prototype TIGA with FPGAs and evaluate it with typical workloads. Experimental results show that TIGA significantly improves the efficiency of NDP tasks in DM systems.	Zhuohui Duan (Huazhong University of Science and Technology); Zelin Yu (Huazhong University of Science and Technology); Haikun Liu (Huazhong University of Science and Technology); Xiaofei Liao (Huazhong University of Science and Technology); Hai Jin (Huazhong University of Science and Technology); Shijie Zheng (Huazhong University of Science and Technology); Sihan Wu (Huazhong University of Science and Technology)
651	LLM-HD: Layout Language Model for Hotspot Detection with GDS Semantic Encoding	With the rapid downscaling of technology nodes, industrial flow such as pitch reduction, patterning flexibility, and lithography processing variability have been challenged. Layout hotspot detection is one of the most challenging and critical steps, which requires technology upgrading. Pattern matching and learning-based detectors are proposed as quick detection methods. However, these computer vision (CV) model-based detectors use images transformed from layout GDS files as their inputs. It leads to foreground information (e.g. metal polygons) loss and even distortion when shrinking the image size to fit the model input. Moreover, plenty of irrelevant background information such as non-polygon pixels are also fed into the model, which hinders the fitting of the model and results in a waste of computational resources. Concerning the disadvantage of the traditional CV model, we propose a new layout hotspot detection paradigm, which directly detects hotspots on GDS files by exploiting a hierarchical GDS semantic representation scheme and a well-designed pre-trained natural language processing (NLP) model. Compared with state-of-the-art works, ours achieves better results both on the ICCAD2012 metal layer benchmark and the more challenging ICCAD2020 via layer benchmark, which demonstrates the effectiveness and efficiency of our approach.	yuyang chen (Shanghaitech University); Yiwen Wu (ShanghaiTech University); Jingya Wang (ShanghaiTech University); Tao Wu (ShanghaiTech University); Xumin He (ShanghaiTech University); Jingyi Yu (ShanghaiTech University); Hao Geng (ShanghaiTech University)
652	Combining Parameterized Pulses and Contextual Subspace for More Practical VQE	We explore the integration of parameterized quantum pulses with the contextual subspace method. The advent of parameterized quantum pulses marks a transition from quantum gates to a more efficient approach. Working with pulses allows us to potentially access areas of the Hilbert space that are inaccessible with a CNOT-based circuit decomposition. Compared to the traditional Variational Quantum Eigensolver (VQE), the computation of the contextual correction generally requires fewer qubits and measurements, thus improving computational efficiency. Plus a Pauli grouping strategy, our framework can minimize the quantum resource cost for the VQE and enhance the potential for processing larger molecular structures.	Zhiding Liang (University of Notre Dame); Zhixin Song (Georgia Institute of Technology); Jinglei Cheng (Purdue University); Hang Ren (University of California, Berkeley); Tianyi Hao (University of Wisconsin-Madison); Rui Yang (Peking University); Yiyu Shi (University of Notre Dame); Tongyang Li (Peking University)
653	Partially-Structured Transformer Pruning with Patch-Limited XOR-Gate Compression for Stall-Free Sparse-Model Access	The pruning-based model compression is regarded as an essential technique to deploy the recent large-size transformer models in practical services; however, accessing sparse transformer models cannot reach the ideal speed at all due to the frequent memory stalls for the irregular memory-accessing patterns. Based on the recent XOR-gate compression relaxing the amount of irregular accesses, this work presents a novel partially-structured transformer pruning method dedicated to the interface-friendly compression format. The stall-free memory access is firstly derived by limiting the number of patches per weight, introducing a new trade-off between model quality and effective memory bandwidth. Then, the partially-structured pruning patterns are deployed to provide better accuracy-bandwidth trade-off by significantly reducing the number of correction patches. Adjusting the patch distribution per weight in an aggressive way, the number of limited patches can be even smaller than that of weight bits, further increasing the effective bandwidth for achieving the similar model accuracy. We demonstrate the proposed stall-free XOR-gate compression schemes at pruned DeiT/BERT models on ImageNet/SQuAD datasets, presenting the highest effective bandwidth for accessing sparse transformers compared to the existing stall-based solutions.	Younghoon Byun (POSTECH); Youngjoo Lee (Pohang University of Science and Technology (POSTECH))

Submissio	Title	Abstract	Authors with Affiliations
655	E-Syn: E-Graph Rewriting with Technology-Aware Cost Functions for Logic Synthesis	Logic synthesis plays a crucial role in the digital design flow. It has a decisive influence on the final Quality of Results (QoR) of the circuit implementations. However, existing multi-level logic optimization algorithms often employ greedy approaches with a series of local optimization steps. Each step breaks the circuit into small pieces (e.g., k-feasible cuts) and applies incremental changes to individual pieces separately. These local optimization steps could limit the exploration space and may miss opportunities for significant improvements. To address the limitation, this paper proposes using e-graph in logic synthesis. The new workflow, named E-Syn, makes use of the well-established e-graph infrastructure to efficiently perform logic rewriting. It explores a diverse set of equivalent Boolean representations while allowing technology-aware cost functions to better support delay-oriented and area-oriented logic synthesis. Experiments over a wide range of benchmark designs show our proposed logic optimization approach reaches a wider design space compared to the commonly used AIG-based logic synthesis flow. It achieves on average 15.29% delay saving in delay-oriented synthesis and 6.42% area saving for area-oriented synthesis.	Chen Chen (The Hong Kong University of Science and Technology(Guangzhou)); Guangyu Hu (The Hong Kong University of Science and Technology); Dongsheng Zuo (The Hong Kong University of Science and Technology (Guangzhou)); Cunxi Yu (University of Maryland, College Park); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou)); Hongce Zhang (Hong Kong University of Science and Technology (Guangzhou))
656	Balloon-ZNS: Constructing High-Capacity and Low-Cost ZNS SSDs with Built-in Compression	This paper proposes Balloon-ZNS that enables transparent compression in emerging storage devices ZNS SSDs to enhance cost efficiency. ZNS SSDs require data pages to be stored and aligned in logical zones and flash blocks, conflicting with the management of variable-length compressed pages. Motivated by an observation that compressibility locality widely exists in data streams, Balloon-ZNS performs compressibility-adaptive, slot-aligned storage management to address the conflict. Evaluation with RocksDB shows Balloon-ZNS can reap more than 80% of the compression gain while achieving -7% to 14% higher throughput than a vanilla ZNS SSD, on average, when data compressibility is not poor.	Yu Wang (Huazhong University of Science and Technology); Zibin Sun (Huazhong University of Science and Technology); You Zhou (Huazhong University of Science and Technology); Tao Lu (DapuStor Corporation); Changsheng Xie (Huazhong University of Science and Technology); Fei Wu (Huazhong University of Science and Technology)
676	Optimal Transistor Folding and Placement for Synthesizing Standard Cells of Complementary FET Technology	As the VLSI technology continues to scale beyond 5nm, a strong demand on the continuing layout reduction of standard cells is required. However, the standard cells with conventional FinFET or nanosheet-FET structure are becoming much hard to meet this requirement due to the lateral P-FET and N-FET separation. It has been widely accepted that Complementary-FET (CFET) is a promising technology, which stacks P-FET on N-FET or vice versa, to achieve this objective. In comparison with synthesizing the conventional FET based standard cells, two prominent optimization tasks in CFET based multi-row cell synthesis that significantly affect the cell quality, in terms of area and routability, are (1) determining transistor folding shapes and (2) determining placement order of transistors with fully secured vertical i.e., z-directional routing space on the stacked FETs as well as buried power rail (BPR). In this work, we propose an optimal solution to the combined problem of tasks 1 and 2. Precisely, we develop a search tree-based area-optimal method of transistor folding and placement, in which we accelerate the cost computation of partial solutions by formulating it into dynamic programming while performing a strict feasibility checking of securing in-cell vertical routing space of partial solutions by formulating and solving it into an instance of network flow problem. Through experiment with benchmark circuits, it is shown that the CFET cells produced by our cell synthesizer are 14% smaller in size on average even with 31% shorter total metal length and 52% less use of metal2 for in-cell routing over the cells produced by the recent state-of-the-art CFET cell generator.	Suwan Kim (Seoul National University); Taewhan Kim (Seoul National University)
677	DACPara: A Divide-and-Conquer Parallel Approach for High-Quality Logic Rewriting in Large-Scale Circuits	Logic rewriting is a critical and time-consuming task in logic synthesis, which determines the area and delay of the synthesized circuit. However, existing parallel solutions for this task suffer from limitations in terms of runtime or quality in large-scale complex circuits. In this paper, we propose a divide-and-conquer parallel approach namely DACPara for high-quality logic rewriting in large-scale circuits. Specifically, after nodes in AIG are divided in each level, dynamic global information is considered to divide and conquer rewriting into three stages for parallel processing. Experiments show that DACPara using 40 CPU physical cores can be 34.36x and 1.96x faster than logic rewriting in ABC and the state-of-the-art CPU parallel method on large benchmarks, respectively, with extremely comparable quality of result. Also, for large-scale complex benchmarks, compared with state-of-the-art GPU accelerated method ours can achieve 1.1% quality improvement.	Nanjiang Qu (Xidian University); Cong Tian (Xidian University); Zhenhua Duan (Xidian University)
682	Double-Win NAS: Towards Deep-to-Shallow Transformable Neural Architecture Search for Intelligent Embedded Systems	Thanks to the evolving network depth, convolutional neural networks (CNNs) have achieved impressive performance across various intelligent embedded scenarios towards embedded intelligence. Nonetheless, this trend also leads to degraded hardware efficiency as the network evolves deeper and deeper. In contrast, shallow networks exhibit superior hardware efficiency, which, unfortunately, suffer from inferior accuracy. To tackle this dilemma, we establish the first deep-to-shallow transformable neural architecture search (NAS) paradigm, namely Double-Win NAS (DW-NAS), which is dedicated to automatically exploring deep-to-shallow transformable networks to marry the best of both worlds. Extensive experiments on two NVIDIA Jetson intelligent embedded systems clearly show the superiority of DW-NAS over previous state-of-the-art methods.	Xiangzhong Luo (Nanyang Technological University); Di Liu (Norwegian University of Science and Technology); Hao Kong (Nanyang Technological University); Shuo Huai (Nanyang Technological University); Weichen Liu (Nanyang Technological University)
697	CAMPER: Exploring the Potential of Content Addressable Memory for 3D Point Cloud Efficient Range Search	Range search is the key part of the point cloud processing pipeline. CAM has proven its efficiency for search tasks on switches. In this work, we propose CAMPER, aiming to explore the potential of CAM for point cloud range search. We developed a ripple comparison 13T CAM cell for distance comparison, designed a spatial approximation search algorithm based on Chebyshev distance, and discussed the flexibility and scalability of the architecture. The results show that in the 64k@64k task, CAMPER achieves a latency of 0.83ms and a power consumption of 114.6mW, increased by 10.4x and 228x, respectively.	Jiapei Zheng (Fudan University); Lizhou Wu (State Key Laboratory of Integrated Chipsand Systems, Frontier Institute of Chip and System, Fudan University); Yutong Su (Fudan University); Jingyi Wang (Fudan University); Zhangcheng Huang (Fudan University); Chixiao Chen (Fudan University); Qi Liu (Fudan University)
698	OTPlace-Vias: A Novel Optimal Transport Based Method for High Density Vias Placement in 3D Circuits	Three-dimensional integrated circuit (3D IC) is an important manufacturing technology. In particular, the Monolithic 3D (M3D) technology stands out as a cutting-edge approach that provides higher integration density. However, M3D also introduces several challenges in terms of high density and computational complexity. In this paper, we propose a new approach for solving the inter-tier vias placement problem through optimal transport, which can be efficiently implemented in parallel with GPUs and consequently achieves significant speedup. Moreover, comparing with previous methods, our approach can also facilitate the processing of high integration density circuits to be more effective.	Lin Chen (University of Science and Technology of China); Qi Xu (University of Science and Technology of China); Hu Ding (University of Science and Technology of China)
706	NeuroSelect: Learning to Select Clauses in SAT Solvers	Modern SAT solvers depend on conflict-driven clause learning to avoid recurring conflicts. Deleting less valuable learned clauses is a crucial component of modern SAT solvers to ensure efficiency. However, a single clause deletion policy cannot guarantee optimal performance on all SAT instances. This paper introduces a new clause deletion metric to diversify existing clause deletion approaches. Then, we propose to use machine learning to evaluate and select clause deletion policies adaptively based on the input instance. We show that our method can reduce the runtime of the state-of-the-art SAT solver Kissat by 5.8\% on large industry benchmarks.	Hongduo Liu (The Chinese University of Hong Kong); Peng Xu (The Chinese University of Hong Kong); Yuan Pu (The Chinese University of Hong Kong); Lihao Yin (Huawei Noah's Ark Lab); Hui-Ling Zhen (Huawei); Mingxuan Yuan (Huawei Noah's Ark Lab); Tsung-Yi Ho (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong)

Submissio	Title	Abstract	Authors with Affiliations
709	Accelerating DTCO with a Sample-Efficient Active Learning Framework for TCAD Device Modeling	Design-Technology Co-Optimization (DTCO) can be significantly accelerated by employing Neural Compact Models (NCMs). However, the effective deployment of NCMs requires a substantial amount of training data for accurate device modeling. This paper introduces an Active Learning (AL) framework designed to enhance the efficiency of both device modeling and process optimization, particularly addressing the challenges of time-intensive Technology Computer-Aided Design (TCAD) simulations. The framework employs a ranking algorithm that assesses metrics such as the expected variance from the neural tangent kernel (NTK), TCAD simulation time, and the complexity of I-V curves. This strategy considerably reduces the number of required simulations while maintaining high accuracy. Demonstrating the effectiveness of our AL framework, we achieved a 28.5\% improvement in MSE within a 30-minute time budget for device modeling, and an 86.7\% reduction in the data points required for process optimization of a 51-stage ring oscillator (RO). These results offer a streamlined, adaptable solution for rapid device modeling and process optimization in various DTCO applications.	Chanwoo Park (Alsemy Inc.); Junghwan Park (Alsemy Inc.); Premkumar Vincent (Alsemy Inc.); Hyunbo Cho (Alsemy Inc.)
712	Disentangle, Align and Generalize: Learning A Timing Predictor from Different Technology Nodes	In VLSI design, accurate pre-routing timing prediction is paramount. Traditional machine learning-based methods require extensive data, posing challenges for advanced technology nodes due to the time-consuming data preparation. To mitigate this issue, we propose a novel transfer learning framework that uses data from previous nodes for learning on the target node. Our method initially disentangles and aligns timing path features across different nodes, then predicts each path's arrival time employing a Bayesian-based model capable of handling highly variable arrival time and generalizing to new designs. Experimental results on transfer learning from 130nm to 7nm nodes validate our method's effectiveness.	Xinyun Zhang (The Chinese University of Hong Kong); Binwu Zhu (The Chinese University of Hong Kong); Fangzhou Liu (The Chinese University of Hong Kong); Ziyi Wang (The Chinese University of Hong Kong); Peng Xu (The Chinese University of Hong Kong); Hong Xu (CUHK); Bei Yu (The Chinese University of Hong Kong)
714	Fracturing-aware Curvilinear ILT via Circular E-beam Mask Writer	Inverse lithography technology (ILT) is vital in optical proximity correction, tending to generate curvilinear masks for optimal process windows. Traditional curvilinear mask manufacturing involves fracturing into rectangles, requiring expensive mask write times. A novel E-beam mask writer that writes variable radius circles per shot significantly reduces the shot count for curvilinear masks. We present two methods to generate circular fracturing-aware masks. The first one converts pixel-based masks from existing ILT methods into circle-based masks using predefined rules. The second one integrates circular constraints into the ILT process, generating circle-based masks directly via optimization. Extensive experimental results validate both approaches' effectiveness.	Xinyun Zhang (The Chinese University of Hong Kong); Su Zheng (The Chinese University of Hong Kong); Guojin Chen (The Chinese University of HongKong); Binwu Zhu (The Chinese University of Hong Kong); Hong Xu (CUHK); Bei Yu (The Chinese University of Hong Kong)
722	Performance-driven Analog Routing via Heterogeneous 3DGNN and Potential Relaxation	Analog routing is crucial for performance optimization in analog circuit design, but conventionally takes significant development time and requires design expertise. Recent research has attempted to use machine learning (ML) to generate guidance to preserve circuit performance after analog routing. These methods face challenges such as expensive data acquisition and biased guidance. In this paper, we introduce AnalogFold, a new paradigm of analog routing leveraging ML-enabled performance-oriented routing guidance. Our approach learns performance-driven routing guidance and uses it to help automatic routers for performance-driven routing optimization. We propose to use a 3DGNN that incorporates cost-aware distance to make accurate predictions on post-layout performance. A pool-assisted potential relaxation process derives the effective routing guidance. The experimental results on multiple benchmarks under the TSMC 40nm technology node demonstrate the superiority of the proposed framework compared to the cutting-edge works.	Peng Xu (The Chinese University of Hong Kong); Guojin Chen (The Chinese University of HongKong); Keren Zhu (The Chinese University of Hong Kong); Tinghuan Chen (The Chinese University of Hong Kong, Shenzhen); Tsung-Yi Ho (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong)
725	A High-Performance Stochastic Simulated Bifurcation Ising Machine	Ising model-based computers have recently emerged as high-performance solvers for combinatorial optimization problems (COPs). For Ising model, a simulated bifurcation (SB) algorithm searches for the solution by solving pairs of differential equations. The SB machine benefits from massive parallelism but suffers from high energy. Dynamic stochastic computing implements accumulation-based operations efficiently. This article proposes a high-performance stochastic SB machine (SSBM) for solving COPs with efficient hardware. To this end, we develop a stochastic SB (sSB) algorithm such that the multiply-and-accumulate (MAC) operation is converted to multiplexing and addition while the numerical integration is implemented by using signed stochastic integrators (SSIs). Specifically, the sSB stochastically ternarizes position values used for the MAC operation. A stochastic computing SB cell (SC-SBC) is constructed by using two SSIs for area efficiency. Additionally, a binary-stochastic computing SB cell (BSC-SBC) uses one binary integrator and one SSI to achieve a reduced delay. Based on sSB, an SSBM is then built by using the SC-SBC or BSC-SBC as the basic building block. The designs and syntheses of two SSBMs with 2000 fully connected spins require at least 1.13 times smaller area than the state-of-the-art designs.	Tingting Zhang (University of Alberta); Hongqiao Zhang (ShanghaiTech University); Zhengkun Yu (ShanghaiTech University); Siting Liu (ShanghaiTech University); Jie Han (University of Alberta)
731	SPARK: An Efficient Hybrid Acceleration Architecture with Run-Time Sparsity-Aware Scheduling for TinyML Learning	Currently most TinyML devices only focus on inference, as training requires much more hardware resources. In this paper, we introduce SPARK, an efficient hybrid acceleration architecture with run-time sparsity-aware scheduling for TinyML learning. Besides a stand-alone accelerator, an in-pipeline acceleration unit is integrated within the CPU pipeline to support simultaneous forward and backward propagation. To better utilize sparsity and improve hardware utilization, a sparsity-aware acceleration scheduler is implemented to schedule the workload between two acceleration units. A unified memory system is also constructed to support transposable data fetch, reducing memory access. We implement SPARK using TSMC 22nm technology and evaluate different TinyML tasks. Our work is the first architecture to utilize two acceleration units for on-device learning. Compared with the baseline accelerator, SPARK achieves 4.1x performance improvement in average with only 2.27% area overhead. SPARK also outperforms off-shelf edge devices in performance by 9.4x with 446.0x higher efficiency.	Mingxuan Li (Peking University); Qinzhe Zhi (Peking University); Yanchi Dong (Peking University); Le Ye (Peking University); Tianyu Jia (Peking University)
733	Knowing The Spec to Explore The Design via Transformed Bayesian Optimization	AI chip scales expediently in the large language models (LLMs) era. In contrast, the existing chip design space exploration methods, aimed at discovering optimal yet often infeasible or unproduceable Pareto-front designs, are hindered by neglect of design specifications. In this paper, we propose a novel Spec-driven transformed Bayesian optimization framework to find expected optimal RISC-V SoC architecture designs for LLM tasks. The highlights of our framework lie in a tailored transformed Gaussian process (GP) model prioritizing specified target metrics and a customized acquisition function (EHRM) in multi-objective optimization. Extensive experiments on large-scale RISC-V SoC architecture design explorations for LLMs, such as Transformer, BERT, and GPT-1, demonstrate that our method not only can effectively find the design according to QoR values from the spec, but also outperforms 34.59% in ADRS over previous state-of-the-art approach with 66.67% runtime.	Donger Luo (Shanghaitech University); QI SUN (Zhejiang University); Xinheng Li (Shanghai Tech); Chen BAI (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong); Hao Geng (ShanghaiTech University)
737	Mixed-Size 3D Analytical Placement with Heterogeneous Technology Nodes	This paper proposes a mixed-size 3D analytical placement framework for face-to-face stacked integrated circuits fabricated with heterogeneous technology nodes and connected by hybrid bonding technology. The proposed framework efficiently partitions a given netlist into two dies and optimizes the positions of each macro, standard cell, and hybrid bonding terminal (HBT). A multi-technology objective function and a multi-technology density penalty calculation process are adopted to handle the heterogeneous-technology-node constraints during mixed-size 3D global placement. Furthermore, a 3D objective function is used to refine the placement result during HBT-cell co-optimization. Our placer achieves the best results for all contest test cases compared with the participating teams at the 2023 CAD Contest at ICCAD on 3D Placement with Macros.	Yan-Jen Chen (National Taiwan University); Cheng-Hsiu Hsieh (National Taiwan University); Po-Han Su (National Taiwan University); Shao-Hsiang Chen (National Taiwan University); Yao-Wen Chang (National Taiwan University)

Submissio	Title	Abstract	Authors with Affiliations
742	Drift: Leveraging Distribution-based Dynamic Precision Quantization for Efficient Deep Neural Network Acceleration	Quantization is one of the most hardware-efficient ways to reduce inference costs for deep neural network (DNN) models. Nevertheless, with the continuous growth of DNN model size, existing static quantization methods fail to utilize the sparsity of models sufficiently. Motivated by the pervasive dynamism in data tensors across DNN models, we propose a dynamic precision quantization algorithm to further reduce computational costs. Furthermore, to address the shortcomings of existing precision-flexible accelerators, we design a novel accelerator, Drift, and achieve online scheduling to efficiently support dynamic precision execution. Evaluation results show that Drift achieves 2.85x speedup and 3.12x energy saving over existing precision-flexible accelerators.	Lian Liu (State Key Lab of Processors, Institute of Computing Technology, CAS; University of Chinese Academy of Sciences); Zhaohui Xu (School of Information Science and Technology, ShanghaiTech University); Yintao He (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences); Ying Wang (State Key Laboratory of Computer Architecture, 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); yinhe han (Institute of Computing Technology,Chinese Academy of Sciences)
752	Redistribution Layer Routing with Dynamic Via Insertion Under Irregular Via Structure	In modern advanced packaging, redistribution layers (RDLs) are often used for signal transmission among chips, and vias are used for communication among different layers. Most existing RDL routers perform via planning before routing. However, since vias can be placed at arbitrary locations under the irregular via structure, via planning limits the solution space and reduces layout flexibility. This paper proposes a new flow with a novel routing graph model for 90- and 135-degree routing, which allows dynamic via insertion during routing. The proposed algorithm enlarges the solution space by providing more choices during path-finding, achieving higher routing quality. The experimental results based on commonly used benchmark suites show that our router achieves over 10\% better wirelength with over 29X speedup over the state-of-the-art work and even achieves 0.4\% better wirelength with 55X speedup over the state-of-the-art any-angle router.	Je-Wei Chuang (National Taiwan University); Zong-Han Wu (National Taiwan University); Bo-Ying Huang (National Taiwan University); Yao-Wen Chang (National Taiwan University)
756	Efficient Bilevel Source Mask Optimization	Resolution Enhancement Techniques (RETs) are critical to meet the demands of advanced technology nodes. Among RETs, Source Mask Optimization (SMO) is pivotal, concurrently optimizing both the source and the mask to expand the process window. Traditional SMO methods, however, are limited by sequential and alternating optimizations, leading to extended runtimes without performance guarantees. This paper introduces a unified SMO framework utilizing the accelerated Abbe forward imaging to enhance precision and efficiency. Further, we propose the innovative \texttt{BiSMO} framework, which reformulates SMO through a bilevel optimization approach, and present three gradient-based methods to tackle the challenges of bilevel SMO. Our experimental results demonstrate that \texttt{BiSMO} achieves a remarkable 40\% reduction in error metrics and 8$\times$ increase in runtime efficiency, signifying a major leap forward in SMO.	Guojin Chen (The Chinese University of HongKong); Hongquan He (ShanghaiTech University); Peng Xu (The Chinese University of Hong Kong); Hao Geng (ShanghaiTech University); Bei Yu (The Chinese University of Hong Kong)
758	A Cache/Algorithm Co-design for Parallel Real-Time Systems with Data Dependency on Multi/Many-core System-on-Chips	Parallel real-time systems often rely on the shared cache for dependent data transmissions across cores. Conventional shared cache and their management techniques suffer from intensive contention and are markedly inflexible, leading to significant transmission latency of shared data. In this paper, we provide a Virtual Indexed Physical Tagged, Selectively-Inclusive Non-Exclusive L1.5 Cache, offering way-level control and versatile sharing capabilities. Focusing on a common-seen parallel task model, the Directed Acyclic Graph (DAG), we construct a novel scheduling method that exploits the L1.5 Cache to reduce data transmission latency, achieving improved timing performance. As a systematical solution, we build a real system, from the SoC and ISA to the drivers and the programming model. Experiments show that the proposed solution significantly improves the real-time performance of DAG tasks with negligible hardware overhead.	Zhe Jiang (South East University); Shuai Zhao (Sun Yat-sen University); Ran Wei (Lancaster University); Yiyang Gao (Sun Yat-sen University); Jing Li (New Jersey Institute of Technology)
767	FDCA: Fine-grained Digital-CIM based CNN Accelerator with Hybrid Quantization and Weight-Stationary Dataflow	Digital-Compute-In-Memory (DCIM) has demonstrated significant energy and area efficiency in convolutional neural network (CNN) accelerators, particularly for high precision applications. However, to mitigate parasitic effects on word and bit lines, most DCIMs employ fine-grained multiply-accumulate operations, which causes new challenges and opportunities but have not been widely explored. This paper proposes FDCA: a Fine-grained Digital-CIM based CNN Accelerator with hybrid quantization and weight-stationary dataflow, in which the key contributions are :1) a hybrid quantization approach for CNNs leveraging hessian trace and approximation is utilized. This method incorporates the ratio of computation time and storage time into quantization, achieving high efficiency while maintaining accuracy; 2) a Cartesian Genetic Programming based approximate shift and accumulate unit with error compensation is proposed, where an approximate adder tree is generated to compensate for errors introduced by DCIM; 3) a weight-stationary dataflow optimized for fine-grained DCIM is used to improve the utilization of CIM macro and eliminate dataflow stalls. The experimental results demonstrate that under 28-nm process, when running VGG16 and ResNet50 on CIFAR100, the proposed FDCA achieves 17.1TOPS/W and 18.79TOPS/W with accuracy loss by 0.71% and 0.98%, respectively. Compared to previous works, this work achieves 1.76× and 1.29× improvements in energy efficiency with less accuracy loss.	Bo Liu (Southeast University); Qingwen Wei (Southeast University); Yang Zhang (Southeast University); Xingyu Xu (Southeast University); Zihan Zou (Southeast University); Xinxiang Huang (Southeast University); Xin Si (Southeast University); Hao Cai (Southeast University)
768	AIG-CIM: A Scalable Chiplet Module with Tri-Gear Heterogeneous Compute-in-Memory for Diffusion Acceleration	The emergence of Diffusion models has gained significant attention in the field of Artificial Intelligence Generated Content. While Diffusion demonstrates impressive image generation capability, it faces hardware deployment challenges due to its unique model architecture and computation requirement. In this paper, we present a hardware accelerator design, i.e. AIG-CIM, which incorporates tri-gear heterogeneous digital compute-in-memory to address the flexible data reuse demands in Diffusion models. Our framework offers a collaborative design methodology for large generative models from the computational circuit-level to the multi-chip-module system-level. We implemented and evaluated the AIG-CIM accelerator using TSMC 22nm technology. For several Diffusion inferences, scalable AIG-CIM chiplets achieve 21.3× latency reduction, up to 231.2× throughput improvement and three orders of magnitude energy efficiency improvement compared to the NVIDIA RTX 3090 GPU.	Yiqi Jing (Peking University); Meng Wu (Peking University); Jiaqi Zhou (Peking University); Yiyang Sun (Peking University); Yufei Ma (Peking University); Ru Huang (Peking University); Le Ye (Peking University); Tianyu Jia (Peking University)
770	FastQuery: Communication-efficient Embedding Table Query for Private LLMs inference	With the fast evolution of large language models (LLMs), privacy concerns with user queries arise as they may contain sensitive information. Private inference based on homomorphic encryption (HE) has been proposed to protect user query privacy. However, private embedding table query has to be formulated as a HE-based matrix-vector multiplication problem and hence, suffers from enormous computation and communication overhead. We observe the overhead mainly comes from the neglect of 1) the one-hot nature of user queries and 2) the robustness of the embedding table to low-precision quantization noise. Hence, in this paper, we propose a private embedding table query optimization framework, dubbed FastQuery. FastQuery features a communication-aware embedding table quantization algorithm and a one-hot-aware dense packing algorithm to simultaneously reduce both the computation and communication costs. Compared to prior-art HE-based frameworks, e.g., CrypTFlow2, Iron, Cheetah, and CHAM, FastQuery achieves 2.7 ~ 4.5× computation and 75.1 ~ 84.4× communication reduction on both LLAMA-7B and LLAMA-30B.	Chenqi Lin (Peking University); Tianshi Xu (Peking University); Zebin Yang (Peking University); Runsheng Wang (Peking University); Ru Huang (Peking University); Meng Li (Institute for Artificial Intelligence and School of Integrated Circuits, Peking University)
774	PIVOT- Input-aware Path Selection for Energy-efficient ViT Inference	The sophisticated self-attention-based spatial correlation entails a high inference delay cost in vision transformers. To this end, we propose PIVOT-a hardware-algorithm co-optimization framework for input-difficulty-aware attention skipping for attention bottleneck optimization. The attention-skipping configurations are obtained via an iterative hardware-in-the loop co-search method. On the ZCU102 MPSoC FPGA, PIVOT achieves 2.7×(1.73×) lower EDP at 0.2%(0.4%) accuracy reduction compared to standard LVViT-S (DeiT-S) ViTs. Unlike prior works that require nuanced hardware support, PIVOT is compatible with traditional GPU and CPU platforms- 1.8× higher throughput at 0.4-1.3% higher accuracy compared to prior works.	Abhishek Moitra (Yale University); Abhiroop Bhattacharjee (Yale University); Priyadarshini Panda (Yale University)

Submissio	Title	Abstract	Authors with Affiliations
781	EMOGen: Enhancing Mask Optimization via Pattern Generation	Layout pattern generation via deep generative models is a promising methodology for building practical large-scale pattern libraries. However, although improving optical proximity correction (OPC) is a major target of existing pattern generation methods, they are not explicitly trained for OPC and integrated into OPC methods. In this paper, we propose EMOGen to enable the co-evolution of layout pattern generation and learning-based OPC methods. With the novel co-evolution methodology, we achieve up to 39% enhancement in OPC and 34% improvement in pattern legalization.	Su Zheng (The Chinese University of Hong Kong); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou)); Bei Yu (The Chinese University of Hong Kong); Martin Wong (The Chinese University of Hong Kong)
785	Enhancing 3-D Random Walk Capacitance Solver with Analytic Surface Green's Functions of Transition Cubes	The complicated dielectric profile under advanced process technologies challenges the accuracy of floating random walk (FRW) based capacitance extraction, as the latter pre-computes the surface Green's functions for a finite set of multi-dielectric transition cubes and makes approximations of transition cubes during the FRW process. In this work, we derive analytic surface Green's functions for transition cubes with arbitrary stratified dielectrics and propose a fast algorithm named AGF to compute them. A capacitance solver named FRW-AGF is then proposed to incorporate AGF into the FRW process to accurately model realistic transition cubes. Experimental results show that the proposed AGF is over 100x, faster than the state-of-the-art, and FRW-AGF largely improves the accuracy of RWCap4 [3, 17] (making all errors to golden values below 5%) without degrading computational speed and parallel scalability.	Jiechen Huang (Tsinghua University); Wenjian Yu (Tsinghua University)
794	How to Steal CPU Idle Time When Synchronous I/O Mode Becomes Promising	The advent of ultra-low-latency storage devices has narrowed the performance gap between storage and CPU in computing platforms, facilitating synchronous I/O adoption. Yet, this approach introduces substantial busy waiting time and underutilizes computing units. To address this, we propose a light-weighted Idle-Time-Stealing (ITS) design. This involves a self-improving thread conducting pre-fetching for high-priority processes during synchronous I/O, and an I/O-waiting process continuing subsequent instruction executions when justifiable. Another thread, the self-sacrificing thread, proactively switches low-priority process I/O requests from synchronous to asynchronous mode, prioritizing high-priority executions. Experimental results demonstrate the effectiveness of our ITS design in reducing CPU idle time.	Chun-Feng Wu (National Yang Ming Chiao Tung University); Yuan-Hao Chang (Academia Sinica); Ming-Chang Yang (The Chinese University of Hong Kong); Tei-Wei Kuo (National Taiwan University)
799	A High-Throughput Private Inference Engine Based on 3D Stacked Memory	Fully Homomorphic Encryption (FHE) enables unlimited computation depth, allowing for privacy-enhanced neural network inference tasks directly on the ciphertext. However, existing FHE architectures suffer from the memory access bottleneck due to the significant data consumption. This work proposes a High-throughput FHE engine for private inference (PI) based on 3D stacked memory (H3). H3 adopts software-hardware co-design that dynamically adjusts the polynomial decomposition during the PI process to minimize the computation and storage overhead at a fine granularity. With 3D hybrid bonding, H3 integrates a logic die with a multi-layer embedded DRAM, routing data efficiently to the processing unit array through an efficient broadcast mechanism. H3 consumes 192mm^2 of the area when implemented using a 28nm logic process. H3 achieves a throughput of 1.36 million LeNet-5 or 920 ResNet-20 PI per minute, surpassing existing 7nm accelerators by 52%. This demonstrates that 3D memory is a promising technology to promote the performance of FHE.	Zhaohui Chen (Alibaba Group); LING LIANG (Peking University); Qi Liu (Alibaba Group); Zhirui Li (Alibaba Group); Fahong Zhang (Alibaba Groups); Yanheng Lu (Alibaba Groups); Zhen Gu (Alibaba Group)
803	Synthesis of Resource-Efficient Superconducting Circuits with Clock-Free Alternating Logic	Gate-level clocking, typical in traditional approaches to Single Flux Quantum (SFQ) technology, makes the effective synthesis of superconducting circuits a significant engineering hurdle. This paper addresses this challenge by employing the recently introduced xSFQ logic family. xSFQ leverages dual-rail alternating encoding to eliminate the clock dependency from the superconducting gate semantics. This obviates the need for ad hoc modifications to existing synthesis tools and avoids unnecessary circuit resource overheads, marking a significant advancement in superconducting circuit design automation. Our implementation results demonstrate an average reduction of over 80% in the Josephson junction count for circuits from the ISCAS85, EPFL, and ISCAS89 benchmark suites.	Jennifer Volk (University of California, Santa Barbara; University of Michigan); Panagiotis Papanikolaou (University of Michigan; University of Patras); Georgios Zervakis (University of Patras); Georgios Tzimpragos (University of Michigan)
812	CAMO: Correlation-Aware Mask Optimization with Modulated Reinforcement Learning	Optical proximity correction (OPC) is a vital step to ensure printability in modern VLSI manufacturing. Various OPC approaches have been proposed, which are typically data-driven and hardly involve particular considerations of the OPC problem, leading to potential performance bottlenecks. In this paper, we propose CAMO, a reinforcement learning-based OPC system that integrates important principles of the OPC problem. CAMO explicitly involves the spatial correlation among the neighboring segments and an OPC-inspired modulation for movement action selection. Experiments are conducted on via patterns and metal patterns. The results demonstrate that CAMO outperforms state-of-the-art OPC engines from both academia and industry.	Xiaoxiao Liang (The Hong Kong University of Science and Technology (Guangzhou)); Haoyu Yang (NVIDIA Corp.); Kang Liu (Huazhong University of Science and Technology); Bei Yu (The Chinese University of Hong Kong); Yuzhe Ma (The Hong Kong University of Science and Technology (Guangzhou))
833	NSPG: Natural language Processing-based Security Property Generator for Hardware Security Assurance	The efficiency of validating complex System-on-Chips (SoCs) is contingent on the quality of the security properties provided. Generating security properties with traditional approaches often requires expert intervention and is limited to a few IPs, thereby resulting in a time-consuming and non-robust process. To address this issue, we, for the first time, propose a novel and automated Natural Language Processing (NLP)-based Security Property Generator (NSPG). Specifically, our approach utilizes hardware documentation in order to propose the first hardware security-specific language model, HS-BERT, for extracting security properties dedicated to hardware design. It is capable of phasing a significant amount of hardware specification, and the generated security properties can be easily converted into hardware assertions, thereby reducing the manual effort required for hardware verification. NSPG is trained using sentences from several SoC documentation and achieves up to 88% accuracy for property classification, outperforming ChatGPT. When assessed on five untrained OpenTitan hardware IP documents, NSPG aided in identifying eight security vulnerabilities in the buggy OpenTitan SoC presented in Hack@DAC 2022.	Xingyu Meng (University of Texas at Dallas); Amisha Srivastava (University of Texas at Dallas); Ayush Arunachalam (University of Texas at Dallas); Haoyu Yang (NVIDIA Corp.); Avik Ray (Amazon); Pedro Henrique Silva (Technology Innovation Institute); Rafail Psiakis (Technology Innovation Institute); Yiorgos Makris (The University of Texas at Dallas); Kanad Basu (University of Texas at Dallas)
834	ALVEARE: a Domain-Specific Framework for Regular Expressions	Regular Expression (RE) matching enables the identification of patterns in datastream of heterogeneous fields ranging from proteomics to computer security. These scenarios require massive data analysis that, combined with the high data dependency of the REs, leads to long computational times and high energy consumption. Currently, RE engines rely on either (1) flexibility in run-time RE changes and broad operators support impairing performance or (2) fixed high-performing accelerators implementing few simple RE operators. To overcome these limitations, we propose ALVEARE: a hardware-software approach combining a Domain-Specific Language (DSL) with an embedded Domain-Specific Architecture. We exploit REs as a DSL by translating them into flexible executables through our RISC-based Instruction Set Architecture that expresses from simple to advanced primitives. Then, we design a speculation-based microarchitecture to execute real benchmarks efficiently. ALVEARE provides RE-domain flexibility and broad operators' support and achieves up to 34x speedup and 57x energy efficiency improvements against the state-of-the-art RE2 and Bluefield DPU 2 with its RE accelerator.	Filippo Carloni (Politecnico di Milano); Davide Conficconi (Politecnico di Milano); Marco D. Santambrogio (Politecnico di Milano)

Submissio	Title	Abstract	Authors with Affiliations
839	Addition is Most You Need: Efficient Floating-Point SRAM Compute-in-Memory by Harnessing Mantissa Addition	The compute-in-memory (CIM) paradigm holds great promise to efficiently accelerate machine learning workloads. Among memory devices, static random-access memory (SRAM) stands out as a practical choice due to its exceptional reliability in the digital domain and balanced performance. Recently, there has been a growing interest in accelerating floating-point (FP) deep neural networks (DNNs) with SRAM CIM due to their critical importance in DNN training and high-accurate inference. This paper proposes an efficient SRAM CIM macro for FP DNNs. To achieve the design, we identify a lightweight approach that decomposes conventional FP mantissa multiplication into two parts: mantissa sub-addition (sub-ADD) and mantissa sub-multiplication (sub-MUL). Our study shows that while mantissa sub-MUL is compute-intensive, it only contributes to the minority of FP products, whereas mantissa sub-ADD, although compute-light, accounts for the majority of FP products. Recognizing "Addition is Most You Need", we develop a hybrid-domain SRAM CIM macro to accurately handle mantissa sub-ADD in the digital domain while improving the energy efficiency of mantissa sub-MUL using analog computing. Experiments with the MLPerf benchmark demonstrate its remarkable improvement in energy efficiency by 8.7×~ 9.3× (7.3×~8.2×) in inference (training) compared to a fully digital FP baseline without any accuracy loss, showcasing its great potential for FP DNN acceleration.	Weidong Cao (The George Washington University); Jian Gao (Northeastern University); Xin Xin (University of Central Florida); Xuan Zhang (Northeastern University)
849	Massively Parallel AIG Resubstitution	Resubstitution is a flexible algorithmic framework for circuit restructuring that has been incorporated into many high-effort logic optimization flows. It is thus important to speed up resubstitution in order to obtain high-quality realizations of large-scale designs. This paper proposes a massively parallel AIG resubstitution algorithm targeting GPUs, with effective approaches to addressing cyclic dependencies and restructuring conflicts. Compared with ABC and mockturtle, our algorithm achieves 41.9x and 50.3x acceleration on average without quality degradation. When combining our resubstitution with other GPU algorithms, a GPU-based resyn2rs sequence obtains 46.4x speedup over ABC with 0.8% and 5.8% smaller area and delay respectively.	Yang Sun (The Chinese University of Hong Kong); Tianji Liu (The Chinese University of Hong Kong); Martin Wong (The Chinese University of Hong Kong); Evangeline Young (The Chinese University of Hong Kong)
852	An NTT/INTT Accelerator with Ultra-High Throughput and Area Efficiency for FHE	As a core arithmetic operation and security guarantee of Fully Homomorphic Encryption (FHE), Number Theoretic Transform (NTT) of a large degree is the primary source of computational and time overhead. In this paper, we propose a scalable and conflict-free memory mapping algorithm that breaks the memory bound and releases a large amount of on-chip resources. A flexible and no-stall hardware/software pipeline architecture is designed to boost the throughput of NTT/INTT of $N=2^{16}$ to over 48,543 operations per second with area efficiency, which 4× and 10× speed up the FPGA-based (HPCA'23) and GPU-based (HPCA'23) schemes.	Zhaojun Lu (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); Peng Xu (School of Cyber Science and Engineering, Huazhong University of Science and Technology); Wei Wang (School of Cyber Science and Engineering, Huazhong University of Science and Technology); Jiliang Zhang (College of Integrated Circuits, Hunan University); Dengguo Feng (State Key Laboratory of Cryptology, P.O.Box 5159)
856	Binding Multi-bit Flip-flop Cells through Design and Technology Co-optimization	Though using multi-bit flip-flop (MBFF) cells provide the benefit of saving dynamic power, its big cell size with many D/Q-pins inherently entails two critical limitations, which are (1) the loss of full flexibility in optimizing the wires connecting to the D/Q-pins in MBFFs and (2) the loss of selectively resizing i.e., controlling output driving strength of internal flip flops in MBFFs to optimize timing. In this work, we propose a comprehensive solution to resolving the limitations through design and technology co-optimization (DTCO) in physical design flow. Specifically, to address limitation 1, given an input circuit with MBFF allocation and binding solution, at the post-placement stage we explore diverse layouts of MBFF cells with various D/Q-pin locations and rebind every MBFF instance in the circuit to the MBFF cell layout that is the most suitable for minimizing the wirelength connecting D/Q-pins. Meanwhile, to address limitation 2, at the post-route stage we explore MBFF cell layouts of non-rectangle (precisely, L- or T-shape) to control the driving strength of internal flip-flops selectively, by which we rebind MBFF instances with negative slack to the area-minimal MBFF cells to optimize timing while increasing the power overhead minimally.	Jooyeon Jeong (Seoul National University); Taewhan Kim (Seoul National University)
865	LEAF: An Adaptation Framework against Noisy Data on Edge through Ultra Low-Cost Training	In real-world neural network deployments, incoming data often contains noise and imperfections. Retraining on resource-constrained edge devices becomes essential to maintain performance. To tackle this challenge, we introduce LEAF, a hardware-efficient framework designed for adapting to degraded images. By analyzing neural network behavior on degraded images, we propose two techniques: 1) Selective Experience Replay for skipping unimportant images, reducing computation, and 2) Pseudo Noise Dithering for extremely low precision (3 or 4-bit) gradient quantization, enabling nearly full-integer training. Extensive experiments on CIFAR10 and Tiny ImageNet datasets, with various image degradations, demonstrate LEAF's ultra-low cost with minimal accuracy loss.	Zihan Xia (University of California San Diego); Jinwook Kim (Sk Hynix inc.); Mingu Kang (University of California San Diego)
869	KATO: Knowledge Alignment And Transfer for Transistor Sizing Of Different Design and Technology	Automatic transistor sizing in circuit design continues to be a formidable challenge. Despite that Bayesian optimization (BO) has achieved significant success, it is circuit-specific, limiting the accumulation and transfer of design knowledge for broader applications. This paper proposes (1) efficient automatic kernel construction, (2) the first transfer learning across different circuits and technology nodes for BO, and (3) a selective transfer learning scheme to ensure only useful knowledge is utilized. These three novel components are integrated into BO with Multi-objective Acquisition Ensemble (MACE) to form Knowledge Alignment and Transfer Optimization (KATO) to deliver state-of-the-art performance: up to 2x simulation reduction and 1.2x design improvement over the baselines.	Wei W. Xing (The University of Sheffield); Weijian Fan (Shenzhen University); Zhuohua Liu (Shenzhen University); Yuan Yao (Beihang University); Yuanqi Hu (Beihang University)
872	SPECRUN: The Danger of Speculative Runahead Execution in Processors	Runahead execution is a continuously evolving microarchitectural technique for processor performance. This paper introduces the first transient execution attack on the runahead execution, called SPECRUN, which exploits the unresolved branch prediction during runahead execution. We show that SPECRUN eliminates the limitation on the number of transient instructions posed by the reorder buffer size, enhancing the exploitability and harmfulness of the attack. We concretely demonstrate a proof-of-concept attack that causes leaking secrets from a victim process, validate the merit of SPECRUN, and design a secure runahead execution scheme. This paper highlights the need to consider the security of potential optimization techniques before implementing them in a processor.	Chaoqun Shen (Hunan University); Gang Qu (Univ. of Maryland, College Park); Jiliang Zhang (College of Integrated Circuits, Hunan University)
883	RT-MDM: Real-Time Scheduling Framework for Multi-DNN on MCU Using External Memory	As the application scope of DNNs executed on microcontroller units (MCUs) extends to time-critical systems, it becomes important to ensure timing guarantees for increasing demand of DNN inferences. To this end, this paper proposes RT-MDM, the first real-time scheduling framework for multiple DNN tasks executed on an MCU using external memory. Identifying execution-order dependencies among segmented DNN models and memory requirements for parallel execution subject to the dependencies, we propose (i) a segment-group-based memory management policy that achieves isolated memory usage within a segment group and sharded memory usage across different segment groups, and (ii) an intra-task scheduler specialized for the proposed policy. Implementing RT-MDM on an actual system and optimizing its parameters for DNN segmentation and segment-group mapping, we demonstrate the effectiveness of RT-MDM in accommodating more DNN tasks while providing their timing guarantees.	Sukmin Kang (Sungkyunkwan University (SKKU)); Seongtae Lee (Sungkyunkwan University); Hyunwoo Koo (Sungkyunkwan University); Hoon Sung Chwa (DGIST); Jinkyu Lee (Sungkyunkwan University (SKKU))

Submissio	Title	Abstract	Authors with Affiliations
891	HAIL-DIMM: Host Access Interleaved with Near-Data Processing on DIMM-based Memory System	Near-data processing (NDP), a solution to reduce data movement overhead between host and memory, should not interfere with host access to ensure system fairness. We propose a cost-effective and energy-efficient LRDIMM-based NDP architecture (HAIL-DIMM) that can seamlessly interleave NDP and regular memory access and is a drop-in replacement for existing main memory modules. The proposed NDP exploits the interleaving capability of the memory controller to interleave NDP and host access naturally. To take advantage of bank interleaving, an atomic operation of the proposed NDP, which consists of data movement and computation, is recognized by the memory controller as a DDR READ/WRITE but by the HAIL-DIMM as NDP based on the request's address. We implement a prototype of the proposed NDP architecture on an FPGA platform as proof of concept. The evaluation results show that the NDP system achieves up to 2.19x speedup in latency and up to 45.4 % energy saving for data movement over the baseline system in memory-bound workloads.	Minkyu Lee (Korea Electronics Technology Institute); Sang-Seol Lee (Korea Electronics Technology Institute); Kyungho Kim (Korea Electronics Technology Institute); Eunchong Lee (Korea Electronics Technology Institute); Sung-Joon Jang (Korea Electronics Technology Institute)
895	Fake Node-Based Perception Poisoning Attacks against Federated Object Detection Learning in Mobile Computing Networks	Federated learning (FL) supports massive edge devices to collaboratively train object detection models in mobile computing scenarios. However, the distributed nature of FL exposes significant security vulnerabilities. Existing attack methods either require considerable costs to compromise the majority of participants, or suffer from poor attack success rates. Inspired by this, we devise an efficient fake node-based perception poisoning attacks strategy (FNPPA) to target such weaknesses. In particular, FNPPA poisons local data and injects multiple fake nodes to participate in aggregation, aiming to make the local poisoning model more likely to overwrite clean updates. Moreover, it can achieve greater malicious influence on target objects at a lower cost without affecting the normal detection of other objects. We demonstrate through exhaustive experiments that FNPPA exhibits superior attack impact than the state-of-the-art in terms of average precision and aggregation effect.	Xiong Xiao (Hunan University); Mingxing Duan (Hunan University); Yingjie Song (Hunan University); Zhuo Tang (Hunan University); Wenjing Yang (National University of Defense Technology)
903	Every Failure Is A Lesson: Utilizing All Failure Samples To Deliver Tuning-Free Efficient Yield Evaluation	Yield estimation and optimization have become increasingly important for circuit design as technology nodes scale down. Simple yet well-established minimal norm importance sampling (MNIS) still serves as an industrial standard due to its robustness and reliability. In this study, we generalize the classic MNIS and propose Every Failure Is A Lesson (EFIAL) to utilize every failure sample (instead of one in MNIS) to construct the proposal distribution. EFIAL is completely tuning-free and the update computation complexity is only $\Ocal(M)$ (M is the number of failure samples) by utilizing the blessing of dimensionality. The idea of EFIAL is then extended to the state-of-the-art (SOTA) pre-sampling method, onion sampling, to significantly boost efficiency, by up to 9.08x (4.68x on average). Extensive evaluations against SOTA yield estimation methods reveal that EFIAL achieves a speedup of up to 13.54x (5.16x on average) and an accuracy improvement of up to 24.91\%.	Wei W. Xing (The University of Sheffield); Yanfang Liu (Beihang University); Weijian Fan (Shenzhen University); Lei He (Eastern institute of technology); Ting-Jung Lin (Ningbo Institute of Digital Twin)
909	Cache-aware Task Decomposition for Efficient Intermittent Computing Systems	Energy harvesting offers a scalable and cost-effective power solution for IoT devices, but it introduces the challenge of frequent and unpredictable power failures due to the unstable environment. To address this, intermittent computing has been proposed, which periodically backs up the system state to non-volatile memory (NVM), enabling robust and sustainable computing even in the face of unreliable power supplies. In modern processors, write back cache is extensively utilized to enhance system performance. However, it poses a challenge during backup operations as it buffers updates to memory, potentially leading to inconsistent system states. One solution is to adopt a write-through cache, which avoids the inconsistency issue but incurs increased memory access latency for each write reference. Some existing work enforces a cache flushing before backups to maintain a consistent system state, resulting in significant backup overhead. In this paper, we point out that although cache delays updates to the main memory, it may preserve a recoverable system state in the main memory. Leveraging this characteristic, we propose a cache-aware task decomposition method that divides an application into multiple tasks, ensuring that no dirty cache lines are evicted during their execution. Furthermore, the cache-aware task decomposition maintains a unchanged memory state during the execution of each task, enabling us to parallelize the backup process with task execution and effectively hide the backup latency. Experimental results with different power traces demonstrate the effectiveness of the proposed system.	Shuo Xu (School of Cyber Science and Technology, Shandong University); Wei Zhang (School of Cyber Science and Technology, Shandong University); Mengying Zhao (Shandong University); Zimeng Zhou (School of Cyber Science and Technology, Shandong University); Lei Ju (School of Cyber Science and Technology, Shandong University)
911	SC-GNN: A Communication-Efficient Semantic Compression for Distributed Training of GNNs	Training big graph neural networks (GNNs) in distributed systems is quite time-consuming mainly because of the ubiquitous aggregate operations that involve a large amount of cross-partition communication for collecting embeddings/gradients during the forward and backward propagations. To reduce the volume of the communication, some recent approaches focused on decaying each of connections via sampling, quantifying, or delaying until satisfactory trade-off are obtained between volume and accuracy. However, when applied to popular GNNs, those approaches are found to be bounded by a common volume/accuracy Pareto frontier which shows that the decaying for individual connection cannot further accelerate the aggregate of training. In this work, SC-GNN, a semantic compression of the cross-partition communication, is proposed to concentrate a group of connections as a high-level semantics and transmit to a target partition. Since carrying the overall intent of a group, the semantics can keep transferring the interactions, i.e., embeddings/gradients, between a pair of remote partitions until GNN models converge. In addition, a connection-pattern based differential optimization is proposed to further prune those weak connections, while guaranteeing the training accuracy. The results show that, for multi-field datasets, the compression rate of SC-GNN is 40.8 times higher than SOTA methods and the epoch time is reduced to 31.77% on average.	Jihe Wang (Computer Science Department, Northwestern Polytechnical University); Ying Wu (Computer Science Department, Northwestern Polytechnical University); Danghui Wang (Computer Science Department, Northwestern Polytechnical University)
924	Conclave - Secure and Robust Cooperative Perception for Connected Autonomous Vehicle Using Authenticated Consensus and Trust Scoring	Connected Autonomous Vehicles have great potential to improve automobile safety and traffic flow, especially in cooperative applications where perception data is shared between vehicles. However, this cooperation must be secured from malicious intent and unintentional errors that could cause accidents. In this paper, we propose Conclave -- a tightly coupled authentication, consensus, and trust scoring mechanism that provides comprehensive security and reliability for cooperative perception. Overall, Conclave shows huge promise in preventing security flaws, detecting even relatively minor sensing faults, and increasing the robustness and accuracy of cooperative perception in CAVs while adding minimal overhead.	Edward Andert (Arizona State University); Francis Mendoza (Arizona State University); Hans Walter Behrens (Arizona State University); Aviral Shrivastava (Arizona State University)

Submissio	Title	Abstract	Authors with Affiliations
927	PipeSSD: A Lock-free Pipelined SSD Firmware Design for Multi-core Architecture	Modern SSD firmware is continuously optimized for higher parallelism to match the growing frontend PCIe bandwidth with more backend flash channels. Although a multi-core microprocessor is typically adopted to concurrently process independent NVMe requests from multiple NVMe queues, the existing one-to-many thread-request mapping model with each thread serving one or more incoming I/O requests has poor scalability due to severe lock contention problem, especially in cache management. In this paper, we first conduct preliminary experiments on an open-channel NVMe SSD to exhibit the lock contention problem in the one-to-many thread-request mapping model. When a thread locks a cache line and is waiting for a long-latency flash read to update this cache line, subsequent tasks on other threads that require the same cache line are all blocked to guarantee correctness. To mitigate this, we propose PipeSSD, a lock-free pipeline-based SSD firmware design with a many-to-one thread-request mapping model that assigns multiple threads to serve different stages of each I/O request in a pipelined way. It is worth noting that PipeSSD only performs cache updates in the last pipeline stage to eliminate dependency loops in the pipeline while maintaining a pilot for each cache line in the beginning pipeline stage to indicate the cache line status. With a multi-core architecture, different pipeline stages are processed on different cores communicated via several FIFO queues, which can ensure the processing sequence and data consistency without any cache line locks. We implement PipeSSD on real hardware and evaluate its performance on a multi-core NVMe SSD prototype. The evaluation results show that on an 8-core system, PipeSSD has a significant throughput improvement compared to the state-of-the-art multi-core SSD firmware.	Zelin Du (The Chinese University of Hong Kong); Shaoqi Li (Shenzhen University); Zixuan Huang (The Chinese University of Hong Kong); Jin Xue (The Chinese University of Hong Kong); Kecheng HUANG (The Chinese University of Hong Kong); Tianyu Wang (Shenzhen University); Zili Shao (The Chinese University of Hong Kong)
928	Sting: Near-storage accelerator framework for scalable triangle counting and beyond	One of the most critical limitations to scalable graph mining is memory capacity, as graphs of interest continue to grow while the rate of DRAM scaling diminishes. While high-performance NVMe storage is cheap and dense enough to better support larger graphs, the relative performance limitations of secondary storage force a cost-performance trade-off. We present STING, which uses an asynchronous callback function to provide a general interface to in-storage graphs while allowing transparent near-storage acceleration. Using triangle counting, we show with transparent filtering and sorting acceleration, STING can achieve improve state-of-the-art by 3x for cost and power efficiency.	Seongyoung Kang (University of California, Irvine); Sang-Woo Jun (University of California, Irvine)
933	Graph-Transformer-based Surrogate Model for Accelerated Converter Circuit Topology Design	Unlike circuit parameter and sizing optimizations, the automated design of analog circuit topologies poses significant challenges for learning-based approaches. One challenge arises from the combinatorial growth of the topology space with circuit size, which limits the topology optimization efficiency. Moreover, traditional circuit evaluation methods are time-consuming, while the presence of data discontinuity in the topology space makes the accurate prediction of circuit performance exceptionally difficult for unseen topologies. To tackle these challenges, we design a novel Graph-Transformer-based Network (GTN) as the surrogate model for circuit evaluation, offering a substantial acceleration in the speed of circuit topology optimization without sacrificing performance. Our GTN model architecture is designed to embed voltage changes in circuit loops and current flows in connected devices, enabling accurate performance predictions for circuits with unseen topologies. Taking the power converter circuit design as an experimental task, our GTN model significantly outperforms an analytical approach and baseline methods directly utilizing graph neural networks. Furthermore, GTN achieves less than 5% relative error and 196× speed-up compared with high-fidelity simulation. Notably, our GTN surrogate model empowers an automatic circuit design framework to discover circuits of comparable quality to those identified through high-fidelity simulation while reducing the time required by up to 97.2%.	Shaoze Fan (New Jersey Institute of Technology); Haoshu Lu (New Jersey Institute of Technology); Shun Zhang (IBM); Ningyuan Cao (University of Notre Dame); Xin Zhang (IBM T. J. Watson Research Center); Jing Li (New Jersey Institute of Technology)
941	Order-Preserving Cryptography for the Confidential Inference in Random Forests: FPGA Design and Implementation	Prior work has addressed the problem of confidential inference in decision trees. Both traditional order-preserving cryptography and order-preserving NTRU cryptography have been used to ensure data and model privacy in decision trees. Furthermore, FPGA architectures and implementations have been proposed for implementing such confidential inference algorithms on limited resource, edge platforms such as low-cost FPGA boards. In this paper, we address the challenging problem of scalability of order-preserving confidential inference to random forests, which are ensembles of decision trees that are meant to improve their classification accuracy and reduce their overfitting. The paper develops a methodology and an FPGA implementation strategy for scaling up order-preserving cryptography to random forests. In particular, a framework is used to study the multifaceted tradeoffs that exist between the number of trees in the random forest, the strength of the encryption, the accuracy of the inferences, and the resources of the edge platform. Extensive experiments are conducted using the MNIST dataset and the Intel DE10 Standard FPGA board.	Rupesh Raj Karn (New York University); Kashif Nawaz (Technology Innovation Institute); Ibrahim (Abe) M. Elfadel (Khalifa University)
954	LIVAK: A High-Performance In-Memory Learned Index for Variable-Length Keys	In-memory learned index has been an efficient approach supporting in-memory fast data access. However, existing learned indexes are inefficient in supporting variable-length keys. To address this issue, we propose a new in-memory learned index called LIVAK that adopts a hybrid structure involving trie, learned index, and B+-tree. Each node indexes an 8-byte slice of keys, and we use learned indexes for large nodes but B+-trees for small nodes. Also, LIVAK presents a character re-encoding mechanism to avoid performance degradation. We compare LIVAK with B+-tree, Masstree, and SIndex on various datasets and workloads, and the results suggest the efficiency of LIVAK.	Zhaole Chu (ustc); Zhou Zhang (University of Science and Technology of China); Peiquan Jin (University of Science and Technology of China); Xiaoliang Wang (University of Science and Technology of China); Yongping Luo (University of Science and Technology of China); Xujian Zhao (Southwest University of Science and Technology of China)
965	Less is More: Hop-Wise Graph Attention for Scalable and Generalizable Learning on Circuits	While graph neural networks (GNNs) have gained popularity for learning circuit representations in various electronic design automation (EDA) tasks, they face challenges in scalability when applied to large graphs and exhibit limited generalizability to new designs. These limitations make them less practical for addressing large-scale, complex circuit problems. In this work we propose HOGA, a novel attention-based model for learning circuit representations in a scalable and generalizable manner. HOGA first computes hop-wise features per node prior to model training. Subsequently, the hop-wise features are solely used to produce node representations through a gated self-attention module, which adaptively learns important features among different hops without involving the graph topology. As a result, HOGA is adaptive to various structures across different circuits and can be efficiently trained in a distributed manner. To demonstrate the efficacy of HOGA, we consider two representative EDA tasks: quality of results (QoR) prediction and functional reasoning. Our experimental results indicate that (1) HOGA reduces estimation error over conventional GNNs by 46.76% for predicting QoR after logic synthesis; (2) HOGA improves 10.0% reasoning accuracy over GNNs for identifying functional blocks on unseen gate-level netlists after complex technology mapping; (3) The training time for HOGA almost linearly decreases with an increase in computing resources.	Chenhui Deng (Cornell University); Zichao Yue (Cornell University); Cunxi Yu (University of Maryland, College Park); Gokce Sarar (Qualcomm Inc.); Ryan M. Carey (Qualcomm Technologies, Inc.); Rajeev Jain (Qualcomm); Zhiru Zhang (Cornell University)

Submissio	Title	Abstract	Authors with Affiliations
966	Size-Optimized Depth-Constrained Large Parallel Prefix Circuits	Binary adders are a critical building block in integrated circuit (IC) design. In addition to the widely used 32/64/128-bit adders, large (1024/2048 bits) adders are important in applications such as cryptography. However, most current adder design methods target regular bitwidths, and cannot efficiently generate large adders with good performance. In practice, adders are often integrated into circuits such as a multiplier-accumulator (MAC), resulting in complex non-uniform input arrival times. To address these challenges, we propose a new algorithm for efficiently generating high-quality adders for non-uniform input arrival times. It is based on a novel divide-and-conquer-friendly problem formulation, and can effectively generate and maintain the most useful adder structures through dynamic programming. Experimental results show that it outperforms the current state-of-the-art methods in both quality and runtime. The adders generated by our algorithm have 2.8%, 8.3%, and 10.3% reductions in delay, area, and power, respectively, compared to those generated by a commercial synthesis tool.	Shiju Lin (The Chinese University of Hong Kong); Bentian Jiang (Huawei Hong Kong Research Center); Weihua Sheng (Huawei Hong Kong Research Center); Evangeline Young (The Chinese University of Hong Kong)
970	Efficient Approximate Decomposition Solver using Ising Model	Computing with memory is an energy-efficient computing approach. It pre-computes a function and store its values in a lookup table (LUT), which can be retrieved at runtime. Approximate Boolean decomposition has been recently proposed to reduce the LUT size for implementing complex functions, but it takes a long time to find a decomposition with a minimized error. As a parallel algorithm developed based on the Ising model, simulated bifurcation (SB) is promised to be a high-performance approach for combinatorial optimization. In this paper, we propose an efficient SB-based approximate function decomposition approach. Specifically, a new approximate disjoint decomposition method, called column-based approximate disjoint decomposition, is first proposed to fit the Ising model. Then, it is adapted to the Ising model-based optimization solver. Moreover, two improvement techniques are developed for an efficient search of the approximate disjoint decomposition when using SB. The experiment results shows that compared to the state-of-the-art work, our approach achieves a 11% smaller mean error distance with an average 1.16× speedup when approximately decomposing 16-input 16-output Boolean functions.	Weihua Xiao (Shanghai Jiao Tong University); Tingting Zhang (University of Alberta); Xingyue Qian (Shanghai Jiao Tong University); Jie Han (University of Alberta); Weikang Qian (Shanghai Jiao Tong University)
979	3D-Carbon: An Analytical Carbon Modeling Tool for 3D and 2.5D Integrated Circuits	Environmental sustainability is a critical concern for Integrated Circuits (ICs) throughout their entire life cycle, particularly in manufacturing and use. Meanwhile, ICs using 3D/2.5D integration technologies have emerged as promising solutions to meet the growing demands for computational power. However, there is a distinct lack of carbon modeling tools for 3D/2.5D ICs. Addressing this, we propose 3D-Carbon, an analytical carbon modeling tool designed to quantify the carbon emissions of 3D/2.5D ICs throughout their life cycle. 3D-Carbon factors in both potential savings and overheads from advanced integration technologies, considering practical deployment constraints like bandwidth. We validate 3D-Carbon's accuracy against established baselines and illustrate its utility through case studies in autonomous vehicles. We believe that 3D-Carbon lays the initial foundation for future innovations in developing environmentally sustainable 3D/2.5D ICs.	Yujie Zhao (Georgia Institute of Technology); Yang (Katie) Zhao (University of Minnesota, Twin Cities); Cheng Wan (Georgia Tech); Yingyan (Celine) Lin (Georgia Institute of Technology)
980	Architectural Whispers: Robust Machine Learning Models Fingerprinting via Frequency Throttling Side-Channels	Security practices in the field of Machine learning (ML) encompass a range of measures, with one notable strategy that involves concealing the architecture of ML models from users, thereby adding an extra layer of protection. This proactive strategy serves multiple key purposes, including safeguarding intellectual property, mitigating model vulnerabilities, and preventing adversarial attacks. In this work, we propose a novel fingerprinting attack that identifies a given ML model's architecture family, from among the latest categories. To this aim, we are the first to leverage a Frequency Throttling Side-Channel Attack, a method that enables us to convert power side-channel information into timing variations at the user-space level. We utilize the timing information of crafted adversary kernels combined with a supervised machine learning classifier to identify the ML model architecture. In particular, our proposed method involves capturing timing information by monitoring an adversary kernel's execution time while a specific ML model runs, unveiling distinctive timing patterns. This process involves initiating the frequency throttling side-channel effect and transforming it into timing information. Subsequently, we employ a specialized machine learning classifier trained on this timing data to precisely identify the victim's ML model architecture. With this approach, we achieve 98% accuracy in correctly classifying a known ML model into its corresponding architecture family. Furthermore, our attack demonstrates transferability by accurately assigning the correct family to unseen models with 90.6% accuracy on average. Additionally, for the purpose of thorough analysis, we have reproduced this attack across 3 different platforms, with comparable results underscoring the attack's platform portability. Finally, it is notable that we intend to publicly release our work, making it accessible to the research community for the purpose of reproducibility.	Najmeh Nazari (UC Davis); Chongzhou Fang (University of California, Davis); Hosein Mohammadi Makrani (University of California, Davis); Behnam Omidi (George Mason University); Mahdi Eslamimehr (UCLA); Setareh Rafatirad (University of California Davis); Avesta Sasan (UC Davis); Hossein Sayadi (California State University, Long Beach); Khaled N. Khasawneh (George Mason University); Houman Homayoun (University of California Davis)
982	Hyb-Learn: A Framework for On-Device Self-Supervised Continual Learning with Hybrid RRAM/SRAM Memory	While RRAM crossbar-based In-Memory Computing (IMC) has proven highly effective in accelerating Deep Neural Networks (DNNs) inference, RRAM-based on-device training is less explored due to its high energy consumption of weight re-programming and cells' low endurance problem. Besides, emerging trends indicate a need for on-device continual learning which sequentially acquires knowledge from multiple tasks to enhance user's experiences and eliminate data privacy concerns. However, learning on each new task leads to forgetting prior learned knowledge on prior tasks, which is known as catastrophic forgetting. To address these challenges, we are the first to propose a novel training framework, Hyb-Learn, for enabling on-device continual learning with a hybrid RRAM/SRAM IMC architecture design. Specifically, when training each new arriving task, our approach first partitions the model into two groups based on the proposed task-correlated PE-wise correlation to freeze or re-training, and correspondingly mapping to RRAM and SRAM, respectively. In practice, the RRAM stores frozen weights with strong task correlation to prior tasks to eliminate the high cost of weight reprogramming issue of RRAM, while the SRAM stores the remaining weights that will be updated. Furthermore, to maximize the freezing ratio for improving training efficiency while maintaining accuracy and mitigating catastrophic forgetting, we incorporate self-supervised learning algorithms that are initialized from a pre-trained model for training each new task.	Fan Zhang (Johns Hopkins University); Li Yang (University of North Carolina at Charlotte); Deliang Fan (Johns Hopkins University)
989	SpARC: Token Similarity-Aware Sparse Attention Transformer Accelerator via Row-wise Clustering	In this paper, we propose SpARC, a sparse attention transformer accelerator that enhances throughput and energy efficiency by reducing the computational complexity of the self-attention mechanism. Our approach exploits inherent row-level redundancies in transformer attention maps to reduce the overall self-attention computation. By employing row-wise clustering, attention scores are calculated only once per cluster to achieve approximate attention without seriously compromising accuracy. To leverage the high parallelism of the proposed clustering approximate attention, we develop a fully pipelined accelerator with a dedicated memory hierarchy.	Han Cho (Korea University); Dongjun Kim (Korea University); Seungeon Hwang (Korea University); Jongsun Park (Korea University)

Submissio	Title	Abstract	Authors with Affiliations
991	Efficient Memory Integration: MRAM-SRAM Hybrid Accelerator for Sparse On-Device Learning	With the prosperous development of Deep Neural Network (DNNs), numerous Process-In-Memory (PIM) designs have emerged to accelerate DNN models with exceptional throughput and energy-efficiency. PIM accelerators based on Non-Volatile Memory (NVM) or volatile memory offer distinct advantages for computational efficiency and performance. NVM based PIM accelerators, demonstrated success in DNN inference, face limitations in on-device learning due to high write energy, latency, and instability. Conversely, fast volatile memories, like SRAM, offer rapid read/write operations for DNN training, but suffer from significant leakage currents and large memory footprints. In this paper, for the first time, we present a fully-digital sparse processing in hybrid NVM-SRAM design, synergistically combines the strengths of NVM and SRAM, tailored for on-device continual learning. Our designed NVM and SRAM based PIM circuit macros could support both storage and processing of N:M structured sparsity pattern, significantly improving the storage and computing efficiency. Exhaustive experiments demonstrate that our hybrid system effectively reduces area and power consumption while maintaining high accuracy, offering a scalable and versatile solution for on-device continual learning.	Fan Zhang (Johns Hopkins University); Amitesh Sridharan (Arizona State University); Wilman tsai (Stanford); Yiran Chen (Duke University); Shan X. Wang (Stanford University); Deliang Fan (Johns Hopkins University)
995	Accel-NASBench: Sustainable Benchmarking for Accelerator-Aware NAS	One of the primary challenges impeding the progress of Neural Architecture Search (NAS) is its extensive reliance on exorbitant computational resources. NAS benchmarks aim to simulate runs of NAS experiments at zero cost, remediating the need for extensive compute. However, existing NAS benchmarks use synthetic datasets and model proxies that make simplified assumptions about the characteristics of these datasets and models, leading to unrealistic evaluations. We present a technique that allows searching for training proxies that reduce the cost of benchmark construction by significant margins, making it possible to construct realistic NAS benchmarks for large-scale datasets. Using this technique, we construct an open-source bi-objective NAS benchmark for the ImageNet2012 dataset combined with the on-device performance of accelerators, including GPUs, TPUs, and FPGAs. Through extensive experimentation with various NAS optimizers and hardware platforms, we show that the benchmark is accurate and allows searching for state-of-the-art hardware-aware models at zero cost.	Afzal Ahmad (Hong Kong University of Science and Technology); Linfeng Du (The Hong Kong University of Science and Technology); Zhiyao Xie (Hong Kong University of Science and Technology); Wei Zhang (Hong Kong University of Science and Technology)
998	AdvHunter: Detecting Adversarial Perturbations in Black-Box Neural Networks through Hardware Performance Counters	The paper introduces AdvHunter, a novel strategy to detect adversarial examples (AEs) in Deep Neural Networks (DNNs). AdvHunter operates effectively in practical black-box scenarios, where only hard-label query access is available, a situation often encountered with proprietary DNNs. This differentiates it from existing defenses, which usually rely on white-box access or need to be integrated during the training phase - requirements often not feasible with proprietary DNNs. AdvHunter functions by monitoring data flow dynamics within the computational environment during the inference phase of DNNs. It utilizes Hardware Performance Counters to monitor microarchitectural activities and employs principles of Gaussian Mixture Models to detect AEs. Extensive evaluation across various datasets, DNN architectures, and adversarial perturbations demonstrate the effectiveness of AdvHunter.	Manaar Alam (New York University Abu Dhabi); Michail Maniatakos (New York University Abu Dhabi)
1009	FinerDedup: Sifting Fingerprints for Efficient Data Deduplication on Mobile Devices	Data deduplication is promised to extend the lifetime and capacity of storage on mobile devices. However, existing data deduplication works show high memory consumption and indexing costs for maintaining a fingerprint for each data block, especially when the duplicate ratio of data blocks on mobile systems is about 10% to 30%. In this paper, we propose a novel approach called FinerDedup to optimize the memory costs and retrieval efficiency of data deduplication. FinerDedup drastically reduces the number of fingerprints by screening out the duplicate data blocks via random forest and Bloom filter. We implement FinerDedup on real mobile devices with Android 10 and evaluate it with real workloads. Extensive experimental results show that FinerDedup can reduce 85% of fingerprints and 20% of I/O latency over the widely-used DmDedup.	Xianzhang Chen (Chongqing University); Xingjie Zhou (Chongqing University); Wei Li (Chongqing University); Xi Yu (Chongqing University); Duo Liu (Chongqing University); Yujuan Tan (Chongqing University); Ao Ren (Chongqing University)
1011	FRM-CIM: Full-Digital Recursive MAC Computing in Memory System Based on MRAM for Neural Network Applications	Computing in memory (CIM) realizes energy-efficient neural network algorithms by implementing highly parallel multiply-and-accumulate (MAC) operation. However, the MAC delay of CIM will sharply increase with the improvement of computing precision, which restricts its development. In this work, we propose a full-digital recursive MAC (FRM) operation based on spin-transfer-torque magnetic random access memory (STT-MRAM) CIM system to enable fast and energy-efficient image recognition application. First, the fast FRM scheme is proposed by utilizing the recursive operations of read and addition in segmented bit-line array, which effectively reduces the delay of MAC operations to 3.5ns and 4ns for 8-bit and 16-bit input and weight precision, respectively. Second, we design an image recognition system using FRM-CIM architecture as the processing element (PE), where the adaptive pruning method for layers is proposed to improve the compatibility of it with the neural network. By performing image recognition for the MNIST and CIFAR-10 datasets, results show that the throughput and energy efficiency of the FRM-CIM system are 58.51TOPS/mm2 and 11.3–56.72 TOPS/W under 8–16-bit precision, which are improved by 4.3 times and 2.6 times compared with the state-of-the-art works. Finally, the recognition accuracy can reach 96.65% and 82.7% on MNIST and CIFAR-10, respectively.	Jinkai Wang (Beihang University); Zekun Wang (Beihang University); Bojun Zhang (Beihang University); Zhengkun Gu (Beihang University); Youxiang Chen (Beihang University); Weisheng Zhao (Beihang University); Yue Zhang (Beihang University)
1015	Graph Neural Networks Automated Design and Deployment on Device-Edge Co-Inference Systems	The key to device-edge co-inference paradigm is to partition models into computation-friendly and computation-intensive parts across device and edge, respectively. However, for Graph Neural Networks (GNNs), partitioning without architecture exploration is ineffective due to various computational-communication overheads of GNN operations over heterogeneous devices. We present GCoDE, the first automatic framework that co-designs the GNN architecture and operation mapping. GCoDE abstracts communication process into explicit operation, fuses architecture search and operations mapping in a joint-optimization space. Also, the performance-awareness approach enables effective evaluation of architecture efficiency. Experiments show GCoDE achieves up to 44.9x speedup and 98.2% energy reduction across various systems.	Ao Zhou (Beihang University); Jianlei Yang (Beihang University); Tong Qiao (Beihang University); Yingjie Qi (Beihang University); Zhi Yang (Peking University); Weisheng Zhao (Beihang University); Chunming Hu (Beihang University)
1027	Whisper: Timing the Transient Execution to Leak Secrets and Break KASLR	The vulnerabilities of transient execution have been exploited in many side-channel attacks (SCA). We report Whisper, a novel transient execution timing (TET) side channel, which is based on the execution time difference of transient execution under different conditions. We develop TET version of SCAs including Meltdown, Zombieload, and Spectre-RSB that use Whisper as covert channel to leak information. We further propose TET-KASLR to break the kernel address space layout randomization (KASLR) mechanism under the protection of KPTI and FLARE. These attacks are simple to implement and can bypass the existing mitigation methods because the TET side channel relies on execution time that can be conveniently obtained by architectural level timing analysis. We demonstrate the correctness and effectiveness of these attacks on various x86-64 CPUs. The root cause of Whisper is analyzed with our toolset built on performance monitor unit (PMU) and potential defense against Whisper is also discussed.	Yu Jin (Beijing University of Posts and Telecommunications); Chunlu Wang (Beijing University of Posts and Telecommunications); Pengfei Qiu (Tsinghua University); Chang Liu (Department of Computer Science and Technology, Tsinghua University); Yihao Yang (Beijing University of Posts and Telecommunications); Hongpei Zheng (Tsinghua University); Yongqiang Lyu (Tsinghua University); Xiaoyong Li (Beijing University of Posts and Telecommunications); Gang Qu (Univ. of Maryland, College Park); Dongsheng Wang (Tsinghua University)

Submissio	Title	Abstract	Authors with Affiliations
1029	ModSRAM: Algorithm-Hardware Co-Design for Large Number Modular Multiplication in SRAM	Elliptic curve cryptography (ECC) is widely used in security applications such as public key cryptography (PKC) and zero-knowledge proofs (ZKP). ECC is composed of modular arithmetic, where modular multiplication takes most of the processing time. Computational complexity and memory constraints of ECC limit the performance. Therefore, hardware acceleration on ECC is an active field of research. Processing-in-memory (PIM) is a promising approach to tackle this problem. In this work, we design ModSRAM, the first 8T SRAM PIM architecture to compute large-number modular multiplication efficiently. In addition, we propose R4CSA-LUT, a new algorithm that reduces the cycles for an interleaved algorithm and eliminates carry propagation for addition based on look-up tables (LUT). ModSRAM is co-designed with R4CSA-LUT to support modular multiplication and data reuse in memory with 52% cycle reduction compared to prior works with only 32% area overhead.	Jonathan Hao-Cheng Ku (Duke University); Junyao Zhang (Duke University); Haoxuan Shan (Duke University); Saichand Samudrala (Texas A&M University); Jiawen Wu (Texas A&M University); Qilin Zheng (Duke University); Ziru Li (Duke University); Jeyavijayan Rajendran (Texas A&M University); Yiran Chen (Duke University)
1031	PowerRChol: Efficient Power Grid Analysis Based on Fast Randomized Cholesky Factorization	Efficient power grid analysis is critical in modern VLSI design. It is computationally challenging because it requires solving large linear equations with millions of unknowns. Iterative solvers are more scalable, but their performance relies on preconditioners. Existing preconditioning approaches suffer from either high construction cost or slow convergence rate, both resulting in unsatisfactory total solution time. In this work, we propose an efficient power grid simulator based on fast randomized Cholesky factorization, named PowerRChol. We first propose a randomized Cholesky factorization algorithm with provable linear-time complexity. Then we propose a randomized factorization oriented matrix reordering approach. Experimental results on large-scale power grids demonstrate the superior efficiency of PowerRChol over existing iterative solvers, showing 1.51X, 1.93X and 3.64X speedups on average over the original RChol [3], feGRASS [11] and AMG [14] based PCG solvers, respectively. For instance, a power grid matrix with 60 million nodes and 260 million nonzeros can be solved (at a 1E-6 accuracy level) in 148 seconds on a single CPU core.	Zhiqiang Liu (Tsinghua University); Wenjian Yu (Tsinghua University)
1036	FNM-Trans: Efficient FPGA-based Transformer Architecture with Full N:M Sparsity	Transformer models have become popular in various AI applications due to their exceptional performance. However, their impressive performance comes with significant computing and memory costs, hindering efficient deployment of Transformer-based applications. Many solutions focus on leveraging sparsity in weight matrix and attention computation. However, previous studies fail to exploit unified sparse pattern to accelerate all three modules of Transformer (QKV generation, attention computation, FFN). In this paper, we propose FNM-Trans, an adaptable and efficient algorithm-hardware co-design aimed at optimizing all three modules of the Transformer by fully harnessing N:M sparsity. At the algorithm level, we fully explore the interplay of dynamic pruning with static pruning under high N:M sparsity. At the hardware level, we develop a dedicated hardware architecture featuring a custom computing engine and a softmax module, tailored to support varying levels of N:M sparsity. Experiment results show that, our algorithm optimizes accuracy by 11.03% under 2:16 attention sparsity and 4:16 weight sparsity, compared to other methods. Additionally, FNM-Trans achieves speedups of 27.13× and 21.24× over Intel i9-9900X and NVIDIA RTX 2080 Ti, respectively, and outpaces current FPGA-based Transformers by 1.88× to 36.51×.	Manting Zhang (Fudan University); Jialin Cao (Fudan University); Kejia Shi (Fudan University); Keqing Zhao (Fudan University); Genhao Zhang (Fudan University); Jun Yu (Fudan Universiy); Kun Wang (Fudan University)
1037	Minimizing Labeling, Maximizing Performance: A Novel Approach to Nanoscale Scanning Electron Microscope (SEM) Defect Segmentation	In semiconductor manufacturing, pinpointing nanoscale wafer defects is crucial for yield and reliability. Deep learning methods for defect segmentation rely heavily on large, labor-intensive datasets and focus mainly on macroscopic wafer defects, not nanoscale morphology. Our research introduces a hybrid weakly supervised scanning electron microscope (SEM) defect segmentation system with two sub-networks: one for accurate defect localization and image cropping, another for detailed segmentation. Validated on 1,328 SEM image defects from a real facility, our model surpasses existing weakly supervised methods and equals fully supervised models in accuracy, with 10% labeling effort, providing a novel approach for high-precision defect segmentation.	Yibo Qiao (Zhejiang University); Weiping Xie (Zhejiang University); Shunyuan Lou (Zhejiang University); Qian Jin (Zhejiang University); Lichao Zeng (University of Science and Technology of China); yining chen (Zhejiang University); QI SUN (Zhejiang University); Cheng Zhuo (Zhejiang University)
1039	Series-Parallel Hybrid SOT-MRAM Computing-in-Memory Macro with Multi-Method Modulation for High Area and Energy Efficiency	MRAM is one of the most promising candidates for CIM. This paper proposes a series-parallel hybrid SOT-MRAM-CIM macro to solve the shortcomings of existing MRAM-CIM structures, like high energy cost and low operating frequency in traditional parallel or serial architecture. Additionally, we incorporate a multi-method modulation scheme, allowing for configurable precision (2/4/6/8-bit). We experimentally verified the performance of SOT-MRAM devices at 180-nm process node and design the macro at 28-nm node based on the test parameters of fabricated SOT devices. The simulation shows this macro can achieve energy efficiency of 23.7~29.6-Tops/W and computing frequency of 164.5-MHz/Bit at 8-bit precision.	Weiliang Huang (Beihang University); Jinyu Bai (Beihang University); Wang Kang (Beihang University); Zhaohao Wang (Beihang University); kaihua cao (Beihang University); hongxi liu (Truth Memory Corporation); He Zhang (Beihang University); Weisheng Zhao (Beihang University)
1041	Fused Sampling and Grouping with Search Space Reduction for Efficient Point Cloud Acceleration	Point-based deep neural networks have demonstrated remarkable ability in analyzing point cloud. However, challenges arise in sampling and grouping layers, particularly in terms of time and energy consumption. In this paper, we introduce a Morton code-based data structure which stores point data with the shared upper bits together. We also propose a fused sampling and grouping approach with a reduced search space, which reuses the point data and the calculated distances. Additionally, a dedicated hardware supporting the proposed method is introduced. Experimental results show that our approach effectively reduces the number of calculations and data accesses with negligible accuracy loss.	Hyunsung Yoon (Pohang University of Science and Technology); Jae-Joon Kim (Seoul National University)
1053	Efficient ILT via Multigrid-Schwartz Method	Inverse lithography technology (ILT) is one of most powerful resolution enhancement technologies (RETs) used in chip manufacturing. Due to the high computational requirements of ILT, large layouts are often split into smaller tiles and then assembled to obtain the final result. This paper states the challenges that may emerge during layout assembly and proposes to use the multigrid Schwarz method to address these issues. Experimental results show that our method achieves comparable quality results to full-chip correction and exhibits better efficiency.	Shuyuan Sun (Fudan University); Fan Yang (Fudan University); Bei Yu (The Chinese University of Hong Kong); Li Shang (fudan university); Dian Zhou (Fudan University); Xuan Zeng (Fudan University)
1055	Enabling Low Latency for ECQF based Flow Aggregation Scheduling in Time-Sensitive Networking	Cycle Queuing and Forwarding (CQF) configures the same cycle length on the flow path, resulting in certain flows unschedulable. Enhanced CQF (ECQF) based flow aggregation utilizes variable cycle length to address this issue. However, it remains a conceptual model without a concrete implementation. In this paper, we propose a jointly optimize aggregation cycle and flows' offsets (JACO) mechanism to achieve ECQF-based flow aggregation. We also design an incremental heuristic algorithm for JACO. Finally, we evaluate the performance of JACO in different scenarios using OMNet++ simulation platform. Compared with ECQF, the results show that JACO reduces latency and improves resource utilization.	Ping Liu (Lanzhou University); Tong Zhang (Nanjing University of Aeronautics and Astronautics); Xiaoqin Feng (Lanzhou University); Yanying Ma (Lanzhou University); Fengyuan Ren (Lanzhou University/Tsinghua University)
1061	LaMUX: Optimized Logic-Gate-Enabled High-Performance Microfluidic Multiplexer Design	To meet the increasingly complex experimental demands, the number of microvalves in flow-based microfluidic biochips has increased significantly, making it necessary to adopt multiplexers (MUXes) to actuate microvalves. However, existing MUX designs have limited coding capacities, resulting in excessive chip-to-world interface. This paper proposes a novel gate structure for modifying the current MUX architecture, along with a mixed coding strategy achieving the maximum coding capacity within the modified architecture. Additionally, a synthesis tool for the mixed-coding-based MUXes (LaMUXes) is presented. Experimental results demonstrate that the LaMUX is exceptionally efficient, substantially reducing the usage of pneumatic controllers and microvalves in MUXes.	Siyuan Liang (The Chinese University of Hong Kong); Yushen Zhang (Technical University of Munich); Rana Altay (Santa Clara University); Hudson Gasvoda (Santa Clara University); Mengchu Li (Technical University of Munich); Ismail Emre Araci (Santa Clara University); Tsun-Ming Tseng (Technical University of Munich); Ulf Schlichtmann (Technical University of Munich); Tsung-Yi Ho (The Chinese University of Hong Kong)

Submissio	Title	Abstract	Authors with Affiliations
1062	Multi-order Differential Neural Network for TCAD Simulation of the Semiconductor Devices	Technology Computer Aided Design (TCAD) is a crucial step in the design and manufacturing of semiconductor devices. It involves solving physical equations that describe the behavior of semiconductor devices to predict various device parameters. Traditional TCAD methods, such as finite volume and finite element methods, discretize relevant physical equations to achieve numerical simulations of devices, significantly burdening the computation resources. For the first time, this paper proposes a novel method for TCAD simulation based on Physics-Informed Neural Networks (PINNs). We proposed Multi-order Differential Neural Network (MDNN), an improved Radial Basis Function Neural Network (RBFNN) model. By training MDNN, it achieves the coupled solution of the Poisson equation and drift-diffusion equation under steady-state conditions, without the need for a pre-existing dataset. To the best of our knowledge, this marks the first instance of an ML-TCAD simulation that does not require any pre-existing data. For an example of PN junction diode, this method effectively simulates the basic physical characteristics of the device, with a self-consistent solution error of less than 1×10^-5.	Zifei Cai (Huazhong University of Science and Technology); AnAoxue Huang (Huazhong University of Science and Technology); Yifeng Xiong (Huazhong University of Science and Technology); Dejiang Mu (Huazhong University of Science and Technology); Xiangshui Miao (Huazhong University of Science and Technology); Xingsheng Wang (Huazhong University of Science and Technology)
1067	LOTUS: learning-based online thermal and latency variation management for two-stage detectors on edge devices	Two-stage object detectors exhibit high accuracy and precise localization, especially for identifying small objects that are favorable for various edge applications. However, the high computation costs associated with two-stage detection methods cause more severe thermal issues on edge devices, incurring dynamic runtime frequency change and thus large inference latency variations. Furthermore, the dynamic number of proposals in different frames leads to various computations over time, resulting in further latency variations. The significant latency variations of detectors on edge devices can harm user experience and waste hardware resources. To avoid thermal throttling and provide stable inference speed, we propose LOTUS, a novel framework that is tailored for two-stage detectors to dynamically scale CPU and GPU frequencies jointly in an online manner based on deep reinforcement learning. To demonstrate the effectiveness of LOTUS, we implement it on NVIDIA Jetson Orin Nano and Mi 11 Lite mobile platforms. The results indicate that LOTUS can consistently and significantly reduce latency variation, achieve faster inference, and maintain lower CPU and GPU temperatures under various settings.	Yifan Gong (Northeastern University); Yushu Wu (Northeastern University); PU ZHAO (Northeastern.edu); zheng zhan (Northeastern University); Liangkai Liu (University of Michigan); Chao Wu (Chinese Academy of Sciences); Xulong Tang (University of Pittsburgh); Yanzhi Wang (Northeastern University)
1080	A Holistic Functionalization Approach to Optimizing Imperative Tensor Programs in Deep Learning	As deep learning empowers various fields, many new operators have been proposed to improve the accuracy of deep learning models. Researchers often use imperative programming diagrams (PyTorch) to express these new operators, leaving the fusion optimization of these operators to deep learning compilers. Unfortunately, the inherent side effects introduced by imperative tensor programs, especially tensor-level mutations, often make optimization extremely difficult. We present a holistic functionalization approach (TensorSSA) to optimizing imperative tensor programs beyond control flow boundaries. We achieve a 1.79X (1.34X on average) speedup in representative deep learning tasks than state-of-the-art works.	Jinming Ma (Shanghai Artificial Intelligence Laboratory); Xiuhong Li (Peking University, Beijing); Zihan Wang (Shanghai Jiao Tong University); Xingcheng Zhang (SenseTime & Shanghai Artificial Intelligence Laboratory); Shengen Yan (National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing, China); Yuting Chen (Shanghai Jiao Tong University); Yueqian Zhang (Shanghai Artificial Intelligence Laboratory); Minxi Jin (Shanghai Artificial Intelligence Laboratory); Lijuan Jiang (Shanghai Artificial Intelligence Laboratory); Yun (Eric) Liang (Peking University); Chao Yang (National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing, China); Dahua Lin (The Chinese University of Hong Kong & Shanghai Artificial Intelligence Laboratory)
1096	ElasticZRAM: Revisiting ZRAM for Swapping on Mobile Devices	Modern mobile devices adopt two-level memory swapping consisting of ZRAM and storage devices to relieve memory pressure. In the swap subsystem, ZRAM can improve application responsiveness and reduce write traffic to storage devices while consuming physical memory and additional CPU cycles. To better utilize ZRAM and improve system performance, we propose ElasticZRAM, an elastic ZRAM to redesign the traditional memory swapping with full awareness of the characteristics of applications and NAND flash-based storage devices on mobile devices. Experimental results on Google Pixel 6 demonstrate that ElasticZRAM improves application response time by up to 24.8\% with negligible overhead compared with state-of-the-arts.	Wentong Li (Software/Hardware Co-design Engineering Research Center, Ministry of Education, and School of Computer Science and Technology, East China Normal University); Dingcui Yu (Software/Hardware Co-design Engineering Research Center, Ministry of Education, and School of Computer Science and Technology, East China Normal University); Yunpeng Song (Software/Hardware Co-design Engineering Research Center, Ministry of Education, and School of Computer Science and Technology, East China Normal University); Longfei Luo (Software/Hardware Co-design Engineering Research Center, Ministry of Education, and School of Computer Science and Technology, East China Normal University); Liang Shi (Software/Hardware Co-design Engineering Research Center, Ministry of Education, and School of Computer Science and Technology, East China Normal University)
1098	SpaHet: A Software/Hardware Co-design for Accelerating Heterogeneous-Sparsity based Sparse Matrix Multiplication	Sparse general matrix-matrix multiplication is widely used in data mining applications. Its irregular memory access patterns limit the performance of general-purpose processors, thus motivating many FPGA-based hardware innovations. Nevertheless, existing accelerators fail to efficiently support heterogeneous input matrix sparsity, which is universal in various real-world applications. With in-depth experimental analysis, we observe that their performance is bottlenecked by their fixed tiling mechanisms, which only alleviate the irregularity of one input matrix. Based on the observation, we propose SpaHet, a software/hardware co-design to accelerate heterogeneous-sparsity based sparse matrix multiplication. Our experimental results show that SpaHet outperforms state-of-the-art FPGA-based solutions by 2.74× in performance.	Haoqin Huang (Huazhong University of Science and Technology); Pengcheng Yao (Huazhong University of Science and Technology); Zhaozeng An (Huazhong University of Science and Technology); Yufei Sun (Huazhong University of Science and Technology); Ao Hu (Huazhong University of Science and Technology); Peng Xu (Huazhong University of Science and Technology); Long Zheng (Huazhong University of Science and Technology); Xiaofei Liao (Huazhong University of Science and Technology); Hai Jin (Huazhong University of Science and Technology)
1116	High-Performance and Resource-Efficient Dynamic Memory Management in High-Level Synthesis	The usability and popularity of high-level synthesis (HLS) tools are still limited due to lack of support for dynamic memory management (DMM). Though HLS-compatible DMM solutions have been proposed recently, nevertheless, based on our investigation, none of them can hit high performance (i.e., minimal memory (de-)allocation latency) and resource efficiency (i.e., managing arbitrarily sized memory with minimal FPGA resource consumption) with one stone, seriously limiting their practicality. In response, we propose HeroDMM, a high-performance and resource-efficient dynamic memory manager for HLS. Results show that HeroDMM outperforms state-of-the-art HLS-compatible DMM solutions by 61.69%--99.99% in performance improvement and 23.79%--97.22% in resource consumption savings.	Qinggang Wang (Huazhong University of Science and Technology); Long Zheng (Huazhong University of Science and Technology); Zhaozeng An (Huazhong University of Science and Technology); Haoqin Huang (Huazhong University of Science and Technology); Haoran Zhu (Huazhong University of Science and Technology); Yu Huang (Huazhong University of Science and Technology); Pengcheng Yao (Huazhong University of Science and Technology); Xiaofei Liao (Huazhong University of Science and Technology); Hai Jin (Huazhong University of Science and Technology)
1117	TATOO: A Flexible Hardware Platform for Binary-Only Fuzzing	Hardware-based tracing, being efficient, can be a good alternative to the computationally-expensive software-based instrumentation in binary-only greybox fuzzing. However, it only records all branches within a specified address range, lacking the flexibility to re-filter them. This paper introduces TATOO, a hardware platform employing tagged architectures and hardware tracing to enhance binary-only fuzzing. TATOO stands out by enabling users to tag instructions at the instruction level, significantly reducing the volume of traced data and improving fuzzing efficiency. TATOO also supports recording the dataflow information for smart mutations. Implemented on a real hardware FPGA platform, TATOO demonstrates a mere 8.7% performance overhead.	Jinting Wu (Southern University of Science and Technology); Haodong Zheng (Southern University of Sciencec and Technology); Yu Wang (Southern University of Science and Technology); Tai Yue (Southern University of Science and Technology); Fengwei Zhang (Southern University of Science and Technology)

Submissio	Title	Abstract	Authors with Affiliations
1124	PMP: Pattern Morphing-based Memory Partitioning in High-Level Synthesis	Memory partitioning is a widely used technique to reduce access conflicts on multi-bank memory in high-level synthesis. Previous memory partitioning methods mainly focus on a given access pattern extracted from stencil applications. Restricted by the pattern shape, these methods are prone to sub-optimal bank numbers or large overhead on address generation. In this work, we propose a pattern-morphing-based memory partitioning method, PMP, that only requires reduced hyperplane families to achieve the minimal bank number. To reduce the side effect of extra data padding, an integer linear programming problem is formulated for pattern morphing. Compared to the previous hyperplane-based memory partitioning, the experimental results show that our approach could achieve the optimal partition factor while saving 22% in LUTs, 21% in FlipFlops, 10% in DSPs, and 40% in memory overhead, on average.	Dajiang Liu (Chongqing University); Decai Pan (Chongqing University); Xiao Xiong (Chongqing University); Jiaxing Shang (Chongqing University); Shouyi YIN (Tsinghua University)
1130	Graph Learning-based Fault Criticality Analysis for Enhancing Functional Safety of E/E Systems	The increasing complexity of Electrical and Electronic (E/E) systems underscores the need for protective measures to ensure functional safety (FuSa) in high-assurance environments. This entails the identification and fortification of vulnerable nodes to enhance system reliability during mission-critical scenarios. Traditionally, the assessment of E/E system reliability has relied on fault injection (FI) techniques and simulations. However, FI faces challenges in coping with escalating design complexity, including resource demands and timing overheads. Furthermore, it falls short in identifying critical components that may lead to functional failures. To address these challenges, we propose a Machine Learning (ML)-based framework for predicting critical nodes in hardware designs. The process begins with constructing a graph from the design netlist, forming the foundation for training a Graph Convolutional Network (GCN). The GCN model utilizes graph node attributes, node labels, and edge connections to learn and predict critical nodes in the circuit. The model furnishes up to 93.7% accuracy in identifying vulnerable circuit nodes during evaluation on diverse designs such as Synchronous Dynamic Random Access Memory (SDRAM) controller, OpenRISC 1200 (OR1200) modules. Furthermore, we incorporate an explainability analysis to interpret individual node predictions. This analysis discerns the critical design factors influencing fault criticality in the design. Moreover, to the best of our knowledge, we, for the first time, perform a regression analysis to generate node criticality scores, quantifying the degrees of criticality, that can enable prioritizing resources towards critical nodes.	Sanjay Das (University of Texas at Dallas); Shamik Kundu (University of Texas at Dallas); Pooja Madhusoodhanan (Texas Instruments); Prasanth Viswanathan Pillai (Texas Instruments); Rubin Parekhji (Texas Instruments); ARNAB RAHA (Intel Corporation); Suvadeep Banerjee (Intel Labs, Intel); Suriya Natarajan (Intel Corporation); Kanad Basu (University of Texas at Dallas)
1147	Planaria: Pattern Directed Cross-page Composite Prefetcher	Due to the memory wall, memory system performance significantly impacts the user experience of mobile phones. The system cache (SC) locates on the memory side and is shared by all the central processing units (CPUs) and graph processing units (GPUs) within the mobile phone and is the last defense line before resorting to the time-consuming off-chip memory access. However, it is challenging to manage SC, due to the memory-side large working set and irregular accessing patterns. Although SC takes up a considerable on-chip area, the effectiveness of SC in terms of hit rate is rather low. It is observed that neither using the state-of-the-art cache replacement policies nor enlarging cache size can significantly benefit SC. The prefetchers designed for higher-level caches cannot be used by SC, because the required program counter (PC) is not available on the memory-side and/or the aggressive prefetch traffic violates the stringent power constraints of mobile phones. In this study, we propose Planaria, which includes two sub-prefetchers (SLP and TLP) and a coordinator (POC) to simultaneously achieve high accuracy and coverage of prefetching. The two sub-prefetchers exploit the intra- and inter-page regularities via self and transfer learning, respectively. The coordinator POC explicitly decouples the learning and issuing phases of the sub-prefetchers. The sub-prefetchers are directed by the full pattern, but are enabled in an irreversible order. The working fashion of "parallel training and serial issuing" effectively increases useful prefetches and reduces useless prefetches. Experimental results show that, Planaria has improved the overall system performance in terms of instructions per cycle (IPC) by 28.9%, 21.9% and 15.3% on average over no prefetcher and BOP and SPP, respectively. Moreover, Planaria only incurs 0.5% power consumption overhead, while BOP and SPP increase the power consumption by 13.5% and 9.7%, respectively.	Yuhang Liu (ICT, CAS); Mingyu Chen (ICT, CAS)
1153	Predicting Lemmas in Generalization of IC3	The IC3 algorithm, also known as PDR, has made a significant impact in the field of safety model checking in recent years due to its high efficiency, scalability, and completeness. The most crucial component of IC3 is inductive generalization, which involves dropping variables one by one and is often the most time-consuming step. In this paper, we propose a novel approach to predict a possible minimal lemma before dropping variables by utilizing the counterexample to propagation (CTP). By leveraging this approach, we can avoid dropping variables if predict successfully. The comprehensive evaluation demonstrates a commendable success rate in lemma prediction and a significant performance improvement achieved by our proposed method.	Yuheng Su (University of Chinese Academy of Sciences; Institute of Software, Chinese Academy of Sciences); Qiusong Yang (Institute of Software, Chinese Academy of Sciences); Yiwei Ci (Institute of Software, Chinese Academy of Sciences)
1157	Towards Efficient SRAM-PIM Architecture Design by Exploiting Unstructured Bit-Level Sparsity	Bit-level sparsity in neural network models harbors immense untapped potential. Eliminating redundant calculations of randomly distributed zero-bits significantly boosts computational efficiency. Yet, traditional digital SRAM-PIM architecture, limited by rigid crossbar architecture, struggles to effectively exploit this unstructured sparsity. To address this challenge, we propose Dyadic Block PIM (DB-PIM), a novel algorithm-architecture co-design framework. It preserves the random distribution of non-zero bits to maintain accuracy while restricting the number of non-zero bits in each weight of the filter to improve regularity. DB-PIM improves both performance and energy efficiency, achieving a remarkable speedup of up to 6.53x and energy savings of 77.50%.	Cenlin Duan (Beihang University); Jianlei Yang (Beihang University); Yiou Wang (Beihang University); Yikun Wang (Beihang University); Yingjie Qi (Beihang University); Xiaolin He (Beihang University); Bonan Yan (Peking University); Xueyan Wang (Beihang University); Xiaotao Jia (School of Integrated Circuit Science and Engineering, Beihang University); Weisheng Zhao (Beihang University)
1159	zeroTT: A Two-Step State Transition Avoidance Scheme for MLC STT-RAM	Compared with conventional SRAM, Spin-Transfer Torque Random Access Memory(STT-RAM) is expected to play a crucial role in future memory technologies with the increasing demands for higher storage density and lower power consumption for modern embedded systems. Moreover, Multi-Level Cell (MLC) STT-RAM outperforms Single-Level Cell (SLC) STT-RAM since it can store multiple bits per cell. However, MLC STT-RAM suffers from the occurrence of two-step state transitions (TTs) due to additional flipping of soft domains. Existing approaches mitigate this problem by reducing TTs with data coding. However, none of them can eliminate all the TTs. In this work, we propose a two-step transition avoidance scheme, referred to as zeroTT, for MLC STT-RAM. We show why the existing (2,3)-based coding methods cannot avoid TTs. Then, we refine the problem of expansion coding and present how to find zeroTT coding methods. Lastly, we propose an optimal (3,4)-based coding method considering the issues of space overhead and coding complexity. The experimental results demonstrate that zeroTT can completely avoid TTs, leading to a more efficient MLC STT-RAM in terms of latency, energy consumption, and lifetime.	Dong Yin (Hunan University); Huizhang Luo (Hunan University); Jeff Zhang (Arizona State University); Mingxing Duan (Hunan University); Wangdong Yang (Hunan University); Zhuo Tang (Hunan University); Kenli Li (Hunan University)

Submissio	Title	Abstract	Authors with Affiliations
1164	UpDLRM: Accelerating Personalized Recommendation using Real-World PIM Architecture	Deep Learning Recommendation Models (DLRMs) have gained popularity in recommendation systems due to their effectiveness in handling large-scale recommendation tasks. The embedding layers of DLRMs have become the performance bottleneck due to their intensive needs on memory capacity and memory bandwidth. In this paper, we propose UpDLRM, which utilizes real-world processing-in-memory (PIM) hardware, UPMEM DPU, to boost the memory bandwidth and reduce recommendation latency. The parallel nature of the DPU memory can provide high aggregated bandwidth for the large number of irregular memory accesses in embedding lookups, thus offering great potential to reduce the inference latency. To fully utilize the DPU memory bandwidth, we further studied the embedding table partitioning problem to achieve good workload-balance and efficient data caching. Evaluations using real-world datasets show that, UpDLRM achieves much lower inference time for DLRMs compared to both CPU-only and CPU-GPU hybrid counterparts.	Sitian Chen (Hong Kong Baptist University); Haobin Tan (Shenzhen University); Amelie Chi Zhou (Hong Kong Baptist University); Yusen Li (Nankai University); Pavan Balaji (Meta)
1169	ChatCPU: An Agile CPU Design and Verification Platform with LLM	The increasing complexity of semiconductor designs necessitates agile hardware development methodologies to keep pace with rapid technological advancements. Following this trend, Large Language Models (LLMs) emerge as a potential solution, providing new opportunities in hardware design automation. However, existing LLMs exhibit challenges in HDL design and verification, especially for complicated hardware systems. Addressing this need, we introduce ChatCPU, the first end-to-end agile hardware design and verification platform with LLM. ChatCPU streamlines the ASIC design and verification process, guiding it from initial specifications to the final RTL implementations with enhanced design agility. Incorporating the LLM fine-tuning and the processor description language design for CPU design automation, ChatCPU significantly enhances the hardware design capability using LLM. Utilizing ChatCPU, we developed a 6-stage in-order RISC-V CPU prototype, achieving successful tape-out using SkyWater 130nm MPW project with Efabless, which is currently the largest CPU design generated by LLM. Our results demonstrate a remarkable improvement in CPU design efficiency, accelerating the design iteration process by an average of 3.81X, and peaking at 12X and 9.33X in HDL implementations and verification stages, respectively. The ChatCPU also enhances the design capability of LLM by 2.63X as compared to base LLama2. These advancements position ChatCPU as a significant milestone in LLM-driven ASIC design and verification.	Xi Wang (Southeast University & National Center of Technology Innovation for EDA); Gwok-Waa Wan (National Center of Technology Innovation for EDA); Sam-Zaak Wong (National Center of Technology Innovation for EDA); Layton Zhang (National Center of Technology Innovation for EDA); Tianyang Liu (Southeast University); Qi Tian (Southeast University); Jianmin Ye (Southeast University)
1175	Formally Verifying Arithmetic Chisel Designs for All Bit Widths at Once	Efficient verification of ALUs has always been a challenge. Traditionally, they are verified at a low level, leading to state space explosion for larger bit widths. We symbolically can verify ALUs for all bit widths at once. Chisel is a hardware description language embedded in Scala. Our key idea is to transform arithmetic Chisel designs into Scala software programs that simulate their behavior, then apply Stainless, a deductive formal verification tool for Scala. We validate the effectiveness by verifying dividers and multipliers in two open-source RISC-V processors, and conclude that our approach requires less manual guidance than others.	Weizhi Feng (Institute of Software, Chinese Academy of Sciences); Yicheng Liu (Institute of Software, Chinese Academy of Sciences); Jiaxiang Liu (Shenzhen University); David N. Jansen (Institute of Software, Chinese Academy of Sciences); Lijun Zhang (Institute of Software, Chinese Academy of Sciences); Zhilin Wu (Institute of Software, Chinese Academy of Sciences)
1179	GSPO: A Graph Substitution and Parallelization Joint Optimization Framework for DNN Inference	This work proposes GSPO, an automatic unified framework that jointly applies graph substitution and parallelization for DNN inference. GSPO uses joint optimization computation graph (JOCG) to represent both graph substitution and parallelization at the operator level. Then, a novel cost model customized for joint optimization is used to quickly evaluate the computation graph execution time. Combined with backtracking search algorithm, GSPO is able to find the optimal joint optimization solution within acceptable search time. Compared to existing frameworks applying equivalent graph substitution or parallelization, GSPO can achieve up to 27.1% end-to-end performance improvement and reduce search time by up to 94.3%.	Zheng Xu (Tsinghua University); Xu Dai (Shanghai Artificial Intelligence Laboratory); Shaojun Wei (Tsinghua University); Yang Hu (Tsinghua University); Shouyi Yin (Tsinghua University)
1183	Look Before You Access: Efficient Heap Memory Safety for Embedded Systems on ARMv8-M	Numerous embedded systems utilize firmware written in memory-unsafe C/C++. So, the firmware may exhibit spatial memory vulnerabilities, such as buffer overflows, which, if exploited by an attacker, can lead to various software attacks. While several studies have proposed defenses against these memory vulnerabilities, they often introduce significant performance and memory overhead or are impractical for application in embedded systems. In this paper, we introduce micro-fat pointer, a novel solution for heap memory safety for embedded systems. Notably, micro-fat leverages TT instructions newly introduced in ARMv8-M to implement an efficient bounds-checking mechanism. Our evaluation results demonstrate that micro-fat pointer exhibits a 41% performance improvement in compared to the existing state-of-the-art heap memory safety solution.	Jeonghwan Kang (Pusan National University); Jaeyeol Park (Pusan National University); Jiwon Seo (Dankook University); Donghyun Kwon (Pusan National University)
1186	Control Flow Divergence Optimization by Exploiting Tensor Cores	Kernels are scheduled on Graphics Processing Units (GPUs) in the granularity of warp, a bunch of concurrently executing threads. When executing kernels with conditional branches, threads within a warp may execute different branches sequentially, resulting in a considerable utilization loss and unpredictable execution time, known as the control flow divergence. This paper proposes a novel method to predict threads' execution path before the kernel launch by deploying a branch prediction network on the GPU's tensor cores, capable of parallel running with CUDA cores. Combined with a well-designed thread data reorganization algorithm, this solution can mitigate GPUs' control flow divergence problem.	Weiguang Pang (Qilu University of Technology (Shandong Academy of Sciences)); Xu Jiang (University of Electronic Science and Technology of China); Songran Liu (Northeastern University); Lei Qiao (Beijing Institute of Control Engineering); kexue fu (Qilu University of Technology (Shandong Academy of Sciences)); longxiang Gao (Qilu University of Technology (Shandong Academy of Sciences)); Wang Yi (Uppsala University)
1219	CEDAR: Computing-in-pixel Edge-aware Detection and Reconstruction Architecture for High-resolution 3D Imaging	Large-format single-photon avalanche diode (SPAD)-based direct time of flight (dToF) sensors are expected to be widely applied in future L5 full driving automation. However, the high-power in-pixel TDCs and the huge amount of data generated by multi-frame histogram sampling impose limitations on the pixel format of SPAD-based dToF sensors. To tackle this challenge, we proposed the Computing-in-pixel Edge-aware Detection and Reconstruction (CEDAR) architecture. In this architecture, edge pixels are recognized by charge-domain convolution (CDC) computing, and noise pixels are eliminated by in-memory denoising (IMD). Only few TDCs in these edge pixels are activated, resulting in significant power and data savings. Afterward, the full format image is reconstructed by a U-Net using the obtained depth information from these edge pixels. For the first time, we proposed a high-resolution 512 × 512 SPAD-based dToF sensor with a low power of 83.3 mW, a distance accuracy of 0.9 cm, and a frame rate of 60 fps. The high-resolution 3D image can be reconstructed by only 3.5% sparse edge pixels, achieving a PSNR of 35.2 dB. The CEDAR architecture can achieve 16× pixel format and image resolution improvement under the same constraint of power dissipation.	Bu Chen (Fudan University); Zhangcheng Huang (Fudan University); Qi Zheng (Fudan University); Weiyi Tang (Fudan University); Jingyi Wang (Fudan University); Hankun Lv (Fudan University); Chixiao Chen (Fudan University); Jianlu Wang (Fudan University); Qi Liu (Fudan University)
1233	Arbitrary-size Multi-layer OARSMT RL Router Trained with Combinatorial Monte-Carlo Tree Search	This paper presents a novel reinforcement-learning-trained router for building a multi-layer obstacle-avoiding rectilinear Steiner minimum tree (OARSMT). The router is trained by our proposed combinatorial Monte-Carlo tree search to select a proper set of Steiner points for OARSMT with only one inference. By using a Hanna-grid graph as the input and a 3D UNet as the network architecture, the router can handle layouts with any dimensions and any routing costs between grids. The experiments on both random cases and public benchmarks demonstrate that our router can significantly outperform previous algorithmic routers and other RL routers using Alpha-Go-like or PPO-based training.	Liang-Ting Chen (Institute of Electronics, National Yang Ming Chiao Tung University); Hung-Ru Kuo (Institute of Electronics, National Yang Ming Chiao Tung University); Yih-Lang Li (National Yang Ming Chiao Tung University); Mango C.-T. Chao (Institute of Electronics, National Yang Ming Chiao Tung University)
1234	RCGP: An Automatic Synthesis Framework for Reversible Quantum-Flux-Parametron Logic Circuits based on Efficient Cartesian Genetic Programming	Reversible computing has gained increasing attention as a prospective solution for energy dissipation, particularly in quantum computing. As the first practical reversible logic gate using adiabatic superconducting devices, reversible quantum-flux-parametron (RQFP) has been experimentally demonstrated in logical and physical reversibility. However, due to its unique logic function and structure, RQFP logic circuit design poses enormous challenges. Furthermore, circuit scale severely limits the existing exact logic synthesis method for RQFP logic. Therefore, this paper proposes an automatic Cartesian genetic programming-based synthesis framework to generate RQFP logic circuits. Experimental results on reversible logic benchmarks demonstrate RCGP's effectiveness.	Rongliang Fu (The Chinese University of Hong Kong); Robert Wille (Technical University of Munich); Tsung-Yi Ho (The Chinese University of Hong Kong)

Submissio	Title	Abstract	Authors with Affiliations
1249	Genetic Quantization-Aware Approximation for Non-Linear Operations in Transformers	Non-linear functions are prevalent in Transformers and their lightweight variants, incurring substantial and frequently underestimated hardware costs. Previous state-of-the-art works optimize these operations by piece-wise linear approximation and store the parameters in look-up tables (LUT), but most of them require unfriendly high-precision arithmetics such as FP/INT 32 and lack consideration of integer-only INT quantization. This paper proposed a genetic LUT-Approximation algorithm namely GQA-LUT that can automatically determine the parameters with quantization awareness. The results demonstrate that GQA-LUT achieves negligible degradation on the challenging semantic segmentation task for both vanilla and linear Transformer models. Besides, proposed GQA-LUT enables the employment of INT8-based LUT-Approximation that achieves an area savings of 81.3~81.7% and a power reduction of 79.3~80.2% compared to the high-precision FP/INT 32 alternatives.	Pingcheng Dong (The Hong Kong University of Science and Technology); Yonghao Tan (The Hong Kong University of Science and Technology); Dong Zhang (The Hong Kong University of Science and Technology); Tianwei Ni (Zhejiang University); Xuejiao Liu (AI Chip Center for Emerging Smart System (ACCESS)); Yu Liu (AI Chip Center for Emerging Smart System (ACCESS)); Peng Luo (AI Chip Center for Emerging Smart System (ACCESS)); Luhong Liang (AI Chip Center for Emerging Smart System (ACCESS)); Shih-Yang Liu (The Hong Kong University of Science and Technology); Xijie Huang (The Hong Kong University of Science and Technology); Huaiyu Zhu (Zhejiang University); Yun Pan (Zhejiang University); Fengwei An (Southern University of Science and Technology); Kwang-Ting Cheng (The Hong Kong University of Science and Technology)
1260	MoNDE: Mixture of Near-Data Experts for Large-Scale Sparse Models	Mixture-of-Experts (MoE) large language models (LLM) have memory requirements that often exceed the GPU memory capacity, requiring costly parameter movement from secondary memories to the GPU for expert computation. In this work, we present Mixture of Near-Data Experts (MoNDE), a near-data computing solution that efficiently enables MoE LLM inference. MoNDE reduces the volume of MoE parameter movement by transferring only the hot experts to the GPU, while computing the remaining cold experts inside the host memory device. By replacing the transfers of massive expert parameters with the ones of small activations, MoNDE enables far more communication-efficient MoE inference, thereby resulting in substantial speedups over the existing parameter offloading frameworks for both encoder and decoder operations.	Taehyun Kim (Department of Electrical and Computer Engineering, Seoul National University); Kwanseok Choi (Seoul National University); Youngmock Cho (Seoul National University); Jaehoon Cho (Seoul National University); Hyuk-Jae Lee (Seoul National University); Jaewoong Sim (Seoul National University)
1264	Defending against Adversarial Patches using Dimensionality Reduction	reliable use of machine learning models. These attacks involve the strategic modification of localized patches or specific image areas to deceive trained machine learning models. In this paper, we propose DefensiveDR, a practical mechanism using a dimensionality reduction technique to thwart such patch-based attacks. Our method involves projecting the sample images onto a lower-dimensional space while retaining essential information or variability for effective machine learning tasks. We perform this using two techniques, Singular Value Decomposition and t-Distributed Stochastic Neighbour Embedding. We experimentally tune the variability to be preserved for optimal performance as a hyper-parameter. This dimension reduction substantially mitigates adversarial perturbations, thereby enhancing the robustness of the given machine learning model. Our defense is model-agnostic and operates without assumptions about access to model decisions or model architectures, making it effective in both black-box and white-box settings. Furthermore, it maintains accuracy across various models and remains robust against several unseen patch-based attacks. The proposed defensive approach improves the accuracy from 38.8% (without defense) to 66.2% (with defense) when performing LaVAN and GoogleAp attacks, which supersedes that of the prominent state-of-the-art like LGS (53.86%) and Jujutsu (60%).	Nandish Chattopadhyay (New York University); Amira Guesmi (NYU Abu Dhabi); Muhammad Abdullah Hanif (New York University Abu Dhabi); Bassem Ouni (Technology Innovation Institute (TII)); Muhammad Shafique (New York University Abu Dhabi (NYUAD))
1267	Auto-ISP: An Efficient Real-Time Automatic Hyperparameter Optimization Framework for ISP Hardware System	Image Signal Processor (ISP) is widely used in intelligent edge devices across various scenarios. The intricate and time-consuming tuning process demands substantial expertise. Current AI-based auto-tuning operates discretely offline, relying on predefined scenes with human intervention, leading to inconvenient manipulation, with potentially fatal impacts on downstream tasks in unforeseen scenes. We propose a real-time automatic hyperparameter optimization ISP hardware system to address real-world scenarios. Our design features a tri-step framework and a hardware accelerator, demonstrating superior performance in human and computer vision tasks, even in real-time unforeseen scenes. Experiments showcase its practicality, achieving 1080P@75FPS/240FPS in FPGA/ASIC, respectively.	Jiaming Liu (Fudan University); Zihao Liu (Alibaba); Xuan Huang (Fudan University); Ruoxi Zhu (Fudan University); Qi Zheng (Fudan University); Zhijian Hao (Fudan University); Tao Liu (Lawrence Technological University); Jun Tao (Fudan University); Yibo Fan (Fudan University)
1274	ZeroTetris: A Spacial Feature Similarity-based Sparse MLP Engine for Neural Volume Rendering	Neural Volume Rendering (NVR), a novel paradigm for the long-standing problem of photo-realistic rendering of virtual worlds, has developed explosively in the past three years. The unique and substantial computational requirements of NVR pose challenge on deploying NVR to existing dedicated accelerator for neural networks. In this work, we propose ZeroTetris, a spacial feature similarity-based sparse multilayer perceptron (MLP) hardware accelerator for NVR. By leveraging the unique similarity-based sparsity between adjacent sampling points in NVR models, ZeroTetris efficiently bypass the computation of zero activations, thereby enhancing energy efficiency. Evaluation results affirm the effectiveness of the proposed design, showcasing ZeroTetris's superior performance in both area and power efficiency compared to other dedicated sparse matrix multiplication or MLP accelerator designs.	Haochuan Wan (ShanghaiTech University); Linjie Ma (ShanghaiTech University); Antong Li (ShanghaiTech University); Pingqiang Zhou (ShanghaiTech University); Jingyi Yu (ShanghaiTech University); Xin Lou (ShanghaiTech University)
1279	Token-Picker: Accelerating Attention in Text Generation with Minimized Memory Transfer via Probability Estimation	The attention mechanism in text generation is memory-bounded due to its sequential characteristics. Therefore, off-chip memory accesses should be minimized for faster execution. Although previous methods addressed this by pruning unimportant tokens, they fall short in selectively removing tokens with near-zero attention probabilities in each instance. Our method estimates the probability before the softmax function, effectively removing low probability tokens and achieving an 12.1x pruning ratio without fine-tuning. Additionally, we present a hardware design supporting seamless on-demand off-chip access. Our approach shows 2.6x reduced memory accesses, leading to an average 2.3x speedup and a 2.4x energy efficiency.	Junyoung Park (KAIST); Myeonggu Kang (Korea Advanced Institute of Science and Technology); Yunki Han (KAIST); Yang-Gon Kim (KAIST); Jaekang Shin (Korea Advanced Institute of Science and Technology (KAIST)); Lee-Sup Kim (KAIST)

Submissio	Title	Abstract	Authors with Affiliations
1295	C-Nash: A Novel Ferroelectric Computing-in-Memory Architecture for Solving Mixed Strategy Nash Equilibrium	The concept of Nash equilibrium (NE), pivotal within game theory, has garnered widespread attention across numerous industries. However, verifying the existence of NE poses a significant computational challenge, classified as an NP-complete problem. Recent advancements introduced several quantum Nash solvers aimed at identifying pure strategy NE solutions (i.e., binary solutions) by integrating slack terms into the objective function, commonly referred to as slack-quadratic unconstrained binary optimization (S-QUBO). However, incorporation of slack terms into the quadratic optimization results in changes of the objective function, which may cause incorrect solutions. Furthermore, these quantum solvers only identify a limited subset of pure strategy NE solutions, and fail to address mixed strategy NE (i.e., decimal solutions), leaving many solutions undiscovered. In this work, we propose C-Nash, a novel ferroelectric computing-in-memory (CiM) architecture that can efficiently handle both pure and mixed strategy NE solutions. The proposed framework consists of (i) a transformation method that converts quadratic optimization into a MAX-QUBO form without introducing additional slack variables, thereby avoiding objective function changes; (ii) a ferroelectric FET (FeFET) based bi-crossbar structure for storing payoff matrices and accelerating the core vector-matrix-vector (VMV) multiplications of QUBO form; (iii) A winner-takes-all (WTA) tree implementing the MAX form and a two-phase based simulated annealing (SA) logic for searching NE solutions. Evaluations demonstrate that C-Nash has up to 68.6% increase in the success rate for identifying NE solutions, finding all pure and mixed NE solutions rather than only a portion of pure NE solutions, compared to D-Wave based quantum approaches. Moreover, C-Nash boasts a reduction up to 157.9X/79.0X in time-to-solutions in comparison to D-Wave 2000 Q6 and D-Wave Advantage 4.1, respectively.	Yu Qian (Zhejiang University); Kai Ni (University of Notre Dame); Thomas Kämpfe (Fraunhofer IPMS); Cheng Zhuo (Zhejiang University); Xunzhao Yin (Zhejiang University)
1297	ViT-slice: End-to-end Vision Transformer Accelerator with Bit-slice Algorithm	Vision Transformers have demonstrated remarkable performance in various vision tasks. However, general-purpose processors, such as CPUs and GPUs, face challenges in efficiently handling the inference of Vision Transformers. To address the issue, prior works have focused on accelerating only attention due to its high computational cost in NLP Transformers. In contrast, Vision Transformers demonstrate a higher computational cost due to linear modules such as linear transformation, linear projection and Feed-Forward Network (FFN), compared to attention. In this paper, we present ViT-slice, an algorithm-architecture co-design that enhances end-to-end performance and energy efficiency by optimizing not only attention but also linear modules. At the algorithm level, we propose bit-slice compression that avoids storing the redundant most significant bits (MSBs). Additionally, we present bit-slice dot product with early skip to efficiently compute the dot product using bit-sliced data. To enable early skip during the dot product computation, we leverage a trainable threshold. On the hardware level, we introduce a specialized bit-slice dot product unit (BSDPU) to efficiently process the bit-slice dot product with early skip algorithm. Additionally, we present a bit-slice encoder and decoder for on-chip bit-slice compression. ViT-slice achieves 244×, 35.3×, 16.8×, 10.4×, 5.0× end-to-end speedup over Xeon CPU, EdgeGPU, TITAN Xp GPU, Sanger accelerator and ViTCoD accelerator, respectively.	Dongjin Shin (Yonsei University); Insu Choi (Yonsei University); Joon-Sung Yang (Yonsei University)
1307	PowerLens: An Adaptive DVFS Framework for Optimizing Energy Efficiency in Deep Neural Networks	To address the power management challenges in deep neural networks (DNNs), dynamic voltage and frequency scaling (DVFS) technology is garnering attention for its ability to enhance energy efficiency without modifying the structure of DNNs. However, current DVFS methods, which depend on historical information such as processor utilization and task computational load, face issues like frequency ping-pong, response lag, and poor generalizability. Therefore, this paper introduces PowerLens, an adaptive DVFS framework. Initially, we develop a power-sensitive feature extraction method for DNNs and identify critical power blocks through clustering based on power behavior similarity, thereby achieving adaptive DVFS instrumentation point settings. Then, the framework adaptively presets the target frequency for each power block through a decision model. Finally, through a refined training and deployment process, we ensure the framework's effective adaptability across different platforms. Experimental results confirm the effectiveness of the framework in energy efficiency optimization.	Jiawei Geng (The School of Computer Science and Technology and the Suzhou Institute for Advanced Research, University of Science and Technology of China); Zongwei Zhu (School of Software Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China); Weihong Liu (The School of Computer Science and Technology and the Suzhou Institute for Advanced Research, University of Science and Technology of China); Xuehai Zhou (The School of Computer Science and Technology and the Suzhou Institute for Advanced Research, University of Science and Technology of China); Boyu Li (School of Software Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China)
1312	WinoGen: A Highly Configurable Winograd Convolution IP Generator for Efficient CNN Acceleration on FPGA	The convolution neural network (CNN) has been widely adopted in computer vision tasks. In the FPGA-based CNN accelerator design, Winograd convolution can effectively improve computation performance and save hardware resources. However, building efficient and highly compatible IP for arbitrary Winograd convolution on FPGA remains underexplored. To address this issue, we propose a novel and efficient reformulation of Winograd convolution, named Structured Direct Winograd Convolution (SDW). We further develop WinoGen, a Chisel-based highly configurable Winograd convolution IP generator. Given arbitrary input/output tile size and kernel size, it can generate optimized high-performance IP automatically. Meanwhile, our generated IP can be compatible with multiple kernel sizes and tile sizes. Experimental results show that the IP generated by WinoGen achieves DSP efficiency up to 3.80 GOPS/DSP and energy efficiency up to 652.77 GOPS/W while showing 2.45 times and 3.10 times improvements when processing a same CNN model compared with state-of-the-arts.	Mingjun Li (The Chinese University of Hong Kong); Pengjia Li (The Chinese University of Hong Kong, Shenzhen); Shuo Yin (The Chinese University of Hong Kong); Shixin Chen (The Chinese University of Hong Kong); Beichen Li (Chinese University of Hongkong, Shenzhen); Chong Tong (The Chinese University of Hong Kong, Shenzhen); Jianlei Yang (Beihang University); Tinghuan Chen (The Chinese University of Hong Kong, Shenzhen); Bei Yu (The Chinese University of Hong Kong)
1317	FQP: A Fibonacci Quantization Processor with Multiplication-Free Computing and Topological-Order Routing	Neural networks demand increasing computational power and memory access due to growing parameter sizes. A solution is low bit-width quantization, but conventional uniform quantization suffers from distribution mismatches, leading to accuracy loss. We introduce Fibonacci Quantization, closely aligning with neural network data distributions using Fibonacci numbers. Fibonacci Quantization Processor (FQP) features two multiplication-free computing units: the Dualistic-Transformation Adder for large numbers multiplication and the Bit-Exclusive Adder for small numbers multiplication. Additionally, Topological-Order Routing optimizes data mapping onto these units. FQP demonstrates either a 0.98% accuracy improvement or 2.17x higher energy efficiency for ResNet50 on ImageNet1k compared to uniform quantization.	Xiaolong Yang (Tsinghua University); Yang Wang (Tsinghua University); Yubin Qin (Tsinghua University); Jiachen Wang (Tsinghua University); Shaojun Wei (Tsinghua University); Yang Hu (Tsinghua University); Shouyi YIN (Tsinghua University)
1318	BNN-YEO: an efficient Bayesian Neural Network for yield estimation and optimization	Yield estimation and optimization is ubiquitous in modern circuit design but remains elusive for large-scale chips. This is largely due to the mounting cost of transistor-level simulation and one's often limited resources. In this study, we propose a novel framework to estimate and optimize yield using Bayesian Neural Network (BNN-YEO). By coupling machine learning method with Bayesian network, our approach can effectively integrate prior knowledge and is unaffected by the overfitting problem prevalent in most surrogate models. With the introduction of a smooth approximation of the indicator function, it incorporates gradient information to facilitate global yield optimization. We examine its effectiveness via numerical experiments on 6T SRAM and found that BNN-YEO provides 100x speedup (in terms of SPICE simulations) over standard Monte Carlo in yield estimation, and 20x faster than the state-of-the-art method for total yield estimation and optimization with improved accuracy.	Zhenxing Dou (Beihang University); Ming Cheng (University of Bologna); Ming Jia (XC Micro Technologies); Peng Wang (Beihang University)

Submissio	Title	Abstract	Authors with Affiliations
1325	Enabling Multiple Tensor-wise Operator Fusion for Transformer Models on Spatial Accelerators	In transformer models, data reuse within an operator is insufficient, which prompts more aggressive multiple tensor-wise operator fusion (multi-tensor fusion). Due to the complexity in tensor-wise operator dataflow, conventional fusion techniques often fall short by limited dataflow options and short fusion length. In this study, we first identify three challenges on multi-tensor fusion that result in inferior fusions. Then we propose dataflow adaptive tiling (DAT), a novel inter-operator dataflow to enable an efficient fusion of multiple operators connected in any form and chained in any length. Then, we broaden the dataflow exploration from intra-operator to inter-operator and develop an exploration framework to quickly find the best dataflow on spatial accelerators with given on-chip buffer size. Experiment results show that DAT delivers 2.24X and 1.74X speedup and 35.5% and 15.5% energy savings on average for edge and cloud accelerators, respectively, comparing to the state-of-the-art dataflow explorer FLAT. In addition, DAT exploration framework will be open-sourced.	Lei Xu (Shanghai Jiao Tong University); Zhiwen Mo (Shanghai Jiao Tong University); Qin Wang (Shanghai Jiao Tong University); Jianfei Jiang (Shanghai Jiao Tong University); Naifeng Jing (Shanghai Jiao Tong University)
1337	HiMOSS: A Novel High-dimensional Multi-objective Optimization Method via Adaptive Gradient-Based Subspace Sampling for Analog Circuit Sizing	This study presents a novel high-dimensional multi-objective optimization method via adaptive gradient-based subspace sampling for analog circuit sizing. To handle constrained multi-objective optimization, we exploit promising regions from a non-crowded Pareto front, with lightweight Bayesian optimization (BO) based on a novel approximate constrained expected hypervolume improvement. This lightweight BO is computational efficient with constant complexity concerning simulation numbers. To tackle high-dimensional challenges, we reduce the effective dimensionality around promising regions by sampling candidates in an adaptive subspace. The subspace is constructed with gradients and previous success steps with their significance decaying over iterations. The gradients are approximated by sparse regression without additional simulations. The experiments on synthetic benchmarks and analog circuits illustrate advantages of the proposed method over Bayesian and evolutionary baselines.	Tianchen Gu (Fudan University); Ruiyu Lyu (Fudan University); Zhaori Bi (Fudan University); Changhao Yan (Associate Prof. Fudan University); Fan Yang (Fudan University); Dian Zhou (Fudan University); Tao Cui (the Chinese Academy of Sciences); Xin Liu (the Chinese Academy of Sciences); Zaikun Zhang (Hong Kong Polytechnic University); Xuan Zeng (Fudan University)
1344	Net Resource Allocation: A Desirable Initial Routing Step	In modern IC design, routing significantly impacts chip performance, power, area, and design iteration count. Critical challenges in routing include generating rectilinear Steiner minimum tree (RSMT) for each net and handling routing resource among nets. Due to limited resources and net scale, congestion is inevitable in VLSI circuit routing. Most competitive routers address congestion after routing without prior net guidance, leading to difficulty in managing resources among nets. To tackle routing and congestion, we suggest introducing a net resource allocation step as a potentially desirable initial routing stage. Firstly, we introduce the concept of net region probability density (NRPD) to achieve suitable net resource allocation. Using a prior NRPD, we model the resource allocation problem as quadratic programming (QP). We utilize penalty method to solve the QP quickly and obtain a posterior NRPD for each net on each grid. Based on the posterior NRPD and congestion map, we introduce a cost scheme to guide net routing. This cost scheme supports a weighted RSMT construction technique for better topological solutions. Additionally, we propose an iterative method for global routing and track assignment, improving detailed routing quality and optimizing design rule violations. Experimental results show the effectiveness of net resource allocation and demonstrate superior performance of our router over the OpenROAD's router across multiple metrics.	Zhisheng Zeng (SKLP, Institute of Computing Technology, CAS, Peng Cheng Laboratory); Jikang Liu (College of Computer Science and Software Engineering, Shenzhen University); Zhipeng Huang (Peng Cheng Laboratory); Ye Cai (College of Computer Science and Software Engineering, Shenzhen University); Biwei Xie (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences); Yungang Bao (State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences); Xingquan Li (Peng Cheng Laboratory)
1347	VARADE: a Variational-based AutoRegressive model for Anomaly Detection on the Edge	Detecting complex anomalies on massive amounts of data is a crucial task in Industry 4.0, best addressed by deep learning. However, available solutions are computationally demanding, requiring cloud architectures prone to latency and bandwidth issues. This work presents VARADE, a novel solution implementing a light autoregressive framework based on variational inference, which is best suited for real-time execution on the edge. The proposed approach was validated on a robotic arm, part of a pilot production line, and compared with several state-of-the-art algorithms, obtaining the best trade-off between anomaly detection accuracy, power consumption and inference frequency on two different edge platforms.	Alessio Mascolini (Politecnico di Torino); Sebastiano Gaiardelli (University of Verona); Francesco Ponzio (Politecnico di Torino); Nicola Dall'Ora (University of Verona); Enrico Macii (Politecnico di Torino); Sara Vinco (Politecnico di Torino); Santa di Cataldo (Politecnico di Torino); Franco Fummi (University of Verona)
1349	A Deep Reinforcement Learning based Online Scheduling Policy for Deep Neural Network Multi-Tenant Multi-Accelerator Systems	Deep Learning, particularly Deep Neural Networks (DNNs), has emerged as a powerful tool for addressing intricate real-world challenges. Nonetheless, the deployment of DNNs presents its own set of obstacles, chiefly stemming from substantial hardware demands. In response to this challenge, Domain-Specific Accelerators (DSAs) have gained prominence as a means of executing DNNs, especially within cloud service providers offering DNN execution as a service. For service providers, managing multi-tenancy and ensuring high quality service delivery, particularly in meeting stringent execution time constraints, assumes paramount importance, all while endeavoring to maintain cost-effectiveness. In this context, the utilization of heterogeneous multi-accelerator systems becomes increasingly relevant. This paper presents RELMAS, a low-overhead deep reinforcement learning algorithm designed for the real-time scheduling of DNNs in multi-tenant environments, taking into account the dataflow heterogeneity of accelerators and memory bandwidths contentions. By doing so, service providers can employ the most efficient scheduling policy for user requests, optimizing Service-Level-Agreement (SLA) satisfaction rates and enhancing hardware utilization. The application of RELMAS to a heterogeneous multi-accelerator system composed of various instances of Simba and Eyeriss sub-accelerators resulted in up to a 173% improvement in SLA satisfaction rate compared to state-of-the-art scheduling techniques across different workload scenarios, with less than a 1.5% energy overhead.	Francesco Giulio Blanco (University of Catania); Enrico Russo (University of Catania); Maurizio Palesi (University of Catania); Davide Patti (University of Catania); Giuseppe Ascia (University of Catania); Vincenzo Catania (University of Catania)
1351	TraceFormer: S-parameter Prediction Framework for PCB Traces based on Graph Transformer	Signal integrity becomes more critical to modern digital systems such as solid-state drives due to their high-speed operation. However, one of the challenges in signal integrity analysis is S-parameter modeling process for printed circuit boards (PCB). Due to increasing PCB design complexity, existing numerical methods take too long to solve governing equations for S-parameters. To overcome the issue, we present a novel deep learning framework, TraceFormer, to predict S-parameters of PCB traces. Our framework constructs a graph from PCB traces and tokenizes trace segments with geometric and topological information. A transformer encoder produces PCB representations from the tokens, followed by extraction networks which predict four different types of complex-valued S-parameters together. TraceFormer achieved above 0.99 R-squared score up to 15GHz for 4-port PCB designs, resulting in less than 3.1% and 4.2% errors in terms of the eye diagram's width and height, respectively.	Doyun Kim (Samsung Advanced Institute of Technology); Jaemin Park (Samsung Advanced Institute of Technology); Youngmin Oh (Samsung Advanced Institute of Technology); Bosun Hwang (Samsung Advanced Institute of Technology (SAIT), Samsung Electronics)
1360	TAPMM:A Traffic-Aware Page Mapping Method for Multi-level NUMA Systems	With the development of chiplet technology, the architecture of Non-Uniform Memory Access (NUMA) has become increasingly intricate. The placement of memory page significantly influences application performance in NUMA systems. We found that memory access bottlenecks occur between high-level NUMA domains consisting of multiple chiplets. In this paper, we introduce a Traffic-Aware Page Mapping Method (TAPMM) designed for multi-level NUMA systems. TAPMM conceptualizes the multi-level NUMA system as a memory access tree, utilizing hardware performance events to be aware of system traffic and identify the optimal page mapping method for bandwidth efficiency. Our experiments demonstrate that TAPMM achieves a speedup of up to 2.12 times on a real commodity machine compared to existing optimization tools.	Fengkun Dong (Hunan university); Guoqing Xiao (Hunan University); Haotian Wang (Hunan University); Yikun Hu (Hunan University); Kenli Li (Hunan University); Wangdong Yang (Hunan University)

Submissio	Title	Abstract	Authors with Affiliations
1361	GNNavigator: Towards Adaptive Training of Graph Neural Networks via Automatic Guideline Exploration	Graph Neural Networks (GNNs) succeed significantly in many applications recently. However, balancing GNNs training runtime cost, memory consumption, and attainable accuracy for various applications is non-trivial. Previous training methodologies suffer from inferior adaptability and lack a unified training optimization solution. To address the problem, this work proposes GNNavigator, an adaptive GNN training configuration optimization framework. GNNavigator meets diverse GNN application requirements due to our unified software-hardware co-abstraction, proposed GNNs training performance model, and practical design space exploration solution. Experimental results show that GNNavigator can achieve up to 3.1X speedup and 44.9% peak memory reduction with comparable accuracy to state-of-the-art approaches.	Tong Qiao (Beihang University); Jianlei Yang (Beihang University); Yingjie Qi (Beihang University); Ao Zhou (Beihang University); Chen Bai (The Chinese University of Hong Kong); Bei Yu (The Chinese University of Hong Kong); Weisheng Zhao (Beihang University); Chunming Hu (Beihang University)
1363	Unleashing the Power of T1-cells in SFQ Arithmetic Circuits	Superconductive rapid single-flux quantum (RSFQ) ICs dissipate 10-100 smaller power w.r.t. CMOS while operating at tens of GHz. The issue of path balancing in RSFQ systems however incurs significant area overhead, particularly severe due to limited layout density of RSFQ fabrication. The SFQ T1-cell realize the full adder function with 60% less area compared to conventional implementation. This cell however imposes complex input timing constraints. With multiphase clocking, the T1-cell input timing can be efficiently satisfied. Here, we propose SFQ technology mapping methodology supporting T1-cells. The area of the arithmetic SFQ networks is reduced by up to 25%.	Rassul Bairamkulov (EPFL); Mingfei Yu (EPFL); Giovanni De Micheli (École Polytechnique Fédérale de Lausanne (EPFL))
1364	HyCiM: A Hybrid Computing-in-Memory QUBO Solver for General Combinatorial Optimization Problems with Inequality Constraints	Computationally challenging combinatorial optimization problems (COPs) play a fundamental role in various applications. To tackle COPs, many Ising machines and Quadratic Unconstrained Binary Optimization (QUBO) solvers have been proposed, which typically involve direct transformation of COPs into Ising models or equivalent QUBO forms (D-QUBO). However, when addressing COPs with inequality constraints, this D-QUBO approach introduces numerous extra auxiliary variables, resulting in a substantially larger search space, increased hardware costs, and reduced solving efficiency. In this work, we propose HyCiM, a novel hybrid computing-in-memory (CiM) based QUBO solver framework, designed to overcome aforementioned challenges. The proposed framework consists of (i) an innovative transformation method (first to our known) that converts COPs with inequality constraints into an inequality-QUBO form, thus eliminating the need of expensive auxiliary variables and associated calculations; (ii) "inequality filter", a ferroelectric FET (FeFET)-based CiM circuit that accelerates the inequality evaluation, and filters out infeasible input configurations; (iii) a FeFET-based CiM annealer that is capable of approaching global solutions of COPs via iterative QUBO computations within a simulated annealing process. The evaluation results show that HyCiM drastically narrows down the search space, eliminating 2^100 to 2^2536 infeasible input configurations compared to the conventional D-QUBO approach. Consequently, the narrowed search space, reduced to 2^100 feasible input configurations, leads to a substantial hardware area overhead reduction, ranging from 88.06% to 99.96%. Additionally, HyCiM consistently exhibits a high solving efficiency, achieving a remarkable average success rate of 98.54%, whereas D-QUBO implementatoin shows only 10.75%.	Yu Qian (Zhejiang University); Zeyu Yang (Zhejiang University); Kai Ni (University of Notre Dame); Alptekin Vardar (Fraunhofer IPMS); Thomas Kämpfe (Fraunhofer IPMS); Xunzhao Yin (Zhejiang University)
1365	Triplet Network-Based DNA Encoding for Enhanced Similarity Image Retrieval	With the exponential growth of digital data, DNA is emerging as an attractive medium for storage and computing. Thus, design methods for encoding, storing, and searching digital data within DNA storage are of utmost importance. This paper introduces image classification as a measurable task for evaluating the performance of DNA encoders in similar image searches. Furthermore, we propose a novel triplet network-based DNA encoder to improve the accuracy and efficiency. The evaluation using the CIFAR-100 dataset demonstrates that the proposed encoder outperforms existing encoders in retrieving similar images, with an accuracy of 0.77, which is equivalent to 94% of the practical upper limit, and 16 times faster training time.	Takefumi Koike (Kyoto University); Hiromitsu Awano (Kyoto University); Takashi Sato (Kyoto University)
1366	Chiplever: Towards Effortless Extension of Chiplet-based System for FHE	Fully Homomorphic Encryption (FHE) is one of the most promising privacy-preserving techniques, which has drawn increasing attention from both academia and industry due to its ideal security. Chiplet-based designs integrate multiple dies (chiplet) into the package delivering high performance and thereby are embraced by the resources-hungry FHE. Despite the chiplet-based system with various specialized accelerators, it falls short in supporting FHE due to the novel FHE polynomial operations. For a chiplet-based system that is not tailored for FHE, one common approach to make it capable of FHE is designing a new dedicated accelerator, However, this full design-and-build approach overlooks the existing abundant resources of accelerators in the system and thereby incurs repeated customization and resource waste. In this paper, we propose Chiplever, a framework enables effortless extension of Chiplet-based system for FHE. We aim to fully harness the available resources in the room for efficient FHE. To achieve this, Chiplever (1)introduces a specialized extension in I/O Chiplet guided by semantics matching (2)and proposes an efficient allocator featuring specialized dataflow scheduling. (3)Chiplever provides three-step mapping to achieve compiler-level to hardware-level support for FHE and optimizes the data communications.	Yibo Du (Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences); Ying Wang (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences); Bing Li (Capital Normal University); Fuping Li (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, 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); Huawei Li (Institute of Computing Technology, Chinese Academy of Sciences); Xiaowei Li (ICT, Chinese Academy of Sciences); yinhe han (Institute of Computing Technology,Chinese Academy of Sciences)
1369	Toward Controllable Hierarchical Clock Tree Synthesis with Skew-Latency-Load Tree	Clock tree synthesis (CTS) constructs an efficient clock tree, meeting design constraints and minimizing resource usage. It serves as a bridge between placement and routing, facilitating concurrent optimization of multiple design objectives. To construct a clock tree with lower latency and load capacitance while maintaining a specified skew constraint, we introduce skew-latency-load tree (SLLT) which combines the merits of bound skew tree and Steiner shallow-light tree, along with an analysis and demonstration of the boundaries of these two tree types. We propose a method for constructing SLLT, which significantly reduces both the maximum latency and load capacitance compared to previous methods while ensuring skew control. Combining this routing topology generation method, we introduce a hierarchical CTS framework, and it is constructed by integrating partition schemes and buffering optimization techniques. We validate our solution at 28nm process technology, demonstrating superior performance compared to the solutions of OpenROAD and advanced commercial tool. Our approach outperforms in all metrics (max latency, skew, buffer number, clock capacitance), achieving a significant reduction in latency of 29.45% compared to OpenROAD and 6.75% compared to the commercial tool.	Weiguo Li (School of Mathematics and Statistics, Minnan Normal University); Zhipeng Huang (Peng Cheng Laboratory); Bei Yu (The Chinese University of Hong Kong); Wenxing Zhu (Center for Discrete Mathematics and Theoretical Computer Science, Fuzhou University); Xingquan Li (Peng Cheng Laboratory)

Submissio	Title	Abstract	Authors with Affiliations
1375	An In-Memory Computing Accelerator with Reconfigurable Dataflow for Multi-Scale Vision Transformer with Hybrid Topology	Transformer models equipped with multi-head attention (MHA) mechanism have demonstrated promise in computer vision tasks, i.e., vision transformers (ViTs). Nevertheless, the lack of inductive bias in ViTs leads to substantial computational and storage requirements, hindering their deployment on resource-constrained edge devices. To this end, multi-scale hybrid models are proposed to take the advantages of both transformers and CNNs. However, existing domain-specific architectures usually focus on the optimization of either convolution or MHA at the expense of flexibility. In this work, an in-memory computing (IMC) accelerator is proposed to efficiently accelerate ViTs with hybrid MHA and convolution topology by introducing pipeline reordering. SRAM-based digital IMC macro is utilized to mitigate memory access bottleneck, while avoiding analog non-ideality. The reconfigurable processing engines and interconnections are investigated to enable the adaptable mapping of both convolution and MHA. Under typical workloads, experimental results exhibit that our proposed IMC architecture delivers 2.20× to 2.52× speedup and 40.6% to 74.8% energy reduction compared with the baseline design.	Zhiyuan Chen (Peking University); Yufei Ma (Peking University); Keyi Li (Peking University); Yifan Jia (Peking University); Guoxiang Li (Peking University); Meng Wu (Peking University); Tianyu Jia (Peking University); Le Ye (Peking University); Ru Huang (Peking University)
1378	FeBiM: Efficient and Compact Bayesian Inference Engine Empowered with Ferroelectric In-Memory Computing	In scenarios with limited training data or where explainability is crucial, conventional neural network-based machine learning models often face challenges. In contrast, Bayesian inference-based algorithms excel in providing interpretable predictions and reliable uncertainty estimation in these scenarios. While many state-of-the-art in-memory computing (IMC) architectures leverage emerging non-volatile memory (NVM) technologies to offer unparalleled computing capacity and energy efficiency for neural network workloads, their application in Bayesian inference is limited. This is because the core operations in Bayesian inference, i.e., cumulative multiplications of prior and likelihood probabilities, differ significantly from the multiplication-accumulation (MAC) operations common in neural networks, rendering them generally unsuitable for direct implementation in most existing IMC designs. In this paper, we propose FeBiM, an efficient and compact Bayesian inference engine powered by multi-bit ferroelectric field-effect transistor (FeFET)-based IMC. FeBiM effectively encodes the trained probabilities of a Bayesian inference model within a compact FeFET-based crossbar. It maps quantized logarithmic probabilities to discrete FeFET states. As a result, the accumulated outputs of the crossbar naturally represent the posterior probabilities, i.e., the Bayesian inference model's output given a set of observations. This approach enables efficient in-memory Bayesian inference without the need for additional calculation circuitry. As the first FeFET-based in-memory Bayesian inference engine, FeBiM achieves an impressive storage density of 26.32 Mb/mm2 and a computing efficiency of 581.40 TOPS/W in a representative Bayesian classification task. These results demonstrate 10.7x/43.4x improvement in compactness/efficiency compared to the state-of-the-art hardware implementation of Bayesian inference.	Chao Li (Zhejiang University); Zhicheng Xu (The University of Hong Kong); Bo Wen (HKU); Ruibin Mao (The University of Hong Kong); Can Li (The University of Hong Kong); Thomas Kämpfe (Fraunhofer IPMS); Kai Ni (University of Notre Dame); Xunzhao Yin (Zhejiang University)
1381	MENDNet: Just-in-time Fault Detection and Mitigation in AI Systems with Uncertainty Quantification and Multi-Exit Networks	Due to rapid technology scaling in recent years, computation units such as AI systems have become highly susceptible to malfunctions in the hardware. Such malfunctions, when manifested in the accelerator memory, alter the pre-trained Deep Neural Network weight parameters, thereby generating faults, which in turn reduce the inference classification accuracy. To improve the reliability of the AI system, these faults are needed to be detected and mitigated by incorporating just-in-time strategy. Existing approaches for detection/mitigation of faults techniques are not ideal for just-in-time incorporation as the approaches prevents continuous execution or add significant latency overhead. To circumvent this issue, this paper explores uncertainty quantification in deep neural networks as a means of facilitating an efficient and novel fault detection approach in AI systems. Furthermore, in order to mitigate the impact of such faults, we propose MENDNet, which leverages the properties of multi-exit neural networks, coupled with the proposed uncertainty quantification framework. By tuning the confidence threshold for inference in each exit and leveraging the energy-based uncertainty quantification metric, MENDNet can make accurate predictions even in the presence of faults in the computation units. When evaluated on state-of-the-art network-dataset configurations and with multiple fault rate-fault position combinations, our proposed approach furnishes up to 80.42% improvement in inference classification accuracy over a traditional DNN implementation, thereby instilling the reliability of the AI accelerator in mission mode.	Shamik Kundu (University of Texas at Dallas); Mirazul Haque (University of Texas at Dallas); Sanjay Das (University of Texas at Dallas); Wei Yang (University of Texas at Dallas); Kanad Basu (University of Texas at Dallas)
1395	FLAME: Fully Leveraging MoE Sparsity for Transformer on FPGA	MoE (Mixture-of-Experts) mechanism has been widely adopted in transformer-based models to facilitate further expansion of model parameter size and enhance generalization capabilities. However, the practical deployment of MoE mechanism for transformer on resource-constrained platforms, such as FPGA, remains challenging due to heavy memory footprints and impractical runtime costs introduced by the MoE mechanism. Diving into the MoE mechanism, we raise two key observations: (1) Expert weights are heavy but cold, making it ideal to leverage expert weight sparsity. (2) There exists highly skewed expert activation paths for MoE layers in transformer-based models, making it feasible to conduct expert prediction and prefetching. Motivated by these two observations, we propose FLAME, the first algorithm-hardware co-optimized MoE accelerating framework designed to fully leverage MoE sparsity for efficient transformer deployment on FPGA. First, to leverage expert weight sparsity, we integrate an N:M pruning algorithm, allowing for the pruning of expert weights without significantly compromising model accuracy. Second, to settle expert activation sparsity, we propose a circular expert prediction (CEPR) strategy. CEPR prefetches expert weights from external storage to on-chip cache before the activated expert index is determined. Last, we co-optimize both MoE sparsity through the introduction of an efficient pruning-aware expert buffering (PA-BUF) mechanism. Experimental results demonstrate that FLAME achieves 84.4% accuracy of expert prediction with merely two expert caches on-chip. In comparison with CPU and GPU, FLAME achieves 4.12× and 1.49× speedup, respectively.	Xuanda Lin (Fudan University); Huinan Tian (Fudan University); Wenxiao Xue (Fudan University); Lanqi Ma (Fudan University); Jialin Cao (Fudan University); Manting Zhang (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)
1397	CDS: An Anti-Aging Calibratable Digital Sensor for Detecting Multiple Types of Fault Injection Attacks	In this paper, we present CDS, a delay chain based digital sensor that exploits timing variations of both detector and protected object for detecting multiple types of fault injection attacks. To demonstrate its capability, we use CDS to protect the hardware accelerator of PRESENT cryptographic algorithm against multiple glitching attacks. Simulation results show that (1) CDS can detect 100% of voltage and temperature coordinated glitching attacks with 4.1% early warning; (2) CDS can detect 100% of laser glitching attacks with 9.1% early warning; (3) CDS maintains outstanding aging resistance with only 1.1% false alarm rate after 7 years of use.	Zhiyuan Chen (The State Key Laboratory of Blockchain and Data Security, Zhejiang University; School of Cyber Science and Technology, Zhejiang University; ZJU-Hangzhou Global Scientific and Technological Innovation Center); Kun Yang (The State Key Laboratory of Blockchain and Data Security, Zhejiang University; School of Cyber Science and Technology, Zhejiang University; ZJU-Hangzhou Global Scientific and Technological Innovation Center); Kui Ren (The State Key Laboratory of Blockchain and Data Security, Zhejiang University; School of Cyber Science and Technology, Zhejiang University; ZJU-Hangzhou Global Scientific and Technological Innovation Center)

Submissio	Title	Abstract	Authors with Affiliations
1402	Sharry：An Efficient and Sharing Far Memory System	Far Memory System(FMS) allows applications to access memory on remote machines(called memory nodes). However, existing FMSs can`t deal with large loads and have low efficiency in utilizing remote memory, which leads to the inability to share memory nodes among multiple processes, limiting the scalability of FMS. In this paper, we propose Sharry, an efficient Sharing FMS. Sharry manages memory objects from multiple processes within a unified address space, avoiding the overhead of space switching. Sharry also optimizes the utilization of remote memory with fine-grained memory management. Additionally, Sharry offloads memory allocation to dedicated CPU core in order to handle larger loads in the sharing scenario. Compared to state-of-the-art FMS, Sharry improves memory utilisation by 45%, causing only 9% performance degradation when multiple processes sharing single memory node.	Chen Chen (Zhejiang University); Yuhang Huang (Zhejiang University); Shuiguang Deng (Zhejiang University); Jianwei Yin (Zhejiang University); Xinkui Zhao (Zhejiang University)
1403	SARIS: Accelerating Stencil Computations on Energy-Efficient RISC-V Compute Clusters with Indirect Stream Registers	Stencil codes are performance-critical in many compute-intensive applications, but suffer from significant address calculation and irregular memory access overheads. This work presents SARIS, a general and highly flexible methodology for stencil acceleration using register-mapped indirect streams. We demonstrate SARIS for various stencil codes on an eight-core RISC-V compute cluster with indirect stream registers, achieving significant speedups of 2.72x, near-ideal FPU utilizations of 81%, and energy efficiency improvements of 1.58x over an RV32G baseline on average. Scaling out to a 256-core manycore system, we estimate an average FPU utilization of 64%, an average speedup of 2.14x, and up to 15% higher fractions of peak compute than a leading GPU code generator.	Paul Scheffler (Integrated Systems Laboratory, ETH Zurich); Luca Colagrande (ETH Zurich); Luca Benini (Università di Bologna and ETH Zurich)
1424	ML-based Physical Design Parameter Optimization for 3D ICs: From Parameter Selection to Optimization	While various studies have shown effective parameter optimizations for specific designs, there is limited exploration of parameter optimization within the domain of 3D Integrated Circuits. We present the first comprehensive study, both qualitatively and quantitatively, comparing five state-of-the-art (SOTA) techniques for parameter optimization applied to 3D ICs. Additionally, we introduce an end-to-end machine learning-based framework, encompassing important parameter selection through optimization, all without human intervention. Extensive studies across six industrial designs under the TSMC 28nm technology node reveal that our proposed framework outperforms SOTA techniques in three different optimization objectives in both optimization quality and runtime.	Hao-Hsiang Hsiao (Georgia Institute of Technology); Pruek Vanna-iampikul (Georgia Institute of Technology); Yi-Chen Lu (Georgia Institute of Technology); Sung Kyu Lim (Georgia Tech)
1439	LLM-MARK: A Computing Framework on Efficient Watermarking of Large Language Models for Authentic Use of Generative AI at Local Devices	As generative AI such as ChatGPT rapidly evolves, the increasing incidence of data misconduct such as the proliferation of counterfeit news or unauthorized use of Large Language Models (LLMs) presents a significant challenge for consumers to obtain authentic information. While new watermarking schemes are recently being proposed to protect the intellectual property (IP) of LLM, the computation cost is unfortunately too high for the targeted real-time execution on local devices. In this work, a specialized hardware-efficient watermarking computing framework is proposed enabling model authentication at local devices. By employing the proposed hardware hashing for fast lookup and pruned bitonic sorting network acceleration, the developed architecture framework enables fast and efficient watermarking of LLM on the small local devices. The proposed architecture is evaluated on Xilinx XCZU15EG FPGA, demonstrating 30x computing speed-up, making this architecture highly suitable for integration into local mobile devices. The proposed algorithm to architecture codesign framework offers a practical solution to the immediate challenges posed by LLM misuse, providing a feasible hardware solution for Intellectual Property protection in the era of generative AI.	Shiyu Guo (Northwestern University); Yuhao Ju (Northwestern University); Xi Chen (Northwestern University); Jie Gu (Northwestern University)
1446	S2RAM PUF: An Ultra-low Power Subthreshold SRAM PUF with Zero Bit Error Rate	The reliability of physical unclonable function (PUF) has become the biggest challenge for key generation. Existing reliability improvement technologies incur high hardware overhead or testing costs. This paper proposes S2RAM-PUF, a novel, highly reliable and energy-efficient subthreshold SRAM PUF fabricated in 65nm process, with zero bit error rate (BER) across all voltage/temperature corners from 0.5V to 0.8V and from -40℃ to 120℃. The 20480 bits generated by the fabricated 5 S2RAM PUF chips pass the NIST 800-22 randomness test and exhibit almost ideal uniqueness with a mean inter-die hamming distance of 0.5007. The total energy per bit is as low as 3.12fJ at 0.5V supply voltage. Both stabilization BER and energy outperform the two state-of-the-art SRAM-type PUFs reported in JSSC 2020 and 2021.	Li Ni (College of Integrated Circuits, Hunan University); Jiliang Zhang (College of Integrated Circuits, Hunan University)
1448	Towards Redundancy-Free Recommendation Model Training via Reusable-aware Near-Memory Processing	The memory-intensive embedding layer in the recommendation model continues to be the performance bottleneck. While prior works have attempted to improve the embedding layer performance by exploiting the data locality to cache the frequently accessed embedding vectors and their partial sums. However, these solutions rely on the static cache, which is invalidated in the embedding training scenario of the embedding vectors being updated frequently. To this end, this paper proposes ReFree, a redundancy-free near-memory processing (NMP) solution for embedding training. Specifically, ReFree identifies the reusable data in real-time for both forward and backpropagation of the embedding layer training, and leverages a lightweight NMP architecture to enable redundancy-free near-memory acceleration of the entire embedding training process. Evaluation results on real-world datasets show that ReFree outperforms the state-of-the-art solutions by 10.9x and reduces 5.3x energy consumption on average.	Haifeng Liu (Huazhong University of Science and Technology); Long Zheng (Huazhong University of Science and Technology); Yu Huang (Huazhong University of Science and Technology); Haoyan Huang (Huazhong University of Science and Technology); Xiaofei Liao (Huazhong University of Science and Technology); Jin Hai (Huazhong University of Science and Technology)
1455	A HW/SW Co-Design of Video Dehazing Accelerator Using Decoupled Local Atmospheric Light Prior	In this paper, we introduce DLAPID, a novel decoupled parallel hardware-software co-design architecture for real-time video dehazing. From a software point of view, DLAPID isolates the atmospheric light operation from the initial transmission estimation to take full advantage of the hardware accelerators' parallelization features. For the hardware implementation, we deploy DLAPID both on FPGA and GPU platforms and validate its effectiveness. Using both real-world driving scenario testing sets and ground-truth datasets, we quantitatively and qualitatively assess the proposed method against several SOTA (state-of-the-art) video dehazing models. The outcomes of our experiments demonstrate that our approach achieves better dehazing performance with lower power consumption and has real-time processing capabilities, thereby preventing potential accidents in autonomous vehicles.	Yanjie Tan (Hunan University); Yifu Zhu (Hunan University); Zhaoyang Huang (Hunan University); Feiteng Nie (Hunan University); Huailiang Tan (Hunan University)
1457	A Combined Content Addressable Memory and In-Memory Processing Approach for k-Clique Counting Acceleration	k-Clique counting problem plays an important role in graph mining which has seen a growing number of applications. However, current k-Clique counting accelerators cannot meet the performance requirement mainly because they struggle with high data transfer issue incurred by the intensive set intersection operations and the inability of load balancing. In this paper, we propose to solve this problem with a hybrid framework of content addressable memory (CAM) and in-memory processing (PIM). Specifically, we first utilize CAM for binary induced subgraph generation in order to reduce the search space, then we use PIM to implement in-place parallel k-Clique counting through iterative Boolean logic "AND"- like operation. To take full advantage of this combined CAM and PIM framework, we develop dynamic task scheduling strategies that can achieve near optimal load balancing among the PIM arrays. Experimental results demonstrate that, compared with state-of-the-art CPU and GPU platforms, our approach achieves speedups of 167.5× and 28.8×, respectively. Meanwhile, the energy efficiency is improved by 788.3× over the GPU baseline.	Xidi Ma (Beihang University); Weichen Zhang (Beihang University); Xueyan Wang (Beihang University); Tianyang Yu (Nanjing University of Aeronautics and Astronautics); Bi Wu (Nanjing University of Aeronautics and Astronautics); Gang Qu (Univ. of Maryland, College Park); Weisheng Zhao (Beihang University)

Submissio	Title	Abstract	Authors with Affiliations
1459	Symbolic Quick Error Detection by Semantically Equivalent Program Execution	Symbolic quick error detection (SQED) has greatly improved efficiency in formal chip verification. However, it has a limitation in detecting single-instruction bugs due to its reliance on the self-consistency property. To address this, we propose a new variant called symbolic quick error detection by semantically equivalent program execution (SEPE-SQED), which utilizes program synthesis techniques to find sequences with equivalent meanings to original instructions. SEPE-SQED effectively detects single-instruction bugs by differentiating their impact on the original instruction and its semantically equivalent program (instruction sequence). To manage the search space associated with program synthesis, we introduce the CEGIS based on the highest priority first algorithm. The experimental results show that our proposed CEGIS approach improves the speed of generating the desired set of equivalent programs by 50% in time compared to previous methods. Compared to SQED, SEPE-SQED offers a wider variety of instruction combinations and can provide a shorter trace for triggering bugs in certain scenarios.	Yufeng Li (Institute of Software, Chinese Academy of Sciences; University of Chinese Academy of Sciences); Qiusong Yang (Institute of Software, Chinese Academy of Sciences); Yiwei Ci (Institute of Software, Chinese Academy of Sciences); Enyuan Tian (Institute of Software, Chinese Academy of Sciences; University of Chinese Academy of Sciences)
1460	OPAL: Outlier-Preserved Microscaling Quantization Accelerator for Generative Large Language Models	To overcome the burden on the memory size and bandwidth due to ever-increasing size of large language models (LLMs), aggressive weight quantization has been recently studied, while lacking research on quantizing activations. In this paper, we present a hardware-software co-design method that results in an energy-efficient LLM accelerator, named OPAL, for generation tasks. First of all, a novel activation quantization method that leverages the microscaling data format while preserving several outliers per sub-tensor block (e.g., four out of 128 elements) is proposed. Second, on top of preserving outliers, mixed precision is utilized that sets 5-bit for inputs to sensitive layers in the decoder block of an LLM, while keeping inputs to less sensitive layers to 3-bit. Finally, we present the OPAL hardware architecture that consists of FP units for handling outliers and vectorized INT multipliers for dominant non-outlier related operations. In addition, OPAL uses log2-based approximation on softmax operations that only requires shift and subtraction to maximize power efficiency. As a result, we are able to improve the energy efficiency by 1.6~2.2x, and reduce the area by 2.4~3.1x with negligible accuracy loss, i.e., <1 perplexity increase.	Dahoon Park (Korea University); Jahyun Koo (DGIST); Sangwoo Jung (DGIST); Jaeha Kung (Korea University)
1461	Energy-efficient SNN Architecture using 3nm FinFET Multiport SRAM-based CIM with Online Learning	There is an increasing demand for ultra-low power in Edge AI devices, such as smartphones, wearables, and Internet-of-Things sensor systems, with constrained battery budgets. Current AI computation units face challenges, primarily from the memory-wall issue, limiting overall system-level performance. In this paper, we propose a new SRAM-based Compute-In-Memory (CIM) accelerator optimized for Spiking Neural Networks (SNNs) inference. Our proposed architecture employs a multiport SRAM design with multiple decoupled read ports to enhance the throughput and transposable read-write ports to facilitate online learning. Furthermore, we develop an Arbiter circuit for efficient data processing and port allocations during the computation. Results for a 128x128 array in 3nm FinFET technology demonstrate a 3.1x improvement in speed and a 2.2x enhancement in energy efficiency with our 5R1W SRAM design compared to the traditional single-port SRAM design. At the system level, a throughput of 44 MInf/s at 607 pJ/Inf and 29 mW is achieved.	Lucas Huijbregts (TU-Delft, IMEC); Hsiao-Hsuan Liu (IMEC); Paul Detterer (IMEC); Said Hamdioui (Delft University of Technology); Amirreza Yousefzadeh (University of Twente); Rajendra Bishnoi (Delft University of Technology,)
1464	MTL-Split: Multi-Task Learning for Edge Devices using Split Computing	Split Computing (SC), where a Deep Neural Network (DNN) is intelligently split with a part of it deployed on an edge device and the rest on a remote server is emerging as a promising approach. It allows the power of DNNs to be leveraged for latency-sensitive applications that do not allow the entire DNN to be deployed remotely, while not having sufficient computation bandwidth available locally. In many such embedded scenarios, such as those in the automotive domain, computational resource constraints also necessitate MultiTask Learning (MTL), where the same DNN is used for multiple inference tasks instead of having dedicated DNNs for each task, which would need more computing bandwidth. However, how to partition such a multi-tasking DNN to be deployed within a SC framework has not been sufficiently studied. This paper studies this problem and MTL-Split, our novel proposed architecture, shows encouraging results on both synthetic and real-world data. The code implementing this architecture will be made publicly available.	Luigi Capogrosso (University of Verona); Enrico Fraccaroli (University of Verona); Samarjit Chakraborty (UNC Chapel Hill); Franco Fummi (University of Verona); Marco Cristani (University of Verona)
1468	Cross-Layer Exploration and Chip Demonstration of In-Sensor Computing for Large-Area Applications with Differential-Frame ROM-Based Compute-In-Memory	In-sensor computing has emerged as a promising approach to mitigating huge data transmission costs between sensors and processing units. Recently, the emerging application scenarios have raised more demands of sensory technology for large-area and flexible integration. However, with thin-film technologies that are capable of providing flexible and large-area integration support, the implementation of in-sensor computing can be strongly restricted due to the low device performance, large-area integration variation, and costly interface between sensors and CMOS processors. To address this challenge, we propose an in-sensor computing architecture to facilitate high-parallelism NN pre-processing and effective data compression. The boundaries of computing parallelism are expanded by adopting compact ROM-based compute-in-memory scheme next to sensing array. Differential-frame computing provides not only excellent robustness, but also high data sparsity. A bio-inspired data compression method with residual recovery caches and zero-skip circuits further enhances output sparsity without accumulated error. Based on the proposed cross-layer design optimization, an LTPS TFT-based ROM CiM chip has been fabricated and experimentally measured. The system-level evaluation demonstrates 3.85× speedup and 5.10× energy efficiency improvement compared with traditional architecture with separated sensors and processors, outperforming existing in-sensor computing works in large-area thin-film technology scenarios.	Jialong Liu (Tsinghua University); Wenjun Tang (Tsinghua University); Deyun Chen (Tsinghua University); Chen Jiang (Tsinghua University); Huazhong Yang (Tsinghua University); Xueqing Li (Tsinghua University)
1469	CSTrans-OPU: An FPGA-based Overlay Processor with Full Compilation for Transformer Networks via Sparsity Exploration	In this work, we propose CSTrans-OPU, an FPGA-based overlay processor with full compilation for transformer networks via sparsity exploration. Specifically, we customize a multi-precision processing element (PE) array with DSP-packing for unified computation format with full resource utilization. Additionally, the introduced sorting and computation mode selection modules make it possible to explore the token sparsity. Moreover, equipped with a user-friendly compiler, CSTrans-OPU enables model parsing, operation fusion, model quantization, instruction generation and reordering directly from model files. To the best of our knowledge, our CSTrans-OPU is the first overlay processor for transformer networks considering sparsity.	Yueyin Bai (Fudan University); Keqing Zhao (Fudan University); Yang Liu (Fudan University); Hongji Wang (Fudan University); Hao Zhou (Fudan University); Xiaoxing Wu (Fudan University); Jun Yu (Fudan University); Kun Wang (Fudan University)

Submissio	Title	Abstract	Authors with Affiliations
1473	MAUnet: Multiscale Attention U-Net for Effective IR Drop Prediction	The efficient analysis of power grids is a crucial yet computationally challenging task in integrated circuit (IC) design, given the shrinking power supply voltage of ultra deep-submicron VLSI design. Different from conventional modified nodal analysis analytical solving technique, this paper introduces MAUnet, an innovative machine-learning model that redefines state-of-the-art full-chip static IR drop prediction. MAUnet ingeniously integrates multi-scale convolutional blocks, attention mechanisms, and U-Net architecture to optimize prediction accuracy. The multi-scale convolutional blocks significantly enhance feature extraction from image-based data, while the attention mechanism precisely identifies hotspot regions. The U-Net architecture, on the other hand, enables scalable image-to-image prediction applicable to circuits of any size. Uniquely, MAUnet also incorporates a pioneering fusion method that synergies both power grids and image-based data. Additionally, we introduce a low-rank approximation transfer learning technique to extend MAUnet's applicability to unseen test cases. Benchmark tests validate MAUnet's superior performance, achieving an average error of less than 6% relative to the average IR drop on three benchmarks.The performance enhancements offered by our proposed method are substantial, outperforming the current state-of-the-art method, IREDGe, by considerable margins of 29%, 65%, and 68% in three canonical benchmarks. Transfer learning is validated to enable model to achieve effective improvement on real circuit test cases. Compared to commercial tools, which often require hours to deliver results, the proposed method provides orders of magnitude speed-up with negligible error in practice.	Mingyue Wang (Beihang University); Yuanqing Cheng (BeiHang University); Yage Lin (BeiHang University); Kelin Peng (BeiHang University); Shunchuan Yang (BeiHang University); Zhou Jin (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Wei W. Xing (The University of Sheffield)
1475	ScaleFold: Reducing AlphaFold Initial Training Time to 10 Hours	AlphaFold2 has been hailed as a breakthrough in protein folding. It can rapidly predict protein structures with lab-grade accuracy. However, its training procedure is prohibitively time-consuming, and gets diminishing benefits from scaling to more compute resources. In this work, we conducted a comprehensive analysis on the AlphaFold training procedure, identified that inefficient communications and overhead-dominated computations were the key factors that prevented the AlphaFold training from effective scaling. We introduced ScaleFold, a systematic training method that incorporated optimizations specifically for these factors. ScaleFold successfully scaled the AlphaFold training to 2080 NVIDIA H100 GPUs with high resource utilization. In the MLPerf HPC v3.0 benchmark, ScaleFold finished the OpenFold benchmark in 7.51 minutes, shown over 6X speedup than the baseline. For training the AlphaFold model from scratch, ScaleFold completed the pretraining in 10 hours, a significant improvement over the seven days required by the original AlphaFold pretraining baseline.	Feiwen Zhu (NVIDIA); Arkadiusz Nowaczynski (NVIDIA); Rundong Li (NVIDIA); Jie Xin (NVIDIA); Yifei Song (NVIDIA); Michal Marcinkiewicz (NVIDIA); Sukru Burc Eryilmaz (NVIDIA); June Yang (NVIDIA); Michael Andersch (NVIDIA)
1478	HEIRS: Hybrid Three-Dimension RRAM- and SRAM-CIM Architecture for Multi-task Transformer Acceleration	Large-scale transformer with millions of weights achieves great success in multiple natural language processing (NLP) tasks. To release the memory bottleneck of multi-task model deployment, transfer learning tunes part of weights with shared parameters among tasks. Moreover, computing-in-memory (CIM) emerges as an efficient solution for neural network (NN) acceleration. With higher storage density, RRAM-CIM can store the large-scale model without costly weight loading, compared with another mainstream SRAM-CIM. However, the RRAM rewrite for tunning and dynamic weight matrix-vector-multiplication (MVM) in transformers requires high-cost RRAM writing in RRAM-CIM. Current hybrid CIM can compensate the weakness of RRAM-CIM by adding SRAM-CIM with independent MVM operation. However, the tunned weights in transfer learning cannot be implemented due to the demand for the cooperative addition of MVM results from shared weights and tunned weights. In this paper, a hybrid three-dimension RRAM-CIM and SRAM-CIM architecture (HEIRS) is proposed for multi-task transformer acceleration, with the monolithically 3D integration of high-density RRAM-CIM and high-performance SRAM-CIM. The 3D RRAM-CIM with ultra-high density stores the whole NN model with mitigated off-chip weight loading. The SRAM-CIM is employed for efficiently performing dynamic weight MVM without RRAM write operation. Moreover, a novel hybrid-CIM paradigm is proposed with an input selective adder tree, to support cooperative addition in transfer learning. The experiment shows that, compared with RRAM-CIM and SRAM-CIM, the proposed HEIRS improves the energy efficiency by up to 7.83x and 2.29x on BERT, respectively. Meanwhile, the latency is also reduced by up to 85.5% and the storage density is enhanced by 7.2x, compared to RRAM-CIM.	Liukai Xu (Shanghai Jiao Tong University); Shuai Yuan (Department of Micro-Nano Electronics, Shanghai Jiao Tong University); dengfeng wang (student); Yiming Chen (Tsinghua University); Xueqing Li (Tsinghua University); Yanan Sun (Department of Micro-Nano Electronics, Shanghai Jiao Tong University)
1498	Hybrid Circuit Mapping: Leveraging the Full Spectrum of Computational Capabilities of Neutral Atom Quantum Computers	Quantum computing based on Neutral Atoms (NAs) provides a wide range of computational capabilities, encompassing high-fidelity long-range interactions with native multi-qubit gates, and the ability to shuttle arrays of qubits. While previously these capabilities have been studied individually, we propose the first approach of a fast hybrid compiler to perform circuit mapping and routing based on both high-fidelity gate interactions and qubit shuttling. We delve into the intricacies of the compilation process when combining multiple capabilities and present effective solutions to address resulting challenges. The final compilation strategy is then showcased across various hardware settings, revealing its versatility, and highlighting potential fidelity enhancements achieved through the strategic utilization of combined gate- and shuttling-based routing. With the additional multi-qubit gate support for both routing capabilities, the proposed approach is able to take advantage of the full spectrum of computational capabilities offered by NAs.	Ludwig Schmid (Technical University of Munich); Sunghye Park (Pohang University of Science and Technology); Robert Wille (Technical University of Munich)
1518	Beyond Inference: Performance Analysis of DNN Server Overheads for Computer Vision	Deep neural network (DNN) inference has become an important part of many data-center workloads. This has prompted focused efforts to design ever-faster deep learning accelerators such as GPUs and TPUs. However, an end-to-end vision application contains more than just DNN inference, including input decompression, resizing, sampling, normalization, and data transfer. In this paper, we perform a thorough evaluation of computer vision inference requests performed on a throughput-optimized serving system. We quantify the performance impact of server overheads such as data movement, preprocessing, and message brokers between two DNNs producing outputs at different rates. Our empirical analysis encompasses many computer vision tasks including image classification, segmentation, detection, depth-estimation, and more complex processing pipelines with multiple DNNs. Our results consistently demonstrate that end-to-end application performance can easily be dominated by data processing and data movement functions (up to 56% of end-to-end latency in a medium-sized image, and ∼ 80% impact on system throughput in a large image), even though these functions have been conventionally overlooked in deep learning system design. Our work identifies important performance bottlenecks in different application scenarios, achieves 2.25× better throughput compared to prior work, and paves the way for more holistic deep learning system design.	Ahmed F. AbouElhamayed (Cornell University); Susanne Balle (Intel Corporation); Deshanand Singh (Intel); Mohamed S. Abdelfattah (Cornell University)

Submissio	Title	Abstract	Authors with Affiliations
1521	ReCG: ReRAM-Accelerated Sparse Conjugate Gradient	Solving sparse linear systems is crucial in scientific computing. Sparse Conjugate Gradient (CG) is one of the most popular iterative solvers with high efficiency and low storage requirements. However, the performance of sparse CG solvers implemented on storage-compute separated architectures is greatly limited by the irregular memory access and the large amount of data transmission. In this paper, we propose a processing-in-memory (PIM) architecture, ReCG, based on the resistive random-access memory (ReRAM) to accelerate sparse CG solvers. The design of ReCG faces three major challenges: (1) how to make complex CG more suitable for acceleration with ReRAM-based architecture, (2) how to map sparse and irregular operations to regular crossbars that are more suitable for dense operations, and (3) how to coordinate the dataflow among hardware units to minimize the impact of the poor write endurance of ReRAMs on CG acceleration. To address these challenges, we (1) classify the sparse CG kernels by exploring the commonality of operations and design a flexible and dedicated architecture, (2) efficiently implement the sparse and irregular operations by utilizing both content-addressable memory (CAM) and multiply-and-accumulate (MAC) crossbars, and (3) develop a novel scheduling strategy for the dataflow. The experimental results show that ReCG improves the performance by up to three, one and one orders of magnitude compared with PETSc on CPU and GPU and CALLIPEPLA on FPGA, respectively, and the energy consumption is reduced by up to two, two and one orders of magnitude.	Mingjia Fan (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Xiaoming Chen (Institute of Computing Technology, Chinese Academy of Sciences); Dechuang Yang (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Zhou Jin (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Weifeng Liu (Super Scientific Software Laboratory, China University of Petroleum-Beijing)
1522	HTAG-eNN: Hardening Technique with AND Gates for Embedded Neural Networks	Embedded Neural Networks (NNs) face significant challenges due to Single-Event Upsets (SEUs), compromising their reliability. To address this challenge, previous works study SEU layers sensitivity of AI models. Contrary to these techniques, remaining at high level, we propose a more accurate analysis, highlighting that, except for the last layer, faults transitioning from 0 to 1 significantly impact classification outcomes. Based on this specific behavior, we propose a simple hardware block able to detect and mitigate the SEU impact. Obtained results show that HTAG protection efficiency is near 96.85% for the LeNet-5 CNN inference model, suitable for an embedded system. This result can be improved with other protection methods for the classification layer. Additionally, it significantly reduces area overhead and critical path compared to existing approaches.	Wilfread GuillemÃ© (INRIA/IRISA); Angeliki Kritikakou (Univ Rennes, Inria, CNRS, IRISA); Youri Helen (DGA MI); Cedric Killian (Universite Jean Monnet); Daniel Chillet (University of Rennes 1)
1526	TrafficHD: Efficient Hyperdimensional Computing for Real-Time Network Traffic Analytics	With the evolution of network infrastructure, the pattern of network traffic becomes unprecedentedly complex. Conventional machine learning algorithms are struggling to cope with the high-dimensional data and real-time processing speeds required in such complex networks. Fortunately, hyperdimensional Computing (HDC), which is power-efficient and supports parallel processing, provides a potential solution to this challenge. In this paper, we present TrafficHD, a novel classification framework that leverages HDC to analyze network traffic in real-time. By transforming network traffic features into high-dimensional binary vectors, TrafficHD enables the rapid execution of recognition tasks within the constraints of real-time systems. Extensive evaluations on a wide range of network tasks show that TrafficHD achieves 30.57× and 98.32× faster than state-of-the-art (SOTA) machine learning and HDC algorithms while providing 3× higher robustness to network noise.	Haodong Lu (Fudan University); Zhiyuan Ma (Fudan University); Xinran Li (Fudan University); Shiyan Bi (Fudan University); Xiaoming He (Fudan University); Kun Wang (Fudan University)
1528	QUQ: Quadruplet Uniform Quantization for Efficient Vision Transformer Inference	While exhibiting superior performance in many tasks, vision transformers (ViTs) face challenges in quantization. Some existing low-bit-width quantization techniques cannot effectively cover the whole inference process of ViTs, leading to an additional memory overhead (22.3%-172.6%) compared with the corresponding fully quantized models. To address this issue, we propose quadruplet uniform quantization (QUQ) to deal with data of various distributions in ViT. QUQ divides the entire data range into at most four subranges that are uniformly quantized with different scale factors, respectively. To determine the partition scheme and quantization parameters, an efficient relaxation algorithm is proposed accordingly. Moreover, dedicated encoding and decoding strategies are devised to facilitate the design of an efficient accelerator. Experimental results show that QUQ surpasses state-of-the-art quantization techniques; it is the first viable scheme that can fully quantize ViTs to 6-bit with acceptable accuracy. Compared with the conventional uniform quantization, QUQ results in not only a higher accuracy but also an accelerator with lower area and power.	Xinkuang Geng (Shanghai Jiao Tong University); Siting Liu (ShanghaiTech University); leibo liu (Institute of Microelectronics and The National Lab for Information Science and Technology, Tsinghua University); Jie Han (University of Alberta); Honglan Jiang (Shanghai Jiao Tong University)
1530	PPGNN: Fast and Accurate Privacy-Preserving Graph Neural Network Inference via Parallel and Pipelined Arithmetic-and-Logic FHE Accelerator	Graph Neural Networks (GNNs) are increasingly used in fields like social media and bioinformatics, promoting the prosperity of cloud-based GNN inference services. Nevertheless, data privacy becomes a critical issue when handling sensitive information. Fully Homomorphic Encryption (FHE) enables computations on encrypted data, while privacy-preserving GNN inference generally necessitates ensuring graph structure data confidentiality and maintaining computation precision, both of which are computationally expensive in FHE. Existing schemes of GNNs inference with FHE are deterred by either computational overhead, accuracy degradation, or incomplete data protection. This paper presents PPGNN to address these challenges all at once. We first propose a novel privacy-preserving GNN inference algorithm utilizing a high-accuracy arithmetic-and-logic FHE approach, meanwhile only need much smaller parameters, substantially reducing computational complexity and facilitating parallel processing. Correspondingly, a dedicated hardware architecture has been designed to implement these innovations, with featured specialized units for arithmetic and logic FHE operations in a pipelined manner. Collectively, PPGNN achieves 2.7× and 1.5× speedup over state-of-the-art Arithmetic FHE and Logic FHE accelerators while ensuring high accuracy, simultaneously with about 18× energy reduction on average.	Yuntao Wei (Beihang University); Xueyan Wang (Beihang University); Song Bian (Beihang University); Yicheng Huang (Beihang University); Weisheng Zhao (Beihang University); Yier Jin (University of Science and Technology of China)
1533	VITA: ViT Acceleration for Efficient 3D Human Mesh Recovery via Hardware-Algorithm Co-Design	Vision Transformers (ViTs) have emerged as a promising solution to enable efficient 3D Human Mesh Recovery (HMR) in augmented and virtual reality (AR/VR) applications. However, it remains a challenge to efficiently accelerate ViT-based HMR due to high computational complexity and memory access footprint. In this paper, we propose VITA, a hardware and algorithm co-design framework for ViT-based HMR with much-improved performance and energy efficiency. To be specific, on the algorithm side, we proposed a pooling attention model optimized with regular memory access and reduced computation complexity. On the hardware side, we proposed an accelerator architecture capable of adapting various data movement caused by various pooling operations.	Shilin Tian (University of Central Florida); Chase Szafranski (University of Central Florida); Ce Zheng (University of Central Florida); Fan Yao (University of Central Florida); Ahmed Louri (The George Washington University); Chen Chen (University Of Central Florida); Hao Zheng (University of Central Florida)
1535	HiRISE: High-Resolution Image Scaling for Edge ML via In-Sensor Compression and Selective ROI	With the rise of tiny IoT devices powered by machine learning (ML), many researchers have directed their focus toward compressing models to fit on tiny edge devices. Recent works have achieved remarkable success in compressing ML models for object detection and image classification on microcontrollers with small memory, e.g., 512kB SRAM. However, there remain many challenges prohibiting the deployment of ML systems that require high-resolution images. Due to fundamental limits in memory capacity for tiny IoT devices, it may be physically impossible to store large images without external hardware. To this end, we propose a high-resolution image scaling system for edge ML, called HiRISE, which is equipped with selective region-of-interest (ROI) capability leveraging analog in-sensor image scaling. Our methodology not only significantly reduces the peak memory requirements, but also achieves up to 17.7x reduction in data transfer and energy consumption.	Brendan Reidy (University of South Carolina); Sepehr Tabrizchi (University of Nebraska–Lincoln); MohammadReza Mohammadi (University of South Carolina); Shaahin Angizi (New Jersey Institute of Technology); Arman Roohi (University of Nebraska - Lincoln); Ramtin Zand (University of South Carolina)

Submissio	Title	Abstract	Authors with Affiliations
1537	inGRASS: Incremental Graph Spectral Sparsification via Low-Resistance-Diameter Decomposition	This work presents inGRASS, a novel algorithm designed for incremental spectral sparsification of large undirected graphs. The proposed inGRASS algorithm is highly scalable and parallel-friendly, having a nearly linear time complexity for the setup phase and the ability to update the spectral sparsifier in $O(\log N)$ time for each incremental change made to the original graph with N nodes. A key component in the setup phase of inGRASS is a multilevel resistance embedding step for efficiently identifying spectrally critical edges and effectively pruning spectrally similar ones, which is achieved by decomposing the initial sparsifier into node clusters with bounded effective-resistance diameters achieved through a low-resistance-diameter decomposition (LRD) scheme. The update phase of inGRASS exploits low-dimensional node embedding vectors for efficiently estimating the importance and uniqueness of each newly added edge. As demonstrated through extensive experiments, inGRASS achieves state-of-the-art results in incremental spectral sparsification of graphs obtained from various tasks, such as circuit simulations, finite element analysis, and social networks.	Ali Aghdaei (Stevens Institute of Technology); Zhuo Feng (Stevens Institute of Technology)
1549	SGM-PINN: Sampling Graphical Models for Faster Training of Physics-Informed Neural Networks	SGM-PINN is a graph-based importance sampling framework to improve the training efficacy of Physics-Informed Neural Networks (PINNs) on parameterized problems. By applying a graph decomposition scheme to an undirected Probabilistic Graphical Model (PGM) built from the training dataset, our method generates node clusters encoding conditional dependence between training samples. Biasing sampling towards more important clusters allows smaller mini-batches and training datasets, improving training speed and accuracy. We additionally fuse an efficient robustness metric with residual losses to determine regions requiring additional sampling. Experiments demonstrate the advantages of the proposed framework, achieving 3X faster convergence compared to prior state-of-the-art sampling methods.	John M. Anticev (Stevens Institute of Technology); Ali Aghdaei (Stevens Institute of Technology); Wuxinlin Cheng (Stevens Institute of Technology); Zhuo Feng (Stevens Institute of Technology)
1551	SAS - A Framework for Symmetry-based Approximate Synthesis	Approximate Computing is a design paradigm that trades off computational accuracy for gains in non-functional aspects such as reduced area, increased computation speed, or power reduction. The latter is of special interest in the field of Internet of Things. In this paper we present SAS, a framework for symmetry-based approximate logic synthesis. Given a Boolean multi-output function, SAS approximates it by (partially) replacing its output functions by symmetric functions with minimal Hamming distance. The framework is capable of restricting the introduced error with respect to a parameterized error metric that covers many real-word use-cases. Experimental results on common benchmark sets as well as large bit width arithmetic Boolean functions confirm the effectiveness of the proposed framework. SAS is capable of synthesizing Boolean functions with size reductions of up to approximately 45% while, at the same time, respecting the specified threshold on the error metric. The framework is publicly available as open-source software on GitHub.	Niklas Jungnitz (Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)); Oliver Keszocze (Technical University of Denmark)
1555	NOFIS: Normalizing Flow for Rare Circuit Failure Analysis	Accurate estimation of rare failure occurrence probability is crucial for ensuring the proper and reliable functioning of integrated circuits (ICs). Conventional Monte Carlo methods are inefficient, demanding an exorbitant number of samples to achieve reliable estimates. Inspired by the exact sampling capabilities of normalizing flows, we revisit this problem and propose normalizing flow assisted importance sampling, termed NOFIS. NOFIS first learns a sequence of proposal distributions associated with predefined nested subset events by minimizing KL divergence losses. Next, it estimates the rare event probability by utilizing importance sampling in conjunction with the last proposal. The efficacy of our NOFIS method is substantiated through comprehensive qualitative visualizations, affirming the optimality of the learned proposal distribution, as well as 10 quantitative experiments (covering electronic Opamp and Charge Pump circuits, and photonic Y-branch), which highlight NOFIS's superior accuracy over baseline approaches.	Zhengqi Gao (Dept. of EECS, MIT); Dinghuai Zhang (Mila – Quebec AI Institute, Universite´ de Montre´al); Luca Daniel (MIT); Duane Boning (MIT)
1556	Unleashing the Potential of AQFP Logic Placement via Entanglement Entropy and Projection	Adiabatic quantum-flux-parametron (AQFP) logic, known for its remarkable energy efficiency, has emerged as a prominent superconductor-based logic family, surpassing traditional rapid single flux quantum (RSFQ) logic. In AQFP circuits, each cell operates on AC power, serving as both a power supply and clock signal to drive data flow across clock phases. However, signal attenuation with increasing wire length may result in more potential data errors. To address this, rows of buffers are inserted as repeaters to ensure data synchronization and avoid wirelength violations. However, these inserted buffer rows in AQFP placement significantly amplifies power consumption and circuit delay. To resolve these challenges, this paper propose an innovative and analytical method for AQFP placement. The proposed method aims to achieve minimizing the need for additional buffers. The framework incorporates two key features: (1) entanglement entropy for topology initialization and (2) projection for placement and buffering. These features offer advantages such as avoiding intensive computations, including fix-order Lagrangian optimization in large-scale scenarios, while significantly reducing the required number of buffer rows. Experimental results validate the efficiency of the proposed framework, demonstrating an outstanding 29% and 40% reduction in the amount of required buffers and time compared to the state-of-the-art method.	Yinuo Bai (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Enxin Yi (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Wei W. Xing (The University of Sheffield); Bei Yu (The Chinese University of Hong Kong); Zhou Jin (Super Scientific Software Laboratory, China University of Petroleum-Beijing)
1561	A Software-Hardware Co-design Solution for 3D Inner Structure Reconstruction	Volume imaging (3D model with inner structure) is widely applied to various areas, such as medical diagnosis and archaeology. Especially during the COVID-19 pandemic, there is a great demand for lung CT. However, it is quite time-consuming to generate a 3D model by reconstructing the internal structure of an object. To make things worse, due to the poor data locality of the reconstruction algorithm, researchers are pessimistic about accelerating it with ASIC. Besides the locality issue, we find that the complex synchronization is also a major obstacle for 3D reconstruction. To overcome the problems, we propose a holistic solution using software-hardware co-design. We first provide a unified programming model to cover various 3D reconstruction tasks. Then, we redesign the dataflow of the reconstruction algorithm to improve data locality. In addition, we remove unnecessary synchronizations by carefully analyzing the data dependency. After that, we propose a novel near-memory acceleration architecture, called Waffle, for further improvement. Experiment results show that Waffle in a package can achieve 3.51× ~ 3.96× speedup over a cluster of 10 GPUs with 9.35× ~ 10.97× energy efficiency.	Xingchen Li (Peking University); Zhe Zhou (Peking University); Qilin Zheng (Peking University); Guangyu Sun (Peking University); QianKun Wang (Peking University); Chenhao Xue (School of Integrated Circuits, Peking University)
1564	Plug Your Volt: Protecting Intel Processors against Dynamic Voltage Frequency Scaling based Fault Attacks	This work proposes a new countermeasure principle to defend against Dynamic Voltage Frequency Scaling (DVFS) based fault attacks on modern Intel systems. First, we establish that the fundamental cause of DVFS fault attacks is the ability to independently control the frequency and voltage of a processor. Using this observation, we construct a partition of frequency-voltage tuples into unsafe-safe states based on whether a tuple causes timing violations according to switching circuit theoretic principles. Our countermeasure completely prevents DVFS faults on three Intel generation CPUs: Sky Lake, Kaby Lake R, and Comet Lake. Further, it can also be deployed both as microcode or as model-specific registers at the hardware level, unlike previous countermeasures. Finally, we evaluate a minuscule overhead 0.28% of our countermeasure on SPEC2017.	Nimish Mishra (Indian Institute of Technology Kharagpur); Rahul Arvind Mool (Indian Institute of Technology Kharagpur); Anirban Chakraborty (Indian Institute of Technology, Kharagpur); Debdeep Mukhopadhyay (Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur)

Submissio	Title	Abstract	Authors with Affiliations
1570	Safe Controller Synthesis for Nonlinear Systems via Reinforcement Learning and PAC Approximation	Controller synthesis for nonlinear systems is an important research issue. Deep Neural Network (DNN) control policies obtained through reinforcement learning (RL), though exhibiting good performance in simulations, cannot be applied to safety-critical systems for lack of formal guarantee. To address this, this paper considers fully utilizing the advantages of RL for complex control tasks to obtain a well-performing DNN controller. Then, using PAC (Probably Approximately Correct) techniques, a polynomial surrogate controller with probabilistically controllable approximation error is obtained. Finally, the safety of the control system under the designed polynomial controller is verified using barrier certificate generation. Experiments demonstrate the effectiveness of our method in generating controllers with safety guarantees for systems with high dimensions and degrees.	Xia Zeng (Southwest University); Banglong Liu (East China Normal University); Zhenbing Zeng (Shanghai University); Zhiming Liu (Southwest University); Zhengfeng Yang (East China Normal University)
1572	EVDMARL: Efficient Value Decomposition-based Multi-Agent Reinforcement Learning with Domain-Randomization for Complex Analog Circuit Design Migration	Automated analog circuit design migration significantly alleviates the burden on designers in circuit sizing under various operating conditions. Conventional methods model the migration problem as black-box optimization, requiring excessive iterations of costly simulations to converge. Reinforcement learning exhibits significant promise in transfer learning, as it enables the generation of circuits that fulfill specifications efficiently. The paper proposes a novel value decomposition-based multi-agent reinforcement learning framework, aiming to model complex analog circuits and eliminate the need for manually defined specifications of sub-circuits for new operating conditions. Additionally, it incorporates generalized domain randomization techniques to leverage the varying information across diverse domains. Experiment demonstrates that our algorithm can efficiently generate circuits meeting specifications under new operating conditions in few number of steps, outperforming state-of-the-art methods.	Handa Sun (Fudan University); Zhaori Bi (Fudan University); Wenning Jiang (Fudan University); Ye Lu (Fudan University); Changhao Yan (Fudan University); Fan Yang (Fudan University); Wenchuang Hu (Sichuan University); Sheng-Guo Wang (Fudan University); Dian Zhou (Fudan University); Xuan Zeng (Fudan University)
1577	Trapped by Your WORDs: (Ab)using Processor Exception for Generic Binary Instrumentation on Bare-metal Embedded Devices	Analyzing the security of closed-source drivers and libraries in embedded systems holds significant importance, given their fundamental role in the supply chain. Unlike x86, embedded platforms lack comprehensive binary manipulating tools, making it difficult for researchers and developers to effectively detect and patch security issues in such closed-source components. Existing works either depend on full-fledged operating system features or suffer from tedious corner cases, restricting their application to bare-metal firmware prevalent in embedded environments. In this paper, we present PIFER (Practical Instrumenting Framework for Embedded fiRmware) that enables general and fine-grained static binary instrumentation for embedded bare-metal firmware. By abusing the built-in hardware exception-handling mechanism of the embedded processors, PIFER can perform instrumentation on arbitrary target addresses. Additionally, We propose an instruction translation-based scheme to guarantee the correct execution of the original firmware after patching. We evaluate PIFER against real-world, complex firmware, including Zephyr RTOS, CoreMark benchmark, and a close-sourced commercial product. The results indicate that PIFER correctly instrumented 98.9\% of the instructions. Further, a comprehensive performance evaluation was conducted, demonstrating the practicality and efficiency of our work.	Shipei Qu (Shanghai Jiao Tong University); Xiaolin Zhang (Shanghai Jiao Tong University); Chi Zhang (Shanghai Jiao Tong University); Dawu Gu (Shanghai Jiao Tong University)
1586	Evaluating the Security of Logic Locking on Deep Neural Networks	Deep neural networks are susceptible to model piracy and adversarial attacks when malicious end-users have full access to the model parameters. Recently, a logic locking scheme called HPNN has been proposed. HPNN utilizes hardware root-of-trust to prevent end-users from accessing the model parameters. This paper investigates whether logic locking is secure on deep neural networks. Specifically, it presents a systematic I/O attack that combines algebraic and learning-based approaches. This attack incrementally extracts key values from the network to minimize sample complexity. Besides, it employs a rigorous procedure to ensure the correctness of the extracted key values. Our experiments demonstrate the accuracy and efficiency of this attack on large networks with complex architectures. Consequently, we conclude that HPNN-style logic locking and its variants we can foresee are insecure on deep neural networks.	You Li (Northwestern University); Guannan Zhao (Northwestern University); Yunqi He (Northwestern University); Hai Zhou (Northwestern University)
1609	GNN-assisted Back-side Clock Routing Methodology for Advance Technologies	The back-side metal layers exhibit lower parasitics compared to the front-side layers in advanced technologies, making them suitable for clock-net distribution. In this study, we explore the advantages of using back-side metal layers for clock routing, which is shared with a power delivery network. Our Graph Neural Network (GNN) based framework, effectively distributes the clock-tree between the front and back sides. We address the back-side clock nets' creation by incorporating back-side buffers. Our results demonstrate better clock and full-chip metrics represented by an increase of up to 13% in the effective frequency with equivalent power consumption, using 3 nm technology.	Nesara Eranna Bethur (Georgia Institute of Technology); Pruek Vanna-iampikul (Georgia Institute of Technology); Odysseas Zografos (imec); Lingjun Zhu (Georgia Institute of Technology); Giuliano Sisto (IMEC); Dragomir Milojevic (IMEC); Alberto Garcia-Ortiz (ITEM (U.Bremen)); Geert Hellings (IMEC); Julien Ryckaert (IMEC); Francky Catthoor (IMEC); Sung Kyu Lim (Georgia Tech)
1612	Crop: An Analytical Cost Model for Cross-Platform Performance Prediction of Tensor Programs	Learn-based cost models used for tensor compiler auto-tuning often suffer from poor performance when trained on one hardware platform and applied to another. This issue necessitates collecting performance data for each potential platform during model deployment, incurring significant overhead. We propose Crop, a comprehensive and universal analytical cost model designed for cross-platform performance prediction of tensor programs. Crop decouples program features and hardware features, gathering hardware-independent program features on one platform and predicts their performance based on parametric hardware features for given platforms. Crop achieves comparable levels of prediction accuracy to that of a learn-based cost model while ensuring portability.	Xinyu Sun (University of Science and Technology of China); Yu Zhang (University of Science and Technology of China); Shuo Liu (University of Science and Technology of China); Yi Zhai (University of Science and Technology of China)
1614	MASC: A Memory-Efficient Adjoint Sensitivity Analysis through Compression Using Novel Spatiotemporal Prediction	Adjoint sensitivity analysis is critical in modern integrated circuit design and verification, but its computational intensity grows significantly with the size of the circuit, the number of objective functions, and the accumulation of time points. This growth can impede its wider application. The intimate link between the forward integration in transient analysis and the reverse integration in adjoint sensitivity analysis allows for the retention of Jacobian matrices from transient analysis, thereby speeding up sensitivity analysis. However, Jacobian matrices across multiple timesteps are often so large that they cannot be stored in memory during the forward integration process, necessitating disk storage and incurring significant I/O overhead. To address this, we develop a memory-efficient sensitivity analysis method that utilizes data compression to minimize memory overhead during simulation and enhance analysis efficiency. Our compression method can efficiently compress the sparse tensor that contains the Jacobian matrices over time by exploiting the spatiotemporal characteristics of the data and circuit attributes. It also introduces a shared-indices technique, a cutting-edge spatiotemporal prediction model, and robust residual encoding. We evaluate our compression method on 7 datasets from real-world simulations and demonstrate that it can reduce the memory requirements for storing Jacobian matrices by more than 16x on average, which is significantly more efficient than other state-of-the-art compression techniques.	Chenxi Li (13121010307); Boyuan Zhang (Luddy School of Informatics, Computing, & Engineering Indiana University Bloomington); Yongqiang Duan (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Yang Li (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Zuochang Ye (Tsinghua University); Weifeng Liu (Super Scientific Software Laboratory, China University of Petroleum-Beijing); Dingwen Tao (Luddy School of Informatics, Computing, & Engineering Indiana University Bloomington); Zhou Jin (Super Scientific Software Laboratory, China University of Petroleum-Beijing)

Submissio	Title	Abstract	Authors with Affiliations
1615	Energy Efficient Dual Designs of FeFET-Based Analog In-Memory Computing with Inherent Shift-Add Capability	Deep neural networks (DNNs) have significantly advanced over the past decade, embracing diverse artificial intelligence (AI) tasks. In-memory computing (IMC) architecture emerges as a promising paradigm, improving the energy efficiency of multiply-and-accumulate (MAC) operations within DNNs by integrating the parallel computations within the memory arrays. Various high-precision analog IMC array designs have been developed based on both SRAM and emerging non-volatile memories (NVMs). These designs perform MAC operations of partial input and weight, with the corresponding partial products then fed into shift-add circuitry to produce the final MAC results. However, existing works often present intricate shift-add process for weight. The traditional digital shift-add process is limited in throughput due to time-multiplexing of ADCs, and advancing the shift-add process to the analog domain necessitates customized circuit implementations, resulting in compromises in energy and area efficiency. Furthermore, the joint optimization of MAC operations and the weight shift-add process is rarely explored. In this paper, we propose novel, energy efficient dual designs of ferroelectric FET (FeFET) based high precision analog IMC featuring inherent shift-add capability. We introduce a FeFET based IMC paradigm that performs partial MAC in each column, and inherently integrates the shift-add process for 4-bit weights by leveraging FeFET's analog storage characteristics. This effectively eliminates the need for additional dedicated shift-add circuitry in multi-bit weight processing. This paradigm supports both 2's complement mode (2CM) and non-2's complement mode (N2CM) MAC, thereby offering flexible support for 4-/8-bit weight data in 2's complement format. Building upon this paradigm, we propose novel FeFET based dual designs, CurFe for the current mode and ChgFe for the charge mode, to accommodate the high precision analog domain IMC architecture. Evaluation results at circuit and system levels indicate that the circuit/system-level energy efficiency of the proposed FeFET-based analog IMC is 1.56X/1.37X higher when compared to the state-of-the-art analog IMC designs.	Zeyu Yang (Zhejiang University); Qingrong Huang (Zhejiang University); Yu Qian (Zhejiang University); Kai Ni (University of Notre Dame); Thomas Kämpfe (Fraunhofer IPMS); Xunzhao Yin (Zhejiang University)
1617	Cross-Layer Reliability Evaluation and Efficient Hardening of Large Vision Transformers Models	Vision Transformers (ViTs) are highly accurate Machine Learning (ML) models. However, their large size and com- plexity increase the expected error rate due to hardware faults. Measuring the error rate of large ViT models is challenging, as conventional microarchitectural fault simulations can take years to produce statistically significant data. This paper proposes a two-level evaluation based on data collected through more than 70 hours of neutron beam experiments and more than 600 hours of software fault simulation. We consider 12 ViT models executed in 2 NVIDIA GPU architectures. We first characterize the fault model in ViT's kernels to identify the faults that are more likely to propagate to the output. We then design dedicated procedures efficiently integrated into the ViT to locate and correct these faults. We propose Maximum corrupted Malicious values (MaxiMals), an experimentally tuned low-cost mitigation solution to reduce the impact of transient faults on ViTs. We demonstrate that MaxiMals can correct 90.7% of critical failures, with execution time overheads as low as 5.61%.	Lucas Roquet (University of Rennes); Fernando Fernandes dos Santos (INRIA); Paolo Rech (University of Trento); Marcello Traiola (Inria / IRISA); Olivier Sentieys (INRIA); Angeliki Kritikakou (Univ Rennes, Inria, CNRS, IRISA)
1625	MCU-Wide Timing Side Channels and Their Detection	Microarchitectural timing side-channels are known to compromise security in computing systems with shared buffers (like caches) and/or parallel execution of attacker and victim tasks. Counterintuitively, such threats exist even in simple microcontrollers lacking such features. This paper describes previously neglected SoC-wide timing side-channels and presents a new formal method for detection. In a case study on Pulpissimo, our method detected a vulnerability to a previously unknown attack variant that allows an attacker to obtain information about a victim's memory accesses. We applied a conservative fix and verified security of the SoC against the considered class of timing side-channels.	Johannes Müller (RPTU Kaiserslautern-Landau); Anna Lena Duque Antón (RPTU Kaiserslautern-Landau); Lucas Deutschmann (University of Kaiserslautern-Landau); Dino Mehmedagić (University of Kaiserslautern-Landau); Cristiano Rodrigues (Universidade do Minho); Daniel Oliveira (Universidade do Minho); Mohammad Rahmani Fadiheh (Technische Universität Kaiserslautern); Keerthikumara Devarajegowda (Siemens EDA); Sandro Pinto (Universidade do Minho); Dominik Stoffel (TU Kaiserslautern); Wolfgang Kunz (TU Kaiserslautern)
1627	SHERLOCK: Scheduling Efficient and Reliable Bulk Bitwise Operations in NVMs	Bulk bitwise operations are commonplace in application domains such as databases, web search, cryptography, and image processing. The ever-growing volume of data and processing demands of these domains often result in high energy consumption and latency in conventional systems, mainly due to extensive data movement. Non-volatile memory (NVM) technologies, such as RRAM, PCM and STT-MRAM, facilitate conducting bulk-bitwise logic operations in-memory (CIM), eliminating the data movement. However, mapping complex real-world applications to these CIM-capable NVMs is non-trivial and can lead to sub-optimal performance. To address this, we present SHERLOCK, a novel mapping and scheduling method tailored to exploit the unique characteristics of these systems. SHERLOCK collaboratively optimizes reliability and performance, a previously overlooked aspect that significantly affects both the correctness and throughput of these systems. Our method also leverages the granularity of CIM operations to reduce the number of write operations and, hence, energy consumption. Our evaluation on three representative applications from different domains shows that SHERLOCK outperforms the state-of-the-art in terms of performance and energy consumption.	Hamid Farzaneh (Technische Universität Dresden); Joao Paulo Cardoso De Lima (Technische UniversitÃ¤t Dresden); Ali Nezhadi Khelejani (Karlsruhe institute of technology (KIT)); Asif Ali Khan (TU Dresden); Mahta Mayahinia (Karlsruhe institute of technology (KIT)); Mehdi Tahoori (Karlsruhe Institute of Technology); Jeronimo Castrillon (TU Dresden)
1629	Lost and Found in Speculation: Hybrid Speculative Vulnerability Detection	Microarchitectural attacks represent a challenging and persistent threat to modern processors, exploiting inherent design vulnerabilities in processors to leak sensitive information or compromise systems. Of particular concern is the susceptibility of Speculative Execution, a fundamental part of performance enhancement, to such attacks. We introduce Specure, a novel pre-silicon verification method composing hardware fuzzing with Information Flow Tracking (IFT) to address speculative execution leakages. Integrating IFT enables two significant and non-trivial enhancements over the existing fuzzing approaches: i) automatic detection of microarchitectural information leakages vulnerabilities without golden model and ii) a novel Leakage Path coverage metric for efficient vulnerability detection. Specure identifies previously overlooked speculative execution vulnerabilities on the RISC-V Boom processor and explores the vulnerability search space 6.45× faster than existing fuzzing techniques. Moreover, Specure detected known vulnerabilities 20× faster.	Mohamadreza Rostami (Technical University of Darmstadt); Shaza Zeitouni (TU Darmstadt); Rahul Kande (Texas A&M University); Chen Chen (Texas A&M University); Pouya Mahmoody (TU Darmstadt); Jeyavijayan Rajendran (Texas A&M University); Ahmad-Reza Sadeghi (Technische Universitaet Darmstadt)
1630	Using Probabilistic Model Rollouts to Boost the Sample Efficiency of Reinforcement Learning for Automated Analog Circuit Sizing	Despite recent advances in algorithms such as the use of reinforcement learning, analog circuit sizing optimization remains a challenging task that demands numerous circuit simulations, hence extensive CPU times. This paper presents the application of Model-Based Policy Optimization (MBPO) to boost the sample efficiency of reinforcement learning for analog circuit sizing. This method leverages an ensemble of probabilistic dynamic models to generate short rollouts branched from real data for a fast extensive exploration of the design space, thereby speeding up the learning process of the reinforcement learning agent and enhancing its convergence. Integrated in the Twin Delayed DDPG (TD3) algorithm, our new model-based TD3 (MBTD3) approach has been validated on analog circuits of different complexity, outperforming the existing model-free TD3 method by achieving power/area-optimal design solutions with up to 3x fewer simulations and half the run time. In addition, for larger analog circuits, we present a multi-agent version of MBTD3 in which multiple simultaneous agents use global probabilistic models for sizing different blocks within the circuit. Demonstrated for a complex data receiver circuit, it surpasses the model-free multi-agent TD3 method at 2x less simulations and half the run time. These novel methods highly boost the efficiency of automated analog circuit sizing.	Mohsen Ahmadzadeh (KU Leuven); Georges Gielen (KU Leuven)

Submissio	Title	Abstract	Authors with Affiliations
1644	SmartATPG: A Learning-based Automatic Test Pattern Generation with Graph Convolutional Network and Reinforcement Learning	Automatic test pattern generation (ATPG) is a critical technology in integrated circuit testing. It searches for effective test patterns to detect all possible faults in the circuit as entirely as possible, thereby ensuring chip yield and improving chip quality. However, the process of searching for test patterns is NP-complete. At the same time, the large amount of backtracking generated during the search for test patterns can directly affect the performance of ATPG. In this paper, a learning-based ATPG framework SmartATPG is proposed to search for high-quality test patterns, reduce the number of backtracking during the search process, and thereby improve the performance of ATPG. SmartATPG utilizes convolutional network (GCN) to fully extract circuit feature information and efficiently explore the ATPG search space through reinforcement learning (RL). Experimental results show that the proposed SmartATPG can perform better than traditional heuristic strategies and deep learning heuristic strategies on most benchmark circuits.	Wenxing Li (Institute of Computing Technology, Chinese Academy of Sciences); Hongqin Lyu (Institute of Computing Technology, Chinese Academy of Sciences); Shengwen Liang (Institute of Computing Technology, Chinese Academy of Sciences); Tiancheng Wang (Institute of Computing Technology, Chinese Academy of Sciences); Huawei Li (Institute of Computing Technology, Chinese Academy of Sciences)
1649	A Real-time Execution System of Multimodal Transformer through PIM-GPU Collaboration	Multimodal transformer excels in various applications, but faces great challenges such as high memory consumption and limited data reuse that hinder real-time performance. To address these issues, we propose a processing-in-memory (PIM)-GPU collaboration oriented compiler that optimizes the acceleration of multimodal transformers. The PIM-GPU synergy adapts well to multimodal transformers and improves execution time through dynamic programming algorithms. In addition, we introduce a tailored PIM allocation algorithm for variable-length inputs to further increase efficiency. Experimental results show an average end-to-end speedup of 15x.	Shengyi Ji (College of Information Science and Engineering, Hunan University); Chubo Liu (College of Information Science and Engineering, Hunan University); Yan Ding (College of Information Science and Engineering, Hunan University); Qing Liao (School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)); Zhuo Tang (College of Information Science and Engineering, Hunan University)
1676	Design of a Quantum Walk Circuit to Solve the Subset-Sum Problem	Search algorithms based on quantum walks have emerged as a promising approach to solve computational problems across various domains, including combinatorial optimization and cryptography. Stating a generic search problem in terms of a (quantum) search over a graph makes the efficiency of the algorithmic method depend on the structure of the graph itself. In this work, we propose a complete implementation of a quantum walk search on Johnson graphs, speeding up the solution of the subset-sum problem. We provide a detailed design of each sub-circuit, quantifying their cost in terms of gate number, depth, and width. We additionally compare our solution against a Grover quantum search approach, showing a reduction of the T-depth cost compared to it. The proposed design provides a building block for the construction of efficient quantum search algorithms that can be modeled on Johnson graphs, filling the gap with the existing theoretical complexity analyses.	Giacomo Lancellotti (Politecnico di Milano); Simone Perriello (Politecnico di Milano); Alessandro Barenghi (Politecnico di Milano - DEIB); Gerardo Pelosi (Politecnico di Milano)
1677	GATE-SiP: Enabling Authenticated Encryption Testing in Systems-in-Package	A heterogeneous integrated system in package (SIP) system integrates chiplets outsourced from different vendors into the same substrate for better performance. However, during post-integration testing, the sensitive testing data designated for a specific chiplet can be blocked, tampered or sniffed by other malicious chiplets. This paper proposes GATE-SiP which is an authenticated partial encryption protocol to enable secure testing. Within GATE-SiP, the sensitive testing pattern will only be sent to the authenticated chiplet. In addition, partial encryption of the sensitive data prevents data sniff threats without causing significant penalties on timing overhead. Extensive simulation results show the GATE-SiP protocol only brings 6.74% and 14.31% on area and timing overhead, respectively.	GALIB IBNE HAIDAR (University of Florida); Kimia Zamiri Azar (University of Florida); Hadi Mardani Kamali (University of Central Florida); Mark Tehranipoor (Intel Charles E. Young Preeminence Endowed Chair Professor in Cybersecurity, Associate Chair for Research and Strategic Initiatives, ECE Department, University of Florida); Farimah Farahmandi (University of Florida)
1683	DeepRIoT: Continuous Integration and Deployment of Robotic-IoT Applications	We present DeepRIoT, a continuous integration and continuous deployment (CI/CD) based architecture that accelerates the learning and deployment of a Robotic-IoT system trained from deep reinforcement learning (RL). We adopted a multi-stage approach that agilely trains a multi-objective RL controller in the simulator. We then collected traces from the real robot to optimize its plant model, and used transfer learning to adapt the controller to the updated model. We automated our framework through CI/CD pipelines, and finally, with low cost, succeeded in deploying our controller in a real F1tenth car that is able to reach the goal and avoid collision from a virtual car through mixed reality.	Meixun Qu (Vienna University of Technology); Jie He (Vienna University of Technology); Zlatan Tucakovic (Vienna University of Technology); Ezio Bartocci (TU Wien); Dejan Nickovic (AIT Austrian Institute of Technology); Haris Isakovic (Vienna University of Technology); Radu Grosu (Vienna University of Technology)
1693	Hardware-Aware Neural Dropout Search for Reliable Uncertainty Prediction on FPGA	The increasing deployment of artificial intelligence (AI) for critical decision-making amplifies the necessity for trustworthy AI, where uncertainty estimation plays a pivotal role in ensuring trustworthiness. Dropout-based Bayesian Neural Networks (BayesNNs) are prominent in this field, offering reliable uncertainty estimates. Despite their effectiveness, existing dropout-based BayesNNs typically employ a uniform dropout design across different layers, leading to suboptimal performance. Moreover, as diverse applications require tailored dropout strategies for optimal performance, manually optimizing dropout configurations for various applications is both error prone and labor-intensive. To address these challenges, this paper proposes a novel neural dropout search framework that automatically optimizes both the dropout-based BayesNNs and their hardware implementations on FPGA. We leverage one-shot supernet training with an evolutionary algorithm for efficient dropout optimization. A layer-wise dropout search space is introduced to enable the automatic design of dropout-based BayesNNs with heterogeneous dropout settings. Extensive experiments demonstrate that our proposed framework can effectively find design configurations on the Pareto frontier. Compared to manually-designed dropoutbased BayesNNs on GPU, our search approach produces FPGA designs that can achieve up to 33× higher energy efficiency. Compared to state-of-the-art FPGA designs of BayesNN, the solutions from our approach can achieve higher algorithmic performance. Our designs and tools will be open-source upon paper acceptance.	Zehuan Zhang (Imperial College London); Hongxiang Fan (Samsung AI Center, University of Cambridge); Hao (Mark) Chen (Imperial College London); Lukasz Dudziak (Samsung AI Center); Wayne Luk (Imperial College)
1703	CircuitVAE: Efficient and Scalable Latent Circuit Optimization	Automatically designing fast and space-efficient digital circuits is challenging because circuits are discrete, must exactly implement the desired logic, and are costly to simulate. We address these challenges with CircuitVAE, a search algorithm that embeds computation graphs in a continuous space and optimizes a learned surrogate of physical simulation by gradient descent. By carefully controlling overfitting of the simulation surrogate and ensuring diverse exploration, our algorithm is highly sample-efficient, yet gracefully scales to large problem instances and high sample budgets. We test CircuitVAE by designing binary adders across a large range of sizes, IO timing constraints, and sample budgets. Our method excels at designing large circuits, where other algorithms struggle: compared to reinforcement learning and genetic algorithms, CircuitVAE typically finds 64-bit adders which are smaller and faster using less than half the sample budget. We also find CircuitVAE can design state-of-the-art adders in a real-world chip, demonstrating that our method can outperform commercial tools in a realistic setting.	Jialin Song (NVIDIA); Aidan Swope (N/A); Robert S. Kirby (Nvidia); Rajarshi Roy (NVIDIA); Saad Godil (NVIDIA Corporation); Jonathan Raiman (OpenAI); Bryan Catanzaro (NVIDIA)
1720	DGR: Differentiable Global Router	Modern VLSI design flows necessitate fast and high-quality global routers. In this paper, we introduce DGR, a GPU-accelerated, differentiable global router capable of concurrent optimization for millions of nets, which we aim to open-source. Our innovation lies in the development of a routing Directed Acyclic Graph (DAG) forest to represent the 2D pattern routing space for all nets, enabling coordinated selection of Steiner trees and 2-pin routing paths from a global perspective. For efficient search within the DAG forest, we relax the discrete search space to be continuous and develop a differentiable solver accelerated by deep learning toolkits on GPUs. Experimental results demonstrate that DGR substantially mitigates routing overflow while concurrently reducing total wirelengths from 0.95% to 4.08% and via numbers from 1.28% to 2.54% in congested testcases compared to state-of-the-art academic global routers. Additionally, DGR exhibits favorable scalability in both runtime and memory with respect to the number of nets.	Wei Li (Carnegie Mellon University); Rongjian Liang (NVIDIA); Anthony Dimitri Armand Agnesina (NVIDIA); Haoyu Yang (NVIDIA Corp.); Chia-Tung Ho (UCSD); Anand Rajaram (NVIDIA); Haoxing Ren (NVIDIA Corporation)

Submissio	Title	Abstract	Authors with Affiliations
1726	EGMA: Enhancing Data Reuse and Workload Balancing in Message Passing GNN Acceleration via Gram Matrix Optimization	Message Passing-based Graph Neural Networks (GNNs) have been widely used to analyze graph data, in which complex vertex and edge operations are performed via the exchange of information between connected vertices. Such complex GNN operations are highly dependent on the graph structure and can no longer characterized as general sparse-dense or matrix multiplications. Consequently, current data reuse and workload balance optimizations have limited applicability to Message Passing-based GNN acceleration. In this paper, we leverage the mathematical insights from Gram Matrix to simultaneously exploit data reuse and workload balancing opportunities for GNN accelerations. Upon this, we further propose a novel accelerator shortly termed as EGMA that can efficiently facilitate a wide range of GNN models with much-improved data reuse and workload balance. Consequently, EGMA can achieve performance speedup by 1.57×, 1.72×, and 1.43× and energy reduction by 38.19%, 34.02%, and 24.54% on average compared to Betty, FlowGNN, and ReGNN, respectively.	Fangzhou Ye (University of Central Florida); Lingxiang Yin (University of Central Florida); Amir Ghazizadeh Ahsaei (Graduate Student); Hao Zheng (University of Central Florida)
1733	Uncovering Software-Based Power Side-Channel Attacks on Apple M1/M2 Systems	Traditionally, power side-channel analysis requires physical access to the target device, as well as specialized devices to measure the power consumption with enough precision. Recently research has shown that on x86 platforms, on-chip power meter capabilities exposed to a software interface might be used for power side-channel attacks without physical access. In this paper, we show that such software-based power side-channel attack is also applicable on Apple silicon (e.g., M1/M2 platforms), exploiting the System Management Controller (SMC) and its power-related keys, which provides access to the on-chip power meters through a software interface to user space software. We observed data-dependent power consumption reporting from such SMC keys and analyzed the correlations between the power consumption and the processed data. Our work also demonstrated how an unprivileged user mode application successfully recovers bytes from an AES encryption key from a cryptographic service supported by a kernel mode driver in MacOS. We have also studied the feasibility of performing frequency throttling side-channel attack on Apple silicon. Furthermore, we discuss the impact of software-based power side-channels in the industry, possible countermeasures, and the overall implications of software interfaces for modern on-chip power management systems.	Nikhil Chawla (Intel Corp.); Chen Liu (Intel Corp.); Abhishek Chakraborty (Intel Corp.); Igor Chervatyuk (Intel Corp.); Thais Moreira Hamasaki (Intel Corp.); Ke Sun (Intel Corp.); Henrique Kawakami (Intel Corp.)
1760	Q-Pilot: Field Programmable Qubit Array Compilation with Flying Ancillas	Neutral atom arrays, particularly the field programmable quantum array (FPQA) with atom movement, show promise for quantum computing. FPQA has a dynamic qubit connectivity, facilitating cost-effective execution of long-range gates, but also poses new challenges in compilation. Inspired by FPGA compilation strategy, we develop a router, \name, that leverages flying ancillas to implement 2-Q gates between data qubits mapped to fixed atoms. Equipped with domain-specific routing techniques, \name achieves 1.4$\times$, 27.7$\times$, and 6.7$\times$ reductions in circuit depth for 100-qubit random, quantum simulation, and quantum approximate optimization algorithm circuits, respectively, compared to alternative fixed architectures.	Hanrui Wang (Massachusetts Institute of Technology); Daniel Bochen Tan (University of California, Los Angeles); Pengyu Liu (CMU); Yilian Liu (Cornell University); Jiaqi Gu (Arizona State University); Jason Cong (UCLA); Song Han (MIT)
1762	VAE-HDC: Efficient and Secure Hyper-dimensional Encoder Leveraging Variation Analog Entropy	Hyperdimensional computing (HDC) is a bio-inspired machine learning paradigm utilizing hyperdimensional spaces for data representation. HDC significantly improves the ability to learn from sparse data and enhances noise robustness, and also enables parallel computation. Despite these advantages, HDC's reliance on high dimensionality and operational simplicity can lead to increased hardware costs and potential security vulnerabilities. This paper introduces a novel HDC encoding strategy using variation-based analog entropy (VAE), aiming to reduce memory footprint, lower power/energy consumption, and enhance security with physically-unclonable entropy generation. The VAE cell, with high entropy robustness 30.23-57.76 dB SNR and a small footprint 10 transistors, allows HDC to achieve a 14.3x reduction in vector dimensions, a 4.4x decrease in unit entropy cell area, and a 2% increase in accuracy compared to binary/multi-bit HDC. These benefits lead to a 1.3-4.4x area and a 327x leakage power reduction when compared to an SRAM baseline. We have designed custom low-power circuits that enable end-to-end analog entropy storage, distribution management, binding, permutation, and bundling. This analog implementation prevents data conversion during feature vector encoding, thereby significantly enhancing energy efficiency 48.5 nJ per query. Furthermore, with hardware-secured basis vectors, data security is significantly improved, as evidenced by the markedly degraded visual distinguish-ability of retrieved image data and maximum of 11dB lower PSNR.	Boyang Cheng (University of Notre Dame); Jianbo Liu (University of Notre Dame); Steven Davis (University of Notre Dame); Zephan M. Enciso (University of Notre Dame); Yiyang Zhang (University of Notre Dame); Ningyuan Cao (University of Notre Dame)
1776	Data-Efficient Conformalized Interval Prediction of Minimum Operating Voltage Capturing Process Variations	Accurate minimum operating voltage (Vmin) prediction is a critical element in manufacturing tests. Conventional methods lack coverage guarantees in interval predictions. Conformal Prediction (CP), a distribution-free machine learning approach, excels in providing rigorous coverage guarantees for interval predictions. However, standard CP predictors may fail due to a lack of knowledge of process variations. We address this challenge by providing principled conformalized interval prediction in the presence of process variations with high data efficiency, where a few additional chips are utilized for calibration. We demonstrate the superiority of the proposed method on industrial 16nm chip data.	Yuxuan Yin (University of California, Santa Barbara); Rebecca Chen (NXP Semiconductors); Chen He (NXP Semiconductors); Peng Li (University of California, Santa Barbara)
1779	Data-driven HLS optimization for reconfigurable accelerators	High-Level Synthesis (HLS) has played a pivotal role in making FPGAs accessible to a broader audience by facilitating high-level device programming and rapid microarchitecture customization through the use of directives. However, manually selecting the right directives can be a formidable challenge for programmers lacking a hardware background.This paper introduces an ultra-fast, knowledge-based HLS design optimization method that automatically extracts and applies the most promising directive configurations to the original source code. This optimization approach is entirely data-driven, offering a generalized HLS tuning solution without reliance on Quality of Result (QoR) models or meta-heuristics. We design, implement, and evaluate our methodology using over 100 applications sourced from well-established benchmark suites and GitHub repositories, all running on a Xilinx ZCU104 FPGA. The results are promising, including an average geometric mean speedup of $\times$1.35 and $\times$7.2 compared to over-provisioning and designer-optimized designs, respectively. Additionally, it demonstrates a high design feasibility score and maintains an average inference latency of 38ms. Comparative analysis with traditional genetic algorithm-based Design Space Exploration (DSE) methods and State-of-the-Art (SoA) approaches reveals that it produces designs of similar quality but at speeds 2-3 orders of magnitude faster. This suggests that it is a highly promising solution for ultra-fast and automated HLS optimization.	Aggelos Ferikoglou (National Technical University of Athens); Andreas Kakolyris (National Technical University of Athens); Vasilis Kypriotis (National Technical University of Athens); Dimosthenis Masouros (National Technical University of Athens); Dimitrios Soudris (NTUA); Sotirios Xydis (National Technical University of Athens)
1783	TBNet: A Neural Architectural Defense Framework Facilitating DNN Model Protection in Trusted Execution Environments	Trusted Execution Environments (TEEs) have become a promising solution to secure DNN models on edge devices. However, existing solutions either provide inadequate protection or introduce large performance overhead. This paper presents TBNet, a TEE-based defense framework that protects DNN model from a neural architectural perspective. TBNet generates a novel Two-Branch substitution model, to exploit (1) the computational resources in untrusted Rich Execution Environment (REE) for latency reduction and (2) the physically-isolated TEE for model protection. Experimental results on a Raspberry Pi across diverse DNN model architectures and datasets demonstrate that TBNet achieves efficient model protection at a low cost.	Ziyu Liu (Northeastern University); Tong Zhou (Northeastern University); Yukui Luo (University of Massachusetts Dartmouth); Xiaolin Xu (Northeastern University)

Submissio	Title	Abstract	Authors with Affiliations
1784	AdderNet 2.0: Optimal FPGA Acceleration of AdderNet with Activation-Oriented Quantization and Fused Bias Removal based Memory Optimization	Emerging proposals, such as AdderNet, exploit efficient arithmetic alternatives to the Multiply-ACcumulate (MAC) operations in convolutional neural networks (CNNs). AdderNet adopts an ℓ1-norm based feature extraction kernel, which shows nearly identical model accuracy as compared to the CNN counterparts and can achieve considerable hardware savings due to simpler Sum-of-Absolute-Difference (SAD) operations. Nevertheless, existing AdderNet-based accelerator designs still face critical implementation challenges, such as inefficient model quantization, excessive feature memory overheads, and sub-optimal resource utilization. This paper presents AdderNet 2.0, an optimal AdderNet based accelerator architecture with a novel Activation-Oriented Quantization (AOQ) strategy, a Fused Bias Removal (FBR) scheme for on-chip feature memory bitwidth reduction, and an improved PE design to improve resource utilization. The proposed AdderNet 2.0 accelerator designs were implemented on Xilinx Kria KV-260 FPGA. Experimental results show that INT6 accelerator design achieves up to 3.78× DSP density improvement, and 24% LUT, 40% FF, and 2.1× BRAM savings compared to the baseline CNN design.	Yunxiang Zhang (Binghamton University); Omar Kailani (Binghamton University); Wenfeng Zhao (Binghamton University)
1788	Deep Harmonic Finesse: Signal Separation in Wearable Systems with Limited Data	We present a method, referred to as Deep Harmonic Finesse (DHF), for separation of non-stationary quasi-periodic signals when limited data is available. The problem frequently arises in wearable systems in which, a combination of quasi-periodic physiological phenomena give rise to the sensed signal, and excessive data collection is prohibitive. Our approach utilizes prior knowledge of time-frequency patterns in the signals to mask and in-paint spectrograms. This is achieved through an application-inspired deep harmonic neural network coupled with an integrated pattern alignment component. The network's structure embeds the implicit harmonic priors within the time-frequency domain, while the pattern-alignment method transforms the sensed signal, ensuring a strong alignment with the network. The effectiveness of the algorithm is demonstrated in the context of non-invasive fetal monitoring using both synthesized and in vivo data. When applied to the synthesized data, our method exhibits significant improvements in signal-to-distortion ratio (26% on average) and mean squared error (80% on average), compared to the best competing method. When applied to in vivo data captured in pregnant animal studies, our method improves the correlation error between estimated fetal blood oxygen saturation and the ground truth by 80.5% compared to the state of the art.	Mahya Saffarpour (University of California Davis); Weitai Qian (University of California Davis); Kourosh Vali (University of California Davis); Begum Kasap (University of California Davis); Herman L. Hedriana (UC Davis School of Medicine); Soheil Ghiasi (University of California, Davis)
1820	Improving the Efficiency of In-Memory-Computing Macro with a Hybrid Analog-Digital Computing Mode for Lossless Neural Network Inference	Analog in-memory-computing (IMC) is an attractive technique with a higher energy efficiency to process machine learning workloads. However, the analog computing scheme suffers from large interface circuit overhead. In this work, we propose a macro with a hybrid analog-digital mode computation to reduce the precision requirement of the interface circuit. Considering the distribution of the multiplication and accumulation (MAC) value, we propose a nonlinear transfer function of the computing circuits by only accurately computing low MAC value in the analog domain, with a digital mode to deal with the high MAC value with smaller possibility. Silicon measurement results show that the proposed macro could achieve 160 GOPS/mm^2 area efficiency and 25.5 TOPS/W for 8b/8b matrix computation. The architectural-level evaluation for real workloads shows that the proposed macro can achieve up to 2.92x higher energy efficiency than conventional analog IMC designs.	Qilin Zheng (Duke University); Ziru Li (Duke University); Jonathan Hao-Cheng Ku (Duke University); Yitu Wang (Duke University); Brady Taylor (Duke University); Deliang Fan (Johns Hopkins University); Yiran Chen (Duke University)
1840	Beyond Conventional Defenses: Proactive and Adversarial-Resilient Hardware Malware Detection using Deep Reinforcement Learning	This research investigates the vulnerability of ML-enabled Hardware Malware Detection(HMD) methods to adversarial attacks. We introduce proactive and robust adversarial learning and defense based on Deep Reinforcement Learning(DRL). First, highly effective adversarial attacks are employed to circumvent detection mechanisms. Subsequently, an efficient DRL technique based on Advantage Actor-Critic(A2C) is presented to predict adversarial attack patterns in real-time. Next, ML models are fortified through adversarial training to enhance their defense capabilities against both malware and adversarial attacks. To achieve greater efficiency, a constraint controller using Upper Confidence Bounds(UCB) algorithm is proposed that dynamically assigns defense responsibilities to specialized RL agents.	Zhangying He (California State University, Long Beach); Houman Homayoun (University of California Davis); Hossein Sayadi (California State University, Long Beach)
1842	On the Design of Novel Attention Mechanism for Enhanced Efficiency of Transformers	We present a new XOR-based attention function for efficient hardware implementation of transformers. While standard attention relies on matrix multiplication, we propose replacing the computation of this attention function with bitwise XOR operations. We mathematically analyze the information-theoretic properties of multiplication-based attention, demonstrating that it preserves input entropy, and then show that XOR-based attention approximately preserves the entropy of its input. Across various simple tasks, including arithmetic, sorting, translation, and text generation, we show comparable performance to baseline methods using scaled GPT models. XOR-based attention shows substantial improvement in power, latency, and area compared to the multiplication-based attention function.	Sumit K. Jha (Florida International University); Susmit Jha (Computer Science Laboratory, SRI); Rickard Ewetz (University of Central Florida); Alvaro Velasquez (University of Central Florida)
1851	Efficient Open Modification Spectral Library Searching in High-Dimensional Space with Multi-Level-Cell Memory	Open Modification Search (OMS) is a promising algorithm for mass spectrometry analysis that enables the discovery of modified peptides. However, OMS encounters challenges as it exponentially extends the search scope. Existing OMS accelerators either have limited parallelism or struggle to scale effectively with growing data volumes. In this work, we introduce an OMS accelerator utilizing multi-level-cell (MLC) RRAM memory to enhance storage capacity by 3x. Through in-memory computing, we achieve 1.7x to 76.7x faster data processing with two to three orders of magnitude energy efficiency improvement. The functionality is tested on a fabricated MLC RRAM chip. To address errors from memory, we leverage hyperdimensional computing, providing robustness by tolerating up to 10% memory errors while delivering massive parallelism in hardware.	Keming Fan (University of California, San Diego); Wei-Chen Chen (Stanford University); Sumukh Pinge (University of California, San Diego); H.-S. Philip Wong (Stanford University); Tajana Rosing (UCSD)
1852	Bitwise Adaptive Early Termination in Hyperdimensional Computing Inference	Hyperdimensional computing (HDC), a powerful paradigm for cognitive tasks, often demands hypervectors of high dimensions (e.g., 10,000) to achieve competitive accuracy. However, processing such large-dimensional data poses challenges for performance and energy efficiency, particularly on resource-constrained devices. In this paper, We present a framework to terminate bit-serial HDC inference early when sufficient confidence is attained in the prediction. This approach integrates a Naive Bayes model to replace the conventional associative memory in HDC. This transformation allows for a probabilistic interpretation of the model outputs, steering away from mere similarity measures. We reduce more than 70% of bits that need to be processed while maintaining comparable accuracy across diverse benchmarks. In addition, We show the adaptability of our early termination algorithm during on-the-fly learning scenarios.	Wei-Chen Chen (Stanford University); H.-S. Philip Wong (Stanford University); Sara Achour (Stanford University)
1872	Learn-by-Compare: Analog Performance Prediction using Contrastive Regression with Design Knowledge	This paper introduces Learn-by-Compare (LbC), a novel approach for analog performance modeling by employing semi-supervised contrastive regression. LbC employs a deep neural network encoder to come up with latent representations of sizing solutions by comparing similarity/dissimilarity of the underlying performance. Leveraging two levels of transistor-level sizing data augmentation (DA), namely LS-DA and GS-DA, LbC produces new data samples by employing design knowledge. Experimental results highlight LbC's superior predictive accuracy compared to traditional regression methods. Offering a streamlined semi-supervised learning methodology, LbC effectively incorporates simple design knowledge and representation learning for efficient analog performance modeling.	Zihu Wang (University of California, Santa Barbara); Karthik Somayaji N.S. (University of California, Santa Barbara); Peng Li (University of California, Santa Barbara)

Submissio	Title	Abstract	Authors with Affiliations
1889	RexBDDs: Reduction-on-Edge Complement-and-Swap Binary Decision Diagrams	We introduce RexBDDs, binary decision diagrams (BDDs) that exploit reduction opportunities well beyond those of reduced ordered BDDs, zero-suppressed BDDs, and recent proposals integrating multiple reduction rules. RexBDDs also leverage (output) complement flags and (input) swap flags to potentially decrease the number of nodes by a factor of four. We define a reduced form of RexBDDs that ensures canonicity, and use a set of benchmarks to demonstrate their superior storage and runtime requirements compared to previous alternatives.	Gianfranco Ciardo (Iowa State University); Andrew S. Miner (Iowa State University); Lichuan Deng (Iowa State University); Junaid Babar (Collins Aerospacce)
1907	GreenFPGA: Evaluating FPGAs as Environmentally Sustainable Computing Solutions	Growing global concerns about climate change highlight the need for environmentally sustainable computing. The ecological impact of computing, including operational and embodied, is a key consideration. Field Programmable Gate Arrays (FPGAs) stand out as promising sustainable computing platforms due to their reconfigurability across various applications. This paper introduces GreenFPGA, a tool estimating the total carbon footprint (CFP) of FPGAs over their lifespan, considering design, manufacturing, reconfigurability (reuse), operation, disposal, and recycling. Using GreenFPGA, the paper evaluates scenarios where the ecological benefits of FPGA reconfigurability outweigh operational and embodied carbon costs, positioning FPGAs as a environmentally sustainable choice for hardware acceleration compared to Application-Specific Integrated Circuits (ASICs). Experimental results show that FPGAs have lower CFP than ASICs, particularly for multiple distinct, low-volume applications, or short application lifespans.	Chetan Choppali Sudarshan (Arizona State University); Aman Arora (Arizona State University); Vidya A. Chhabria (Arizona State University)
1911	Effective Quantum Resource Optimization via Circuit Resizing in BQSKit	In the noisy intermediate-scale quantum era, mid-circuit measurement and reset operations facilitate novel circuit optimization strategies by reducing a circuit's qubit count in a method called resizing. This paper introduces two such algorithms. The first one leverages gate-dependency rules to reduce qubit count by 61.6% or 45.3% when optimizing depth as well. Based on numerical instantiation and synthesis, the second algorithm finds resizing opportunities in previously unresizable circuits via dependency rules and other state-of-the-art tools. This resizing algorithm, implemented in BQSKit, reduces qubit count by 20.7% on average for these previously impossible-to-resize circuits.	Siyuan Niu (Lawrence Berkeley National Lab); Akel Hashim (University of California, Berkeley); Costin Iancu (Lawrence Berkeley National Lab); Wibe Albert de Jong (Lawrence Berkeley National Lab); Ed Younis (Lawrence Berkeley National Lab)
1912	EdGeo: A Physics-guided Generative AI Toolkit for Geophysical Monitoring on Edge Devices	Full-waveform inversion (FWI) plays a vital role in geoscience to explore the subsurface. It utilizes the seismic wave to image the subsurface velocity map. As the machine learning (ML) technique evolves, the data-driven approaches using ML for FWI tasks have emerged, offering enhanced accuracy and reduced computational cost compared to traditional physics-based methods. However, a common challenge in geoscience --- the unprivileged data --- severely limits ML effectiveness. The issue becomes even worse during model pruning, a step essential in geoscience due to environmental complexities. To tackle this, we introduce the EdGeo toolkit, which employs a diffusion-based model guided by physics principles to generate high-fidelity velocity maps. The toolkit uses the acoustic wave equation to generate corresponding seismic waveform data, facilitating the fine-tuning of pruned ML models. Our results demonstrate significant improvements in SSIM scores and reduction in both MAE and MSE across various pruning ratios. Notably, the ML model fine-tuned using data generated by EdGeo yields superior quality of velocity maps, especially in representing unprivileged features, outperforming other existing methods.	Junhuan Yang (George Mason University); Hanchen Wang (Los Alamos National Laboratory); Yi Sheng (George Mason University); Youzuo Lin (University of North Carolina at Chapel Hill); Lei Yang (George Mason University)
1936	HyCaMi: High-Level Synthesis for Cache Side-Channel Mitigation	Cache side-channels are a major threat to cryptographic implementations, particularly block ciphers. Traditional manual hardening methods transform block ciphers into Boolean circuits, a practice refined since the late 90s. The only existing automatic approach based on Boolean circuits achieves security but suffers from performance issues. This paper examines the use of Lookup Tables (LUTs) for automatic hardening of block ciphers against cache side-channel attacks.We present a novel method combining LUT-based synthesis with quantitative static analysis in our HyCaMi framework. Applied to seven block cipher implementations, HyCaMi shows significant improvement in efficiency, being 9.5× more efficient than previous methods, while effectively protecting against cache side-channel attacks. Additionally, for the first time, we explore balancing speed with security by adjusting LUT sizes, providing faster performance with slightly reduced leakage guarantees, suitable for scenarios where absolute security and speed must be balanced.	Heiko Mantel (Technical University of Darmstadt); Joachim Schmidt (Technical University of Darmstadt); Thomas Schneider (Technical University of Darmstadt); Maximilian Stillger (Technical University of Darmstadt); Tim Weißmantel (Technical University of Darmstadt); Hossein Yalame (Technical University of Darmstadt)
1949	E-DGCN: An Efficient Architecture Design for Accelerating Dynamic Graph Convolutional Network (DGCN) Inference	Dynamic graph neural networks (DGCNs) have been proposed to extend machine learning techniques to applications involving dynamic graphs. Typically, a DGCN model includes a graph convolutional network (GCN) followed by a recurrent neural network (RNN) to capture both spatial and temporal information. To efficiently perform distinct neural network models as well as maximize the data reuse and hardware utilization, customized hardware designs for such applications require a reconfigurable computing engine, flexible dataflow, and efficient data locality exploitation. We propose an efficient DGCN accelerator named E-DGCN. Specifically, E-DGCN includes modified Processing Elements (PEs) with a flexible interconnection design to support diverse computation patterns and various dataflows. Additionally, a lightweight vertex caching algorithm is proposed to exploit data locality, enabling E-DGCN to selectively load required vertices during DGCN inference. These implementations provide benefits in managing data computation and communication.	Yingnan Zhao (The George Washington University); Ke Wang (University of North Carolina at Charlotte); Jiaqi Yang (George Washington University); Ahmed Louri (The George Washington University)
1963	FCM: Wire Cutting For Fusion Reduction in Measurement-based Quantum Computing	Measurement-based quantum computing (MBQC) is a promising quantum computing paradigm that carries out computation through one-way measurements on entangled photon qubits. Practical photonic hardware first generates a 2D mesh of resource states with each being a small number of entangled photon qubits and then exploits fusion operations to connect resource states to scale up the computation. Given that the fusion operation is highly error-prone, it is important to reduce the number of fusions for an MBQC circuit. In this paper, we propose FCM, a fusion-aware scheme that exploits wire cutting to improve the fidelity of MBQC. By cutting a large MBQC circuit into several smaller subcircuits, FCM effectively reduces the number of fusions in each subcircuit and thus improves the computation fidelity. Given circuit cutting requires classical post-processing to combine the results of subcircuits, FCM strives to achieve the best cutting strategy under different settings. Experimental evaluation of representative benchmarks demonstrates that, when cutting a large circuit to two subcircuits, FCM reduces the maximum number of fusions of all subcircuits by 59.6% on average (up to 69.1%).	Zewei Mo (University of Pittsburgh); Yingheng Li (University of Pittsburgh); Aditya Pawar (University of Pittsburgh); Xulong Tang (University of Pittsburgh); Jun Yang (University of Pittsburgh); Youtao Zhang (University of Pittsburgh)
1968	Algorithm-Hardware Co-Design of Distribution-Aware Logarithmic-Posit Encodings for Efficient DNN Inference	Traditional Deep Neural Network (DNN) quantization methods using integer, fixed-point, or floating-point data types struggle to capture diverse DNN parameter distributions at low precision, and often require large silicon overhead and intensive quantization-aware training. In this study, we introduce Logarithmic Posits (LP), an adaptive, hardware-friendly data type inspired by posits that dynamically adapts to DNN weight/activation distributions by parameterizing LP bit fields. We also develop a novel genetic-algorithm based framework, LP Quantization (LPQ), to find optimal layer-wise LP parameters while reducing representational divergence between quantized and full-precision models through a novel global-local contrastive objective. Additionally, we design a unified mixed-precision LP accelerator (LPA) architecture comprising of processing elements (PEs) incorporating LP in the computational datapath. Our algorithm-hardware co-design demonstrates on average <1% drop in top-1 accuracy across various CNN and ViT models. It also achieves ~2x improvements in performance per unit area and 2.2x gains in energy efficiency compared to state-of-the-art quantization accelerators using different data types.	Akshat Ramachandran (Georgia Institute of Technology); Zishen Wan (Georgia Institute of Technology); Geonhwa Jeong (Georgia Institute of Technology); John Gustafson (Arizona State University); Tushar Krishna (Georgia Institute of Technology)

Submissio	Title	Abstract	Authors with Affiliations
1976	EDGE-LLM: Enabling Efficient Large Language Model Adaptation on Edge Devices via Unified Compression and Adaptive Layer Voting	Efficiently adapting Large Language Models (LLMs) on resource-constrained devices, such as edge devices, is vital for applications requiring continuous and privacy-preserving adaptation. However, existing solutions fall short due to the high memory and computational overhead associated with LLMs. To address this, we introduce an LLM tuning framework, Edge-LLM, that features three core components: (1) a unified compression method offering cost-effective layer-wise pruning ratios and quantization policies, (2) an adaptive tuning and voting scheme that selectively adjusts a subset of layers during each iteration and then adaptively combines their outputs for the final inference, thus reducing backpropagation depth and memory overhead during adaptation, and (3) a complementary search space that optimizes device workload and utilization. Experiment results demonstrate that Edge-LLM achieves efficient on-device adaptation with comparable performance with vanilla tuning methods.	Zhongzhi Yu (Georgia Institute of Technology); Zheng Wang (Georgia Institute of Technology); Yuhan Li (Georgia Institute of Technology); Ruijie Gao (Georgoa Institute of Technology); Xiaoya Zhou (University of California, Santa Barbara); Sreenidhi Reddy Bommu (Georgia Institute of Technology); Yang (Katie) Zhao (University of Minnesota, Twin Cities); Yingyan (Celine) Lin (Georgia Institute of Technology)
1978	A High Level Approach to Co-Designing 3D ICs	3D ICs promise increased logic density and reduced routing congestion over conventional monolithic 2D ICs. High level synthesis (HLS) tools promise reduced design complexity by approaching the design from a higher abstraction level and allow for more optimization flexibility. We propose improving timing closure of 3D ICs by co-designing the architecture and physical design by integrating HLS and 3D IC macro placement into the same holistic loop. On average our method is able to reduce estimated total negative slack (TNS) by 62% and 92% when compared to a traditional binding and placement technique for 2D and 3D ICs respectively.	Daniel Xing (University of Maryland); Ankur Srivastava (University of Maryland)
2007	RISC-V Instruction Set Extensions for Multi-Precision Integer Arithmetic	Arithmetic operations on multi-precision integers (MPI) are a performance-critical component of many public-key cryptosystems, including not only classical RSA and ECC, but also post-quantum isogeny-based schemes. In this paper, we analyze and compare two different MPI representations, namely full-radix versus reduced-radix, for efficient modular arithmetic implementations on 64-bit RISC-V (i.e., RV64GC). We then explore how the execution time can be further improved by designing Instruction Set Extensions (ISEs). The ISE we propose can accelerate a CSIDH-512 class group action by a factor of 1.71 compared to an ISA-only implementation on a 64-bit Rocket core. The hardware overhead introduced by our ISE is approximately 10%.	Hao Cheng (University of Luxembourg); Georgios Fotiadis (University of Luxembourg); Johann Groszschaedl (University of Luxembourg); Daniel Page (University of Bristol); Thinh H. Pham (University of Bristol); Peter Y. A. Ryan (University of Luxembourg)
2021	PPA-Relevant Clustering-Driven Placement for Large-Scale VLSI Designs	Today's place-and-route (P&R) flows are increasingly challenged by complexity and scale of modern designs. Often, heuristics must trade off between turnaround time and quality of PPA outcomes. This paper presents a clustered placement methodology that improves both turnaround time and final-routed solution quality. Our PPA-aware clustering considers timing, power and logical hierarchy during netlist clustering, effectively reducing problem size and accelerating global placement runtime while improving post-route PPA metrics. Additionally, our machine learning (ML)-accelerated virtualized P&R (V-P&R) methodology predicts the best cluster shapes (i.e., aspect ratios and utilizations) to use in P&R of the clustered netlist. With the open-source OpenROAD tool, our methods achieve up to 47% (average: 36%) global placement runtime improvement with similar half-perimeter wirelength (HPWL) and 90% (29%) improvement in post-route total negative slack (TNS). With the commercial Cadence Innovus tool, our methods achieve up to 1.68% (0.00%) improvement in power and 94% (44%) improvement in TNS.	Andrew Kahng (UCSD); Seokhyeong Kang (Pohang University of Science and Technology); Sayak Kundu (University of California, San Diego); Kyungjun Min (Pohang University of Science and Technology); Seonghyeon Park (POSTECH); Bodhisatta Pramanik (University of California San Diego)
2033	OPTIMA: Design-Space Exploration of Discharge-Based In-SRAM Computing: Quantifying Energy-Accuracy Trade-offs	In-SRAM computing promises energy efficiency, but circuit nonlinearities and PVT variations pose major challenges in designing robust accelerators. To address this, we introduce OPTIMA, a modeling framework that aids in analyzing bit-line discharge and power consumption in 6T-SRAM-based accelerators. It provides insights into limiting factors and enables fast design-space exploration of circuit configurations. Leveraging OPTIMA for in-SRAM multiplications exhibits ~100× simulation speed-up while maintaining an average modeling error of 0.56mV. Exploration yields an optimized multiplier with 1.02pJ energy consumption per 4-bit operation and classification accuracies of 71.91% (top-1) and 90.72% (top-5) for ImageNet and 92.57% for CIFAR-10 datasets respectively when applied in quantized DNNs.	Saeed Seyedfaraji (Institute of Computer Technology, Technische Universität Wien (TU Wien)); Severin Jäger (TU Wien); Salar Shakibhamedan (TU Wien); Asad Aftab (Technical University (TU) Vienna); Semeen Rehman (TU Wien)
2041	TraiNDSim: A Simulation Framework for Comprehensive Performance Evaluation of Neuromorphic Devices for On-Chip Training	The advancement of neuromorphic devices (NDs) for processing deep neural networks has narrowed the accuracy gap with software-trained models. To accurately assess ND performance, reliable simulation frameworks for on-chip training are crucial. We critically evaluated existing frameworks, identifying key defects in the training process. Consequently, we introduce TraiNDSim, a novel framework that addresses these issues. In refining the training process, we propose an advanced conductance normalization strategy called layer-wise normalization, which limits the weight range by taking the initial weight distribution into account. Additionally, our framework integrates three conductance models, notably refining one of the conventional models to depend solely on nonlinearity. Moreover, it features a bi-directional weight representation method with a unique conductance compensation technique. Our comprehensive analysis using TraiNDSim demonstrates its effectiveness in accurately reflecting the impact of ND parameters on training, promising more precise device performance evaluations. Our framework is available at https://anonymous.4open.science/r/TraiNDSim-FC25.	Donghyeok Heo (Sungkyunkwan University); Hyeonsu Bang (Sungkyunkwan University); Jong Hwan Ko (Sungkyunkwan University (SKKU))
2046	Partitioned Scheduling and Parallelism Assignment for Real-Time DNN Inference Tasks on Multi-TPU	Pipelining on Edge Tensor Processing Units (TPUs) optimizes the deep neural network (DNN) inference by breaking it down into multiple stages processed concurrently on multiple accelerators. Such DNN inference tasks can be modeled as sporadic non-preemptive gangs with execution times that vary with their parallelism levels. This paper proposes a strict partitioning strategy for deploying DNN inferences in real-time systems. The strategy determines tasks' parallelism levels and assigns tasks to disjoint processor partitions. Configuring the tasks in the same partition with a uniform parallelism level avoids scheduling anomalies and enables schedulability verification using well-understood uniprocessor analyses. Evaluation using real-world Edge TPU benchmarks demonstrated that the proposed method achieves a higher schedulability ratio than state-of-the-art gang scheduling techniques.	Binqi Sun (Technical University of Munich); Tomasz Kloda (LAAS-CNRS, Université de Toulouse, CNRS, INSA); Chu-ge Wu (Beijing Institute of Technology); Marco Caccamo (TUM, Germany)
2055	FHE-CGRA: Enable Efficient Acceleration of Fully Homomorphic Encryption on CGRAs	Fully Homomorphic Encryption (FHE) is a privacy-preserving technique that allows computation directly on encrypted data. In this work, we investigate execution the of FHE machine learning (ML) applications. We show that the runtime hardware reconfigurability of the underlying execution units of homomorphic operations is highly desirable for efficient hardware resource utilization. Based on the observation, we propose FHE-CGRA, a coarse-grained reconfigurable architecture (CGRA) acceleration framework for end-to-end homomorphic applications. The experiment shows that FHE-CGRA achieves up to 8.15x speedup against a conventional CGRA for accelerating FHE-encrypted convolution neural network (FHE-CNN) models, and 16.48x power efficiency w.r.t. the state-of-the-art FPGA.	Miaomiao Jiang (Shandong University/Quan Cheng Laboratory); Yilan Zhu (Shandong University); Honghui You (Shandong University); Cheng Tan (Google); Zhaoying Li (National University of Singapore); Jiming Xu (Ant Group); Lei Ju (School of Cyber Science and Technology, Shandong University)

Submissio	Title	Abstract	Authors with Affiliations
2060	4-Transistor Ternary Content Addressable Memory Cell Design using Stacked Hybrid IGZO/Si Transistors	In this paper, we propose a 4T-based paired orthogonally stacked transistors for random access memory (POST-RAM) cell structure and also suggest ternary content addressable memory (TCAM) applications. POST-RAM cells feature vertically stacked read and write transistors, maximizing area efficiency by utilizing only two transistors' space. %POST-RAM cells have read and write transistors stacked vertically, maximizing area efficiency by using the area of only two transistors. POST-RAM employs InGaZnO (IGZO) channels for write transistors and single crystal silicon channels for read transistors, which results in both extremely long memory retention and fast reading performance. A comprehensive 3D-TCAD simulation is conducted to validate the procedural design of the proposed device structure. Furthermore, we introduced a self-clamped searching scheme (SC2S) designed to enhance the efficiency of TCAM operations. The results conclusively demonstrate that operating a TCAM based on the proposed POST-RAM architecture can lead to a 20$\%$ improvement in energy-delay product (EDP). Notably, the delay performance can be enhanced by up to 40$\%$ when compared to a 16T SRAM-based TCAM. Additionally, the proposed scheme enables a more than sixfold reduction in cell area, demonstrating an efficient use of space.	Munhyeon Kim (Seoul National University); Jae-Joon Kim (Seoul National University)
2070	Ev-Edge: Efficient Execution of Event-based Vision Algorithms on Commodity Edge Platforms	Event-based vision sensors have demonstrated great promise in applications like autonomous UAVs. However, deploying event-based algorithms on heterogeneous edge platforms is inefficient due to mismatch between irregular nature of event streams and diverse characteristics of algorithms (mixture of spiking and conventional neural networks) on one hand and the underlying hardware platform on the other. We introduce Ev-Edge, a framework that contains three key optimizations to boost performance of event-based vision systems on edge platforms. Ev-Edge achieves 1.28x-2.05x latency and 1.23x-2.15x energy improvements over an all-GPU implementation and 1.42x-1.98x latency improvements over round-robin scheduling methods in multi-task execution scenarios with negligible accuracy loss on the NVIDIA Jetson Xavier platform.	Shrihari Sridharan (Purdue University); Surya Selvam (Purdue University); Kaushik Roy (Purdue University); Anand Raghunathan (Purdue University)
2078	How accurately can soft error impact be estimated in black-box/white-box cases? -- a case study with an edge AI SoC --	Artificial intelligence (AI) edge devices often feature numerous storage units and sequential logic circuits, making them vulnerable to soft errors. For reliable and critical edge AI applications, assessing System-on-Chip (SoC) reliability in advance is essential. Here, there are two cases: a self-designed SoC (white-box), or a commercial off-the-shelf (COTS) chip (black-box). This study uses alpha particle irradiation results on our 22nm AI SoC as a golden reference to estimate soft error impacts, injecting faults across the entire chip in the white-box case and into the accessible memory and registers in the black-box case. The results demonstrate a high degree of consistency between the white-box case and golden reference, meaning that pre-silicon reliability assessment is feasible. As for the black-box case, the proportion of memory in the SoC remains unchanged and is still significantly larger than that of registers, and hence the simulation results between black-box and white-box are not substantially different.	Quan CHENG (Kyoto University); Qiufeng Li (Southern University of Science and Technology); Longyang Lin (Southern University of Science and Technology); Wang Liao (Kochi University of Technology); Liuyao Dai (University of California, Merced); Hao Yu (Southern University of Science and Technology); Masanori Hashimoto (Kyoto University)