19TH INTERNATIONAL SYMPOSIUM ON EMBEDDED MULTICORE/MANY-CORE SYSTEMS-ON-CHIP

MCSoC-2026 · Program with Abstracts

December 14–17, 2026 · Shanghai Jiao Tong University, Shanghai

TIME ZONE
All times UTC+8
(Shanghai local time)

PAPER TALKS
15 min each
10 min talk + 5 min Q&A

CO-LOCATED SUMMITS — NeuroCore Summit · Mon, Room A — EmbodiCore Summit · Wed, Room B — *mcsoc-forum.org*

Each paper is allotted 15 minutes (10-minute talk and 5 minutes of Q&A). The session chair and co-chair are listed under each session title, and the chair of each keynote is listed with the talk. Click a session title to go to its papers. The presenter of each paper is underlined.

Online participation (Zoom): Room A – Meeting ID **839 8259 2902** · Room B – Meeting ID **865 3881 0892**. The passcode and direct links are sent by email to registered authors and session chairs.

Time (China, UTC+8)	Room A	Room B
MONDAY, DECEMBER 14, 2026		
13:00–13:05	*O1: Kick-off Session*	
13:05–14:05	*K1: Keynote 1 – Hai Jin (Huazhong University of Science and Technology): A Unified Approach of Graph Computing* Session Chair: Tee Hui Teo (Singapore University of Technology and Design)	
14:10–15:10	**A1:** NeuroCore Summit I: Neuromorphic Circuits and Low-Power Devices	**B1:** S-Session 14: Advances in Intelligent Systems and Learning Technologies – I
15:10–16:10	**A2:** NeuroCore Summit II: Spiking Neural Networks and On-Chip Learning	**B2:** S-Session 14: Advances in Intelligent Systems and Learning Technologies – II
16:10–16:20	*Coffee/Tea Break*	
16:20–17:20	**A3:** NeuroCore Summit III: Brain-Inspired Algorithm–Hardware Co-Design and Applications	**B3:** S-Session 25: Governance-Aware Agentic AI for Systems-of-Systems
17:20–18:20	**A4:** NeuroCore Summit IV: Neuromorphic Accelerators and System Architectures	**B4:** S-Session 18: Harnessing AI Tools for Transforming Education
TUESDAY, DECEMBER 15, 2026		
09:00–09:20	*O2: Opening Session*	
09:20–10:20	*K2: Keynote 2 – Mingle Sun (Beijing SemiDrive Semiconductor Technology Co., Ltd.): Next-Generation AI SoCs for Automotive Applications and Embodied Intelligence: Meeting High-Performance, Multi-Task and Real-Time Requirements* Session Chair: Yongfu Li (Shanghai Jiao Tong University)	
10:20–10:40	*Coffee/Tea Break*	
10:40–12:10	**A5:** Hardware Acceleration of AI on Embedded Edge SoCs – I	**B5:** S-Session 1: Performance Optimization and Auto-Tuning of Software on Multicore/Manycore Systems (POAT2026)

Time (China, UTC+8)	Room A	Room B
12:10–13:30	*Lunch*	
13:30–14:30	*K3: Keynote 3 – Yan Xie (Li Auto): The Next Frontier of AI Inference: Why Dataflow is the Future of Intelligent Vehicles* Session Chair: Xinfei Guo (Shanghai Jiao Tong University)	
14:40–15:40	**A6:** Architectural Innovations in Embedded Multicore/Manycore SoCs – I	**B6:** Design Automation and Low-Power Techniques for Embedded Manycore SoCs
15:40–16:00	*Coffee/Tea Break*	
16:00–17:00	**A7:** Architectural Innovations in Embedded Multicore/Manycore SoCs – II	**B7:** Multicore SoCs for Cyber-Physical and Autonomous Systems – II
17:00–18:15	**A8:** Hardware Acceleration of AI on Embedded Edge SoCs – II	**B8:** Application Domains for Embedded Multicore/Manycore SoCs
WEDNESDAY, DECEMBER 16, 2026		
09:00–10:00	*K4: Keynote 4 – Mohamad Sawan (Westlake University): Brain-inspired SoC for Brain-Computer Interfaces and Embodied AI Applications* Session Chair: Zhishang Wang (The University of Aizu)	
10:00–10:15	*Coffee/Tea Break*	
10:15–11:15	**A9:** Design Methodologies for Embedded Multicore/Manycore SoCs – II	**B9:** S-Session 24: Human-Centric AI and Intelligent Systems for Sustainable Industry – I
11:15–12:30	**A10:** Hardware Acceleration of AI and Neuromorphic Edge Systems	**B10:** S-Session 24: Human-Centric AI and Intelligent Systems for Sustainable Industry – II
12:30–13:30	*Lunch*	
13:30–14:30	**A11:** Testing, Security, and Trust in Embedded Multicore/Manycore Systems – I	**B11:** Real-Time Systems for Embedded Multicore/Manycore SoCs
14:30–15:30	**A12:** Testing, Security, and Trust in Embedded Multicore/Manycore Systems – II	**B12:** Multicore SoCs for Cyber-Physical and Autonomous Systems – I
15:30–15:50	*Coffee/Tea Break*	
15:50–16:50	**A13:** S-Session 2: Emerging Machine Learning and Deep Learning Models: Theory and Applications	**B13:** EmbodiCore Global Summit on Embodied AI, Collaborative Humanoid, and Distributed Robotic Computing
18:30–21:30	*Dinner / Banquet*	
THURSDAY, DECEMBER 17, 2026		
09:00–10:00	**A14:** Testing, Security, and Trust in Embedded Multicore/Manycore Systems – III	**B14:** Machine Learning for Energy-Efficient and Reliable Manycore Systems
10:00–10:20	*Coffee/Tea Break*	
10:20–11:20	**A15:** Hardware Acceleration of AI on Embedded Edge SoCs – III	**B15:** Heterogeneous Integration and Advanced On-Chip Interconnects
11:20–12:20	**A16:** Design Methodologies for Embedded Multicore/Manycore SoCs – I	**B16:** Hardware Acceleration of AI on Embedded Edge SoCs – IV
12:20–12:40	*C: Closing Session*	

Sessions and Papers

Monday, December 14, 2026

A1: NeuroCore Summit I: Neuromorphic Circuits and Low-Power Devices

14:10–15:10 · Room A · Chair: Elishai Ezra Tsur (The Open University of Israel) · Co-chair: Pavan Nutalapati (Oracle Corp)

14:10	#128 **Beyond Static Power: Asynchronous vs Clock-Gated Integrate-and-Fire Layers in a Standard-Cell Flow** Zhanbo Shen (Eindhoven University of Technology); Henk Corporaal (Technical University Eindhoven); Sander Stuijk (Eindhoven University of Technology); Federico Corradi (TU Eindhoven) Spiking Neural Networks (SNNs) aim to achieve high energy efficiency through sparse, event-driven computation. However, algorithmic-level evaluations often obscure underlying physical hardware costs, such as wire delays and routing overhead, dictated by the architecture and whether the system uses a synchronous or asynchronous design. To address this, we present a post-layout comparison of optimized SNN architectures exploiting three design styles: bundled-data (BD) asynchronous, quasi-delay insensitive (QDI) asynchronous, and clock-gated synchronous (SYNC), all in a single standard-cell design flow using IHP sg13g2 130nm technology. The architectures implement integrate-and-fire neurons organized in a single layer and are tested under two distinct workloads: index-sparse and broadcast. The results demonstrate that QDI provides lower latency at the available checkpoints, but its area and power costs are substantially higher than those of BD and SYNC. BD significantly outperforms the clock-gated SYNC baseline in sparse workload, achieving a smaller area 1.12 - 1.42×, 1.8 - 18.8× lower idle power, 1.19 - 1.60× lower dynamic energy per synaptic operation (SOP), and up to 1.86× shorter mean event-response latency. Under the boradcast workload, the latency trade-offs become size-dependent; BD achieves lower first-fire latency at larger network sizes, whereas SYNC produces shorter burst-drain intervals. Ultimately, achieving the promised efficiency of SNNs requires workload-aware architecture choices, rather than one-size-fits-all.
14:25	#180 **Aging Analysis of CMOS Synaptic Circuits with Simplified Leaky Integrate-and-Fire Neurons** Jani Babu Shaik (Shanghai jiao tong university); Siona Menezes Picardo (Middlesex University Dubai); Fan Hu and Xinyue Wu (Shanghai Jiao Tong University); Sonal Singhal (Shiv Nadar University); Xinfei Guo (Shanghai Jiao Tong University) CMOS scaling exacerbates transistor reliability issues such as Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI), degrading analog neuromorphic circuits over time. This paper systematically analyzes aging effects on four synaptic topologies, Linear Charge-Discharge (LCD), Current Mirror Integrator (CMI), Log-Domain Integrator (LDI), and Differential Pair Integrator (DPI), paired with conventional and reliability-aware Simplified Leaky Integrate-and-Fire (SLIF) neurons. Pseu-dorandom presynaptic spike trains are used to quantify inter-spike interval (ISI) deviations, spike-count errors, and power consumption under 10-year NBTI/HCI stress. Results reveal that LCD topologies exhibit catastrophic temporal instability. With the reliability-aware SLIF effectively reduces threshold voltage shifts in the critical MOSFET reflecting the lesser deviations compared to conventional after aging. Among all combinations, CMIrel achieves superior temporal stability with minimal power degradation. While for rate coding, the CMIrel achieves lowest ARIrate quantifying with lower spike count errors and power consumption. The quantitative trade-off ARI framework identifies CM Irel as the optimal solution for temporal encoding, while LDIrel is optimal for applications constrained by rate encoding. This work establishes a design methodology for aging-resilient neuromorphic encoders.
14:40	#231 **From Python to Layout: The Spyqe Framework for Generating Asynchronous Neuromorphic Circuits** Zhanbo Shen and Sander Stuijk (Eindhoven University of Technology); Federico Corradi (TU Eindhoven) Event-driven Spiking Neural Networks (SNNs) naturally align with asynchronous circuits, which process sparse spike streams on demand without global clock overhead. Although high-level Python frameworks effectively capture SNN models, they abstract away the control, storage, scheduling, and explicit timing required for physical silicon implementation. To bridge this gap, we present Spyqe: a framework that compiles a supported subset of Python SNN models into bundled-data register-transfer-level (RTL) circuits, automating synthesis, place-and-route, and matched-delay insertion. Generic SNN models are mapped into flattened cell netlists, whereas specialized leaky integrate-and-fire (LIF) layers are compiled into SRAM-backed, time-multiplexed architectures with shared asynchronous processing units. We validated the complete ASIC flow in IHP 130 nm and GF 22 nm Fully Depleted Silicon on Insulator (FDSOI) technologies. Standard Delay Format (SDF) annotated post-layout simulations demonstrate bit-exact numerical agreement with the Python reference. Spyqe enables the rapid and intuitive design of asynchronous neuromorphic circuits directly from high-level Python model descriptions.
14:55	#240 **Resilient On-Chip STDP Learning Under Memory Voltage Scaling for Ultra-Low-Power Neural Network Processors** Subbaiah Ravi Hariprakash and Yuga Hanyu (University of Aizu); Abderazek Ben Abdallah (The University of Aizu); Zhishang Wang and Khanh N. Dang (University of Aizu) Edge AI with on-chip learning based on Spiking Neural Networks (SNNs) offers a promising solution by enabling energy-efficient, adaptive intelligence while minimizing data movement. To further reduce power consumption, aggressively undervolted on-chip memory can be employed, but operating near-threshold voltages introduces permanent hardware faults, primarily Stuck-at-0 (SA0) and Stuck-at-1 (SA1). While the network largely absorbs SA0 faults, SA1 faults force neurons to fire continuously, triggering runaway lateral inhibition that permanently suppresses healthy neurons and causes catastrophic collapse of Spike-Timing-Dependent Plasticity (STDP) learning. To address this challenge, we propose STDP-DIP (Spike- Timing-Dependent Plasticity with Diagnostic Isolation Pruning), a fault-aware on-chip learning framework that integrates fault injection with STDP learning. STDP-DIP identifies SA1 faults during learning and structurally severs their inhibitory outputs, preventing pathological inhibition from propagating through the network while preserving normal neuronal dynamics and synaptic plasticity. Experimental results demonstrate that STDP- DIP restores stable STDP learning under severe near-threshold memory fault conditions, achieving >91% accuracy at 20% bit error rates, improving performance by over 80% compared with conventional uncorrected approaches while maintaining the energy benefits of near-threshold operation.

B1: S-Session 14: Advances in Intelligent Systems and Learning Technologies – I

14:10–15:10 · Room B · Chair: Mohamed Badawy (Infineon Technologies AG & Technical University of Munich) · Co-chair: Sandeep Sonawane (Nutanix Inc.)

14:10	#131 **Residual Deep Neural Network for Explainable Student Academic Performance Classification** Yousuf Nasser Al Husaini and Mohammed Abdulla Salim Al Husaini (Arab Open University); Mohammad Abrar (Arab Open University) Early identification of students who may require academic support is an important problem in educational data mining and learning analytics. This paper presents a residual deep neural network (Residual DNN) for multiclass student academic performance classification using academic, behavioral, family, and institutional factors. The target variable is derived from examination scores and grouped into low, medium, and high performance classes using data-driven thresholds. The proposed architecture integrates fully connected representation learning, batch normalization, dropout regularization, and residual shortcut connections to improve optimization stability and preserve useful low-level representations across hidden layers. The model was evaluated on a public student performance factors dataset containing 6,607 student records and 19 predictive attributes. Experimental results show that the proposed Residual DNN achieved 94.86% test accuracy, 94.87% macro F1- score, 94.74% precision, and 95.03% recall, outperforming the other evaluated neural configurations in the same experimental setting. Explainability analysis further shows that attendance, hours studied, previous scores, parental involvement, access to learning resources, and tutoring sessions are the most influential predictors of academic performance. The findings demonstrate that residual neural learning can provide an accurate and interpretable decision-support mechanism for data-driven academic monitoring and timely student intervention.
14:25	#148 **Interpretable Hidden-State Modeling of Programming Learning Progression** Md. Shahajada Mia (The University of Aizu); Yutaka Watanobe (University of Aizu); Md Mostafizer Rahman (University of Notre Dame); Md Faizul Ibne Amin and Daniel Muepu Mukendi (The University of Aizu) Online Judge (OJ) systems record large volumes of learner submissions generated through repeated programming practice, often involving several incorrect attempts before reaching a correct solution. However, relying on the binary final outcome can hide the progression a learner makes before solving a programming problem. To capture this hidden progression, this paper presents an interpretable hidden-state modeling approach for analyzing learner progression in OJ-based programming practice. Using ALDS1 submission records from the Aizu Online Judge platform, each submission is mapped to one of four progression levels: Low, Medium, High, and Solved. A Hidden Markov Model (HMM) is then fitted to learner problem sequences to reveal learner latent progression behavior, with an observed Markov chain serving as a baseline. The six-state HMM uncovers meaningful low-progress, developing, near-solution, and solution states, whereas the Markov chain shows strong persistence in the low-progress level and non-monotonic movement from high to low progress. A Mann Whitney U test comparison demonstrates that solved sequences exhibit substantially greater hidden-state transition activity than unsolved ones (2.155 vs. 0.679; Cliff's δ=0.784), while unsolved sequences remain more concentrated in the low-progress state. Overall, the approach offers interpretable educational insights to identify stuck and near-solution learners who fail to complete a problem, thereby supporting adaptive instructor feedback.
14:40	#307 **Score Agreement and Confidence Calibration in Grounded LLM Short-Answer Scoring** Anoop Yadav, Muhammad Jawad Arshad and Yutaka Watanobe (University of Aizu) Traditional automated tutoring systems estimate student mastery from structured responses and therefore omit evidence contained in open-ended explanations. Large language models (LLMs) enable free-text assessment, but their scores and verbalized confidence may be inaccurate or overconfident. We evaluate a grounded LLM scoring pipeline on 1,000 answers from the Mohler ShortAnswerGrading dataset. Providing the question and instructor reference answer increases correlation with the exact normalized human consensus from r = 0.271 to r = 0.729 (Meng Z = 15.51, p < .001), while raw expected calibration error (ECE) remains 0.195-0.373 when scores are counted as correct within 1.0 or 0.5 point of consensus. A matched GPT- 4o replication (n = 999) reproduces the grounding gain and raw miscalibration. In a post-hoc analysis with 1,000 questiongrouped splits, Platt scaling reduces median ECE to 0.087-0.100 on unseen questions, while isotonic regression reduces it to 0.075- 0.078 but is less stable under log loss. These results establish an agreement-calibration divergence: grounding improves agreement with human consensus but does not calibrate verbalized confidence.
14:55	#216 **Codiflect: An Intelligent Tutoring System for Idiomatic Python with RAG-Based LLM Tutorials and Deterministic Selection and Verification** Jonas Sobotta and Li Ke (University of Aizu); Mohamed Hamada (UoA); Yutaka Watanobe (University of Aizu) Developers who move from Java to Python often write code that works but is not idiomatic Python. We present Codiflect, an Intelligent Tutoring System (ITS) for VS Code that teaches idiomatic Python from the developer's own code. Static analysis finds where the code uses one of eight idioms or misses a chance to use it, and a Bayesian learner model estimates proficiency and mastery per idiom. A Retrieval-Augmented Generation (RAG) pipeline lets a Large Language Model (LLM) write a tutorial on the developer's code. Deterministic code selects the material by proficiency and builds and grades exercises; the LLM only writes and reviews the tutorial text. We validated each component technically, without users. In simulation, at most 4.5% of learners with an idiomatic rate of 70%, below the mastery threshold of 75%, are declared to have mastered an idiom. On Python code by Java programmers, the detectors reach a precision of 0.98 for idiomatic uses and 0.81 for missed idioms. Exercise grading rejected all 2,610 generated wrong answers, and a code check found that 6.1% of checkable tutorials change the behavior of the learner's code. Across three LLMs, stating the learner's proficiency in the prompt lengthens the explanation by a median of 13% to 35% and hardly changes which concepts it covers. Finally, an LLM judge can pass a common qualification test without telling apart the conditions it is meant to compare.

A2: NeuroCore Summit II: Spiking Neural Networks and On-Chip Learning

15:10–16:10 · Room A · Chair: Elishai Ezra Tsur (The Open University of Israel) · Co-chair: Vijayakumar Venganti (Cisco Systems, Inc.)

15:10	#71 **MSTDP-Trained Izhikevich Spiking Neural Networks for Reinforcement Learning** Tuan-Hung Nguyen, Khanh Ngoc Nguyen and Nam Hai Nguyen (University of Transport and Communications); Quang Huy Dong (University of Transportation and Communication); Truong Cong Tan

	(Trường Đại Học Giao Thông Vận Tải); Truong Duc Phuong (University of Transport and Communications); Khanh N. Dang (University of Aizu); Duy-Anh Nguyen (University of Transport and Communications) Spiking Neural Networks (SNNs) provide a biologically inspired and energy-efficient alternative to traditional Artificial Neural Networks (ANNs), but training SNNs remains challenging. This paper presents a Reinforcement Learning (RL) framework for training SNNs in the Atari Breakout environment, evaluating five Izhikevich neuron types (RS, IB, CH, FS, and LTS) together with a Leaky Integrate-and-Fire (LIF) baseline. Using Modulated Spike-Timing-Dependent Plasticity (MSTDP), the proposed system learns directly from pixel-based interaction without labeled data. Experimental results show that Izhikevich dynamics strongly affect RL performance. In particular, the best model of Izhikevich achieves the best overall test reward (1.167 +/- 0.168) with the low failure rate (36.7%), while the LIF model only have overall test reward of 1.047 +/- 0.061 and the failure rate of 41.3%. These results suggest that richer Izhikevich firing dynamics can improve policy learning in bio-inspired reinforcement learning systems. The source code is available at https://github.com/klab-aizu/Izhikevich_RL.
15:25	#139 **Event-Driven Language Models with Sparse Neural Activity for Neuromorphic Hardware** Simon Richter (University of Southern Denmark - SDU); Ruhai Lin (University of California, Santa Cruz); Jason Yik (Harvard University); Taylor Kergan and Rui-Jie Zhu (University of California, Santa Cruz); Farshad Moradi (University of Southern Denmark); Jason K. Eshraghian (University of California, Santa Cruz) Inference with transformer-based large language models (LLMs) is often limited by the memory-bound KV cache and quadratic attention cost. State-space models (SSMs) mitigate this through linear attention and fixed-size recurrent states, but their large dense linear projections remain computationally expensive even after quantization. We introduce a method that induces sparse neural activity in heavily quantized linear-attention models with minimal performance loss. Activations below a per-projection trainable threshold (+-Δ) are nullified while preserving crucial outliers, achieving comparable performance to dense models with up to 4x fewer effective arithmetic operations. Targeting a multi-core, multi-chip neuromorphic platform, where event-driven execution converts unstructured sparsity into throughput at both the compute and communication levels, a capability GPU architectures fundamentally lack, we project up to37x higher throughput and 16x lower power versus edge GPU inference of a comparable transformer-based model, and up to5.4x improvements over the non-sparsified baseline. These results position sparse, quantized linear-attention models as a natural fit for deploying LLMs on event-driven multi-core platforms.
15:40	#160 **IzhiGAN: A Calibration Aware ANN to SNN Framework for Efficient Image Generation** Quang-Thai Pham, Thanh-Dat Nguyen, Van-Vu Luyen and Duy-Anh Nguyen (University of Transport and Communications); Yuga Hanyu and Khanh N. Dang (University of Aizu) Spiking Neural Networks (SNNs) have emerged as an energy-efficient computing paradigm for deploying computationally intensive Generative Adversarial Networks (GANs). However, direct Artificial Neural Network-to-Spiking Neural Network (ANN-to-SNN) conversion for generative models remains challenging due to the mismatch between continuous ANN activations and spike-based representations, which can introduce approximation errors that accumulate across successive layers. To address this issue, this paper proposes an offline dual-stage calibration framework for converting a pre-trained ANN generator into an Izhikevich-based SpikingGAN. The framework combines neuron-wise affine current mapping with iterative layer-wise linear compensation to better align ANN activations with spiking responses and reduce conversion errors. The resulting calibration parameters are folded into the network weights and biases prior to deployment, eliminating the need for additional hidden-layer compensation during inference. Experiments on the MNIST dataset show that the converted SpikingGAN preserves the global structural characteristics of the generated digits, achieving an average Peak Signal-to-Noise Ratio (PSNR) of 23.05 dB, a Structural Similarity Index Measure (SSIM) of 0.5540, and a Pearson correlation coefficient of 0.7883. These results demonstrate that the proposed framework effectively improves ANN-to-SNN conversion fidelity while introducing no additional hidden-layer computational overhead during inference.
15:55	#272 **A Spiking Trajectory Memory for Predicting Repetitive Motion from Event Cameras** Sagi Taizi and Elishai Ezra Tsur (The Open University of Israel) The brain is an exceptional prediction machine, capable of anticipating repetitive motion within a few cycles, yet trackers built from its closest artificial counterparts, event cameras and spiking neural networks (SNNs), mostly react to instantenues motion. Here we propose a spiking framework that learns a repeating trajectory online from event-camera input, predicts the target's position a short horizon ahead, and quickly flags when the motion departs from the learned path. The pipeline is composed of two biologically inspired stages. A spiking localiser, trained on simulated data, reads the target's position from 5 ms event frames as place-cell activity. A spiking memory then learns the path online: an adaptive-frequency oscillator serves as a clock that locks on to the motion's rhythm, a map over a ring of phase cells holds the path, and the target's distance from a slowly updated copy of that map flags deviations. To the best of our knowledge, this is the first spiking memory of a repeating trajectory from event-camera input. On real event-camera recordings of three types of repetitive motion, led by a swinging pendulum, the pipeline runs in real time on a conventional CPU, predicts the pendulum 50 ms ahead as accurately as classical periodic filters, and detects all recorded deviations with low latency and no false alarms.

B2: S-Session 14: Advances in Intelligent Systems and Learning Technologies – II

15:10–16:10 · Room B · Chair: Yousuf Nasser Al Husaini (Arab Open University) · Co-chair: Somesh Rahul (Independent Researcher)

15:10	#264 **Digital-Twin-Driven Experiential Learning for Cyber-Physical Security: Teaching Zero Trust Through Real-Time Industrial Process Consequences** Abdul Khadir Mahamed, Ramesh Subramani and Sai Pavan Kumar Mandapaka (New York University) Industrial cybersecurity is usually taught as separate network, access-control, and process-control topics, so learners rarely see how one cyber decision becomes a physical consequence. This paper presents a digital-twin-driven experiential-learning framework in which learners assess a contextual access request with an explainable six-factor risk score, choose allow, step-up, or deny, and immediately observe the effect on a simulated heated mixing tank protected by an independent safety instrumented function inside an IEC 62443 zone-and-conduit laboratory. Four incident-inspired scenarios cover compromised credentials, unsafe engineering action, ransomware, and an unvetted vendor firmware push. Technical validation shows that: (i) consequences depend measurably on response delay, with production loss kept below one minute only for responses within 58 to 247 s and safety trips in three scenarios beyond 128 to 247 s; (ii) across 400 seeded requests, the Zero-Trust policy executed 16.5% of malicious requests versus 71.0% for perimeter trust and

	100% for implicit trust, reducing simulated production loss by 86.2% at a 300 s delay, a ranking that held across 50 further seeds and stealthier attackers; (iii) runs are bit-identical, and 90.4% of decisions are unchanged under ±20% weight perturbation; and (iv) a vectorized twin processes about 4,000 forty-minute episodes per second on one CPU core and scales across cores. No student data are reported; learning effectiveness is left to a future controlled study.
15:25	#332 **Code Refactoring with LLM: A Comprehensive Evaluation Across Few-Shot Settings** Md. Raihan Tapader (Dhaka University of Engineering & Technology, Gazipur); Md Mostafizer Rahman (University of Notre Dame); Ariful Islam Shiplu, Md Faizul Ibne Amin and Md. Shahajada Mia (The University of Aizu); Yutaka Watanobe (University of Aizu) Programming practice has shifted from focusing solely on raw functionality to writing simple, clear, efficient, and maintainable code that is easy to understand and maintain. Code refactoring plays a central role in this shift, improving the structural organization and runtime performance of code. However, existing refactoring methods do not scale well across programming languages and coding styles, as they rely on hand-crafted transformation rules. To overcome this limitation, we propose a prompt-based framework for multilingual code refactoring with few-shot learning. Our method uses a prompt driven Large Language Model (LLM) to refactor code in five languages, including C, C++, C#, Python, and Java. It evaluates the quality of the resulting transformations using both empirical metrics (compilability, correctness, edit distance, similarity, source lines of code, token count, character count, and cyclomatic complexity) and human assessment. Experimental results show that refactored Java code achieves the highest overall correctness, reaching 99.99% in the 10-shot setting at a sampling temperature of 0.3. Under the same configuration, refactored Java code also attains the highest average compilability (94.78%) relative to the original source code.
15:40	#169 **An LLM-Driven Multi-Agent Intelligent System for Autonomous Smart Home Management** Akila Siriweera (University of Aizu); Janani Rangila (Kotelawala Defence University) Smart homes integrate heterogeneous Internet of Things (IoT) devices, including sensors, actuators, appliances, and safety equipment, which continuously emit telemetry and change state. This dynamic environment requires autonomous coordination across competing objectives such as safety, security, comfort, and energy efficiency, surpassing the capabilities of a single monolithic model. This paper introduces a large language model (LLM)-driven multi-agent intelligent system that coordinates a set of role-specialized agents for autonomous smart-home device management. Each agent, supported by a large language model through a pluggable multi-provider model router, independently analyzes current device telemetry and issues a typed command with an associated confidence score. Commands targeting the same device are reconciled through a prioritized arbitration cascade, where safety-critical decisions take unconditional precedence. An ensemble of machine-learning anomaly-detection models identifies abnormal device readings and proposes corrective actions. Additionally, a natural-language interface enables residents to issue commands directly. Agent decisions are recorded as Ed25519-signed transactions on an in-process blockchain with adaptive proof-of-work, while batched off-chain telemetry is periodically anchored using Merkle roots. An end-to-end prototype has been implemented, and preliminary evaluation on simulated smart-home scenarios across multiple consensus configurations under both baseline and threat conditions demonstrates the feasibility of the proposed system.
15:55	#168 **Automated IELTS Diagnostic Assessment for L1 Sinhala Speakers via Retrieval-Grounded Dual-Rater LLM Scoring with Official Band Descriptors** Akila Siriweera (University of Aizu); Vishmika Devindi (KDU University) English proficiency is widely certified through standardized assessments such as the International English Language Testing System (IELTS). IELTS exams are often high stakes, with results influencing eligibility to apply to study at English-medium tertiary institutions. Currently, native Sinhala speakers in Sri Lanka are having difficulty preparing for the IELTS examinations due to a scarcity of preparatory resources. Preparation for IELTS benefits from knowing a learner's current status in each of the exam's four modules to facilitate targeted instruction on learner weaknesses. This research aims to propose an IELTS diagnostic assessment tool for the Sinhalese community. The writing and speaking modules are often scored without reference to the official band descriptors, although the descriptors define how examiners award bands. The reading and listening modules are often scored without answer-key matching and band conversion. The diagnostic test proposed in this paper covers all four modules in a single session. The Writing and Speaking modules are scored by two independent large language model raters grounded in the official band descriptors. The Reading and Listening modules are scored against answer keys and converted to bands through weighted thresholds. The tool assigns the user to an overall band (proficiency level) and a per-module band. It also provides a skill profile for adaptive curriculum design to assist in targeted study. Evaluation results indicate consistent dual-rater agreement and measurable learner progress.

A3: NeuroCore Summit III: Brain-Inspired Algorithm–Hardware Co-Design and Applications

16:20–17:20 · Room A · Chair: Elishai Ezra Tsur (The Open University of Israel) · Co-chair: Chandrashekhar Medicherla (Salesforce Inc)

16:20	#279 **Investigating the Impact of Mapping-Awareness in Hardware-Aware Spiking Neural Architecture Search** Hella Toto Kiesa and Jonas Bergmann (Karlsruhe Institute of Technology); Sven Nitzsche (FZI Forschungszentrum Informatik); Tanja Harbaum and Juergen Becker (Karlsruhe Institute of Technology) Energy-efficient deployment of neural networks (NNs) at the edge remains a major challenge. Spiking Neural Networks (SNNs) combined with hardware-aware Neural Architecture Search (NAS) are a promising response. Existing hardware-aware NAS (HW-NAS) either applies the mapping of neurons onto neuromorphic cores only after the search or explores it together with further hardware parameters, which makes its individual effect hard to isolate. We present MapSNAS, a HW-NAS framework for SNNs that supports a fixed mapping and a mapping-aware search over predefined mapping strategies on a given neuromorphic platform, evaluated with the SANA-FE simulator targeting Intel's Loihi. We first characterize three mapping strategies across SNN architectures of varying sizes on DVSGesture and SHD. NeuMap offers the best trade-off, with the lowest energy in all cases and the best or second-best traffic-latency product (TLP). Using NeuMap as the fixed mapping, we then conduct an ablation study on DVSGesture comparing accuracy-only search, hardware-aware search and mapping-aware search. Contrary to our expectation, mapping-aware search does not improve over the fixed mapping under the same search budget. Its Pareto fronts consist almost exclusively of NeuMap. We find no lower TLP at matched accuracy, with comparable energy. Retrained with three seeds, the best fixed-mapping architecture reaches a mean accuracy of 0.907 compared to 0.875 for mapping-aware search. For HW-NAS on a given platform with few predefined mapping strategies and a limited search budget, our results suggest characterizing the strategies once upfront and fixing the best one instead of enlarging the search space.

16:35	#82 **An Energy-efficient BitNet LLM Accelerator with a Reconfigurable Bit-cascaded MAC Core and 1.6-bit Online Weight Decompression** Chaoming Fang, Ziyang Shen and Jie Yang (Westlake University); Mohamad Sawan (Westlake University & Polytechnique Montréal) BitNet ternary-weight LLMs promise major memory and energy savings, yet edge inference is jointly constrained by prefill compute and decode off-chip bandwidth. We present a BitNet accelerator with three co-designed components targeting this bottleneck. First, a reconfigurable bit-cascaded MAC core reuses ternary adder-tree lanes for 2b×8b BitLinear operations and, via bit-position shifting and selective negation, for 8b×8b self-attention, unifying both kernels in one datapath. Second, a 1.6-bit base-3 weight encoding packs five ternary weights into one 8-bit symbol, with a fully combinational on-chip decoder whose group size is selected from a systematic sweep of storage cost against decoder area. Third, a bandwidth-aware balance among compute lanes, decoders and DRAM traffic optimizes end-to-end throughput. Implemented in Verilog RTL and synthesized as a 32-core TSMC 28nm instance at 400 MHz with 20 GB/s memory, our design reduces time-to-first-token by 21.6% and raises decode throughput by 21.4% versus a baseline with dedicated non-reconfigurable 2b×8b and 8b×8b MAC arrays and conventional 2-bit-per-weight storage, under identical chip area and memory bandwidth, across BitNet 2B, 3B and 7B over sequences 256-4096.
16:50	#38 **Next-Generation Neurotechnology AI-Augmented Systems for Emotional & Cognitive Enhancement** Manish Meshram (Crave Infotech); Sparsh Potti (Research Assistant) The dissertation studies next-generation neuro-technologies and the Artificial Intelligence (AI) augments responsible for the regulation of emotions and cognition improvement. AI neurotechnology is projected to be the most dominant in the next few years as BCI (brain-computer interfaces) technologies currently raking in $5.46 billion projected revenues by 2030. This dissertation studies 147 research articles in the form of clinical studies for the years 2020 to 2025. The focal point of the studies entails three ADAPT core technologies: neural feedback systems; adaptive neural stimulation systems; and artificial intelligence driven emotional prediction systems. The research shows that AI neurofeedback systems are able to achieve 67% success rates on emotional regulation tasks in comparison to 31% in baseline traditional systems (analyzed emotional regulation system studies on 892 patients). Cognitive system augmentations/AI applications report about 15-22% improvements on target cognitive systems of interest such as working memory and an improvement on the attention span by 28% in clinical trial settings. Deep learning neural network systems analyzing emotional states through pattern recognition on EEGs are able to achieve 89% of real-time prediction of emotional states 4-6 seconds prior to the target emotions becoming conscious.
17:05	#164 **A Resource-Efficient Deepfake Images Detection Pipeline Utilizing Lightweight Convolutional Neural Networks** Rasel Ahmed Raju, Zeba Tashin Oeeshee, Md. Rakibul Islam Reshad and Muhammad Shafayat Oshman (North South University) The proliferation of generative AI has facilitated the creation of hyper-realistic deepfake media, posing a critical threat to digital trust. While existing detection frameworks achieve high accuracy, they often rely on computationally intensive architectures unsuitable for real-time deployment. This paper addresses this gap by proposing a lightweight, custom-built Convolutional Neural Network (CNN) architecture. Trained on a balanced dataset of 190,000 images, the model achieves a test accuracy of 97.89% and an AUC-ROC of 0.93. The study introduces an end-to-end pipeline integrating the model into a real-time web interface via Gradio, demonstrating an inference latency of 40ms per image. This research provides a scalable solution for frontline verification by non-technical stakeholders.

B3: S-Session 25: Governance-Aware Agentic AI for Systems-of-Systems

16:20–17:20 · Room B · Chair: TBC · Co-chair: Abdul Khadir Mahamed (New York University)

16:20	#145 **AI Ambassadors: A Governance-Aware Agentic AI Framework for Trusted and Autonomous Consular Service Delivery in Digital Embassies** Mohammed Saeed Jawad (International University of Sarajevo); Marcin Mrugalski (University of Zielona Góra); Taher Alzahrani (Imam Mohammad Ibn Saud Islamic University (IMSIU)); Dino Kečo; Mohammad Al-Johani (International Burch University (IBU)) All embassies and consulates are under pressure to deliver more efficient and effective services to citizens. They are also under pressure to strengthen transparency and accountability, and comply with regulations. Digital government initiatives allow citizens to access government services more easily. However, many consular services remain fragmented, labour-intensive, and reliant on face-to-face interactions. Recent developments in agentic artificial intelligence (AI) create the possibility of automating the complex and fragmented administrative tasks of consular services through the use of autonomous agents that possess the ability to reason, plan, and make decisions. Introducing autonomous AI to government systems and services creates significant governance challenges regarding trust, explainability, accountability, and compliance. This paper introduces AI Ambassadors, a framework of governance-focused agentic AI, designed to provide autonomous consular services within digital embassies. The framework incorporates governance-focused agentic AI systems to provide autonomous consular services for visa application support, citizen support, document verification, appointment scheduling, emergency assistance, and consular services compliance. A governance orchestration layer provides supervision of agents via policy guidance, risk assessment, explainability, and human oversight. With its integrated trust assurance framework, the proposed system employs blockchain technology to provide unalterable records of decisions and the rationale for making them. The proposed design for a digital embassy system is evaluated through a simulation, the Digital Embassy Twin. Initial findings of the simulation indicate improvements in service efficiency, system governance, explainability, and citizen trust, when compared to traditional embassy systems, rule-based embassy systems, and unaligned agentic AI systems
16:35	#300 **SPARKI: Evidence-Contract Support for Governed AI-Assisted Engineering Workflows** Chad Curbow and Michael C Meyer (Eastern Washington University) AI assistants are very good at generating plausible engineering solutions while ignoring key real-world limitations. Additional Checks must be performed before a final output can be generated in order to reduce hallucinations and create a more complete product. In this paper, we present SPARKI (Systems Platform for AI, Research, Knowledge, and Innovation), an evidence-contract support method that provides bounded context, source-access rules, structured completion requirements, and separately executed task-specific automated checks while reserving consequential decisions for people. We compare SPARKI to the baseline Codex using 12 different benchmark tests. Some of the benchmarks had completion differences of 0 out of 20 on the baseline vs 15 out of 20 when using SPARKI (paired difference 75 percentage points; exact McNemar p=6.10x10^-5; unadjusted paired percentile-bootstrap 95% interval 55–90 points). In several benchmarks, both the baseline and SPARKI successfully completed every individual test. Overall token-use and time to

	result did not yield conclusive evidence that either the baseline or SPARKI had benefits over the other.
16:50	#306 **PAGER: Policy-Aware Routing and Agent Selection for Multi-Agent Systems** Siva Kumar Chintham (LTM); Seema Gangaiah Aarella (Austin College & University of North Texas); Atul Prasad (University of San Diego); Shomik Basu (Workday Inc); Sai Srinivas Choudary Potla (Innovaccer); Pratik Bundela (IEEE); Akshay Deshpande (Waymo); Shiva Kumar Devalokam Subrayappa (Evernorth Health Services) Secure and efficient query routing is a critical challenge in multi-agent systems, particularly in regulated environments where agent selection must satisfy authorization, privacy, and compliance requirements. Existing routing approaches primarily consider semantic similarity, cost, latency, or response quality without explicitly enforcing governance policies. This paper presents PAGER (Policy-Aware Agent Router), a framework that integrates policy enforcement with capability-based agent selection and configurable conflict resolution. PAGER employs Open Policy Agent (OPA) and Rego to evaluate hard constraints before routing requests, ensuring that only policy-compliant agents are considered for execution. A conflict resolution module subsequently selects an eligible agent using cost, latency, quality, weighted scoring, or round-robin strategies. The framework is evaluated using 300 healthcare tasks from MedAgentBench with simulated agents and predefined compliance policies. PAGER achieves 100% routing accuracy on the benchmark tasks with a compliant agent pool. In an adversarial configuration containing non-compliant agents, it prevents all policy violations and correctly denies 30 requests without eligible agents. However, evaluation on 20 held-out paraphrases reveals limitations in rule-based query analysis, reducing routing accuracy to 10% and exposing two policy violations. These findings demonstrate the effectiveness of policy enforcement before agent selection while highlighting the importance of reliable query interpretation. PAGER provides a modular approach to governance-aware routing in multi-agent systems, with potential applications in regulated AI environments.
17:05	#309 **Authorization-Aware Detection of Indirect Prompt Injection in Retrieval-Augmented Generation** Aswathnarayan Muthukrishnan Kirubakaran (Meta Inc); Pratik Bundela (IEEE); Ram Sekhar Bodala (Amtrak); Akshay Deshpande (Waymo); Murali Shankar Dulam (JPMorgan Chase); Suhas Malempati (MetaDigi Labs Inc); Anurag Kumar (Cisco Systems); Nitin Saksena (Albertsons Companies) Retrieval-augmented generation (RAG) improves language-model grounding but creates an untrusted instruction channel: text retrieved as evidence may be interpreted as executable control. Passage-only prompt guards cannot reliably distinguish malicious instructions from legitimate procedural language because the same sentence may be authorized, informational, or task-diverting depending on the requested operation and source authority. This paper presents a task- and authority-aware framework that assigns localized instruction spans one of three roles: informational, authorized, or diverting. A counterfactual benchmark places identical instruction spans under different task and authority conditions, and instruction-family-disjoint splits prevent paraphrase leakage. A controlled evaluation isolates the authorization-classification problem using oracle-localized spans and 1,275 examples from 85 instruction families. At a threshold selected on validation data to maximize attack recall subject to a 1% safe false-positive-rate (FPR) constraint, the full-context classifier achieves 98.3% macro F1, 98.5% diverting-instruction recall, and 0.2% test FPR. Passage-only classification remains non-discriminative (AUROC 0.50), confirming that instruction text alone is insufficient. Metadata stress tests show gradual degradation under missing fields and reduced recall under spoofed authority. The study establishes a reproducible benchmark and evaluation protocol for authorization-aware prompt-injection detection while explicitly separating the validated relation-classification contribution from future predicted-span and end-to-end RAG evaluation.

A4: NeuroCore Summit IV: Neuromorphic Accelerators and System Architectures

17:20–18:20 · Room A · Chair: Elishai Ezra Tsur (The Open University of Israel) · Co-chair: Siddhant Bapusaheb Tanpure (Meta)

17:20	#92 **SCDQ: A Shared Circular Delay Queue for Scalable Synaptic Delays in Neuromorphic Accelerators** Roy Meijer (Eindhoven University of Technology); Paul Detterer (IMEC-NL); Amirreza Yousefzadeh (University of Twente); Alberto Patino-Saucedo (IMSE); Guangzhi Tang (Maastricht University); Kanishkan Vadivel (IMEC); Yingfu Xu (IMEC-NL); Manil Dev Gomony (Eindhoven University of Technology); Federico Corradi (TU Eindhoven); Bernabe Linares-Barranco (Instituto de Microelectronica de Sevilla); Manolis Sifalakis (Innatera) Synaptic delay parameterization of neural network models has shown increasingly promising results in recent literature, suggesting that delay-parameterized models are simpler, smaller, sparser, and thus more energy efficient than similar-performing non-delay-parameterized ones. We introduce the Shared Circular Delay Queue (SCDQ), a novel hardware structure for supporting synaptic delays on digital neuromorphic accelerators. Our analysis shows that the SCDQ delay structure scales better in terms of memory than commonly used Ring Buffer and Shared Delay Queue approaches, and is more amenable to algorithm-hardware co-optimization, as its memory overhead is modulated by model activation sparsity rather than network size alone. We evaluate SCDQ on the Seneca neuromorphic processor and show that its dedicated circuit implementation (Delay IP) reduces inference energy by up to 3.5x and latency by up to 4.3x compared to an equivalent software implementation, while contributing only 2-3% of total system energy and occupying just 3% of the area of a single compute core.
17:35	#99 **A Low-Cost FPGA Implementation of a Time-Division-Multiplexed Spiking Neural Network** Marwan Ahmed Elsayed Fetteha and Ameer Abdelhadi (McMaster University) This paper presents a low-cost hardware implementation for spiking neural network (SNN). The design employs a deterministic spike-based rate encoding scheme and a time-division multiplexed (TDM) leaky integrate-and-fire (LIF) neuron architecture to reduce hardware resource utilization while preserving functional equivalence with a reference software model. The proposed architecture is verified using software model demonstrating exact agreement in membrane potential evolution, spike timing, and classification results. Experimental evaluation on the Modified National Institute of Standards and Technology (MNIST) handwritten digit recognition confirms correct inference behavior. The presented design achieves an accuracy of 95.14% on the MNIST data set. The proposed design was implemented on Artix-7 AC701 Field-Programmable Gate Array (FPGA) and uses 31.40% Look-Up-Tables (LUTs) and 7.47% Flip-Flops (FF) of its resources. The proposed design provides a practical and scalable approach for implementing large scale SNNs on hardware.
17:50	#68 **The Sparsity Tax: Weight Sparsity Trade-offs in Event-Driven SIMD and SIMT Neuromorphic Cores** Mattias Westerink, Sameed Sohail, Berend Jan van der Zwaag, Sabih H. Gerez and Amirreza Yousefzadeh (University of Twente)

	Event-driven neuromorphic inference exploits activation sparsity by updating neuron state only on spikes. However, weight sparsity introduces irregular gather-style updates that undermine lockstep Single Instruction Multiple Data (SIMD) execution. We call the resulting overheads in control, metadata, and memory activity the sparsity tax. This paper quantifies that tax by comparing three closely related accelerators integrated into one neuromorphic core: (i) baseline lockstep SIMD, (ii) bitmap-gated Sparse-SIMD that selectively disables lanes without compressing weights, and (iii) a Single Instruction Multiple Threads (SIMT) style design with per-PE address generation and run-length coded sparse weights. Using an RTL-to-gates flow in GF22FDX+ and activity-driven energy estimation, we evaluate event-driven neural network inference across varying post-training pruning levels. Results show that total core area changes are insignificant because SRAM dominates area, while performance and energy strongly depend on how sparsity is handled: SIMD and Sparse-SIMD exhibit near-constant throughput, Sparse-SIMD achieves limited energy savings due to bitmap and dense-storage overheads, and SIMT provides the strongest energy scaling and substantial speedups at high sparsity, albeit with sublinear gains due to metadata reads, load imbalance, and sparsity-independent phases. The hardware code for the proposed architectures and experiments is publicly accessible for research purposes.
18:05	#119 **An Event-Driven Asynchronous Multi-Lane Serial Interconnect for Chiplet-Scale Neuromorphic Integration** Sameed Sohail, Douwe De Vries, Berend Jan van der Zwaag, Sabih H. Gerez, Ana Lucia Varbanescu and Amirreza Yousefzadeh (University of Twente) Chiplet-based (2.5D/3D) integration is emerging as a scalable approach for neuromorphic systems beyond single-die limits, avoiding yield and cost challenges of large monolithic designs, but requiring interconnects that efficiently support sparse communication patterns. Existing high-performance chiplet interconnects are optimized for dense, continuous traffic, either using a large number of parallel I/O wires or specialized mixed-signal PHYs for serial communication, which limit scalability and make them inefficient for sparse workloads. In contrast, event-driven serial links are either not fully synthesizable due to reliance on analog front-ends or are limited in scalability. This paper presents an open-source, fully digital, asynchronous, event-driven multi-lane serial interconnect for chiplet-scale neuromorphic integration, using a protocol with in-band flow control and configurable lane parallelism to achieve scalable throughput with minimal I/O count and no analog components. Synthesized in GlobalFoundries 22-nm FDSOI, a 4-lane configuration occupies 817 um2 and consumes 0.23 pJ/bit. FPGA-based validation demonstrates robust operation under chip-to-chip boundary conditions, while ASIC simulations show a compact, energy-efficient, and portable interconnect suitable for chiplet-scale neuromorphic systems.

B4: S-Session 18: Harnessing AI Tools for Transforming Education

17:20–18:20 · Room B · Chair: Yousuf Nasser Al Husaini (Arab Open University) · Co-chair: Youngsoo Kim (University of Minnesota Duluth)

17:20	#149 **Hybrid FPGA-Based Random Number Generator Using Linear Feedback Shift Registers, Thermal Noise Injection, Von Neumann Correction, and Nonlinear Post-Processing for Conway's Game of Life Visual Display** Teo Tee Hui, Vincent Ho and Khant Htet Kaung (Singapore University of Technology and Design); Binghe Li (ZJU-UIUC Institute, Zhejiang University, Haining); Rey Lam, Jeremiah He, Logavarshini Kandaraj and Kulkarni Ojas Jaydeep (Singapore University of Technology and Design) This study presents a hybrid hardware-based random number generator (RNG) implemented on a CMOD A7 FPGA, which is designed to overcome the deterministic limitations of conventional linear feedback shift registers (LFSRs). By integrating raw thermal noise from the FPGA's internal XADC temperature sensor and applying a Von Neumann corrector, the design injects true stochastic entropy into a modified 16-bit LFSR architecture. Additional nonlinear post-processing, including S-box substitution and arithmetic mixing with a free-running LFSR, enhances the output entropy and breaks the linear correlations. The RNG output drives unique, non-reproducible visual patterns, and Conway's Game of Life is displayed via a TFT screen, with data transmitted over UART for logging and analysis. Statistical evaluation using NIST SP 800-22 tests confirms the generator's uniformity, unpredictability, and high entropy, passing key tests such as Monobit, Runs, Approximate Entropy, and Linear Complexity, while highlighting areas for future improvement. This system effectively integrates physical entropy sources to produce aperiodic, high-quality random sequences that are suitable for embedded applications requiring secure randomness. This study underscores the potential of combining physical noise and algorithmic techniques to enhance the randomness of FPGA-based random number generators (RNGs).
17:35	#150 **Novel Interactive Reaction Game Design Incorporating Real-Time FPGA-Based Multiple-Recursive Pseudo-Random Number Generator for Psychomotor Training** Teo Tee Hui, Zhao En Koh, HoI Chia, Renzo Ricardo King, Francine Flores Casana and Joel Yeo (Singapore University of Technology and Design); Yifei Yang (University of Waterloo); Bo Wen Heng (Singapore University of Technology and Design) The design, development, and integration of a Pseudo-Random Number Generator within the framework of an interactive reaction game is presented in this work. The background motivation stems from the need for unpredictability and fairness in game mechanics, where randomization plays a critical role in enhancing user engagement and challenge. The methods involved the selection and implementation of an efficient PRNG algorithm tailored to the game's real-time requirements, ensuring the rapid and reliable generation of random sequences. Integration was achieved by embedding the PRNG into the game's core logic, enabling dynamic stimulus presentation based on random timing and event occurrence. The performance was evaluated through a series of tests that measured the response times, randomness quality, and system stability during gameplay. The results demonstrated that the integrated PRNG maintained high-quality randomness without compromising the game's responsiveness or the user experience. The system exhibited consistent behavior across multiple sessions, thereby validating the robustness of the approach. In conclusion, the successful integration of the PRNG significantly contributed to the functionality of the interactive reaction game by providing essential randomization features that support fair and engaging gameplay. This implementation offers a scalable model for incorporating pseudo-random number generation in similar real-time interactive applications, with the potential for further optimization and enhancement based on user feedback and evolving technical standards in the future.
17:50	#151 **Design and Implementation of a Correlation-Aware Digital Shaker System for Pseudo-Random Number Generation Using Piezoelectric Sensing and Congruential Generators** Teo Tee Hui, Harsh Agarwal and Benjamin Lee (Singapore University of Technology and Design); Vanessa Slavov (University of Waterloo); Sanika Kalyan Sonpethkar, Mohamed Danial B Hamdan and Aryan Gupta (Singapore University of Technology and Design)

	This paper documents the design of a digital shaker system for a pseudo-Random Number Generator (PRNG) using piezoelectric sensing, with a focus on evaluating entropy extraction methods from mechanically induced signals. The system comprises piezoelectric sensors, an analog-to-digital converter (ADC), and a Verilog Hardware Description (Verilog HDL) processing pipeline within a 3D printed housing. The experimental results show that the piezoelectric signals suggest the presence of temporal correlations across independent channels. An XOR bit-shifting-based combination of sensor signals for processing entropy was explored, and it was observed to better preserve higher-order signal information under temporally correlated conditions. The signals were further conditioned using multiplicative and linear congruential generators to produce randomized numbers. The results obtained via the conducted tests demonstrate the limitations of the selected randomness extraction methods when applied to correlated physical signal sources, highlighting the importance of having a correlation-aware processing pipeline.
18:05	#152 **Programmable Anomaly Detector System: An Embedded, Real-Time IoT Acoustic Fault Detection using Hybrid FFT, CNN Autoencoder on a Resource Constrained FPGA** Teo Tee Hui and Gavin Tan (Singapore University of Technology and Design); Eric Aleong (University of Waterloo); Keming Su, Wesley Teo and Haneeshwar Reddy (Singapore University of Technology and Design) This project experiments with cost- and energy-efficient methods for autonomously detecting mechanical failures. The central idea behind our approach is that audio-based detection can be significantly more energy- and hardware-efficient than visual-based methods when monitoring failures, such as 3D printers. A single microphone can simultaneously monitor numerous 3D printers, whereas individual cameras would be required at each printer if a visual approach were used. Furthermore, standard microphone technology (MEMS/electret) is often 1–2 orders of magnitude less energy-intensive than standard CMOS cameras. In this study, we designed and built an end-to-end embedded system that captures and processes audio, stores operational data, and communicates with users when anomalies are detected. A hybrid model integrating FFT and a CNN autoencoder was implemented as a programmable detector core on an FPGA board for real-time processing. An anomaly detector equipped with a 3D printed enclosure, a display with a GUI, and an APP on both local devices and the cloud web was designed and demonstrated, and was named the Programmable Anomaly Detector System.

Tuesday, December 15, 2026

A5: Hardware Acceleration of AI on Embedded Edge SoCs – I

10:40–11:55 · Room A · Chair: Mushtaq Ahmed (Malaviya National Institute of Technology Jaipur) · Co-chair: Md Farhadur Reza (Eastern Illinois University)

10:40	#87 **Mamba-Family State-Space Model Kernels on a Programmable CGLA** Takuto Ando and Yasuhiko Nakashima (Nara Institute of Science and Technology) Edge and embedded inference is constrained by power and data movement. Mamba-family state-space models replace attention with sequence-linear recurrence, but their inference path combines dense projections, short-reduction SSD kernels, and recurrent-state updates. This paper maps these kernel groups onto IMAX, a programmable CPU-Grounded Linear Array (CGLA), and measures them from kernel execution to token-level integration. Projection kernels match the long-reduction IMAX pipeline, whereas SSD Step-1 is limited by short reductions and kernel-boundary overheads. Mamba-130M token-level integration identifies projection GEMV as the decode bottleneck. Across 36 matched SSD kernel shapes, IMAX has lower modeled energy in 24 shapes and lower latency in 6 shapes relative to Jetson AGX Orin under the stated IMAX ASIC and Jetson power assumptions. These results show that programmable CGLAs fit long-reduction projection kernels, while SSD and decode-time projection support require boundary reduction and persistent-weight execution.
10:55	#162 **FPGA Implementation and Performance Evaluation of Memory-Loaded Ternary Linear Layers under Varying Sparsity** Ryota Kayanoma and Hiroki Nakahara (Tohoku University); Yumiko Uchida and Yoshitaka Miyatani (Sony Semiconductor Solutions Corporation) Transformer models repeatedly execute large linear layers, whose computation and memory costs make hardware implementation difficult at scale. This paper presents a scalable FPGA architecture for sparse ternary linear layers that does not rely on LUT-based computation. The proposed processing element eliminates multipliers by exploiting ternary weights and sequentially accumulates only non-zero weights, while the combination of patch-level and channel-level parallelism with broadcast-oriented on-chip memory improves throughput. As an evaluation target, we apply ternary weight training and 4-bit activation quantization to a Vision Transformer using sparse event-derived inputs, which are suitable for model sparsification and low-bit inference. Experiments on the DVS128 dataset show that a 70% sparsity ratio provides an effective accuracy-throughput trade-off, with limited additional accuracy degradation among the evaluated sparse settings despite an accuracy loss from the FP32 baseline. At this operating point, our FPGA implementation achieves 2254 FPS, corresponding to 1.03 times the throughput of an embedded GPU baseline of 2183 FPS; at 80% sparsity, the FPGA reaches 3382 FPS, corresponding to 1.55 times the GPU throughput. These results demonstrate that the proposed LUT-independent architecture effectively exploits ternary weights and structured sparsity to reduce memory usage and improve throughput for large-scale Transformer linear layers.
11:10	#123 **Accuracy-Aware, Hardware-Adaptive Co-Optimization of Pruning and Dataflow Mapping for Modern DNNs** Md Farhadur Reza (Eastern Illinois University) The growing deployment of deep neural networks (DNNs) on resource-constrained devices demands high computational efficiency and low energy consumption. Pruning and dataflow mapping are key optimization techniques, but they are typically applied independently, leaving significant synergy unutilized. This paper introduces a reinforcement learning (RL) framework that jointly explores pruning strategies and dataflow mappings, guided by a hardware-aware cost model and a real accuracy feedback loop. Unlike prior works that treat sparsity statically, our approach dynamically adjusts pruning granularity, sparsity levels, tile sizes, and dataflow types on a per-layer basis while preserving accuracy through fine-tuning. We formulate co-optimization as a Markov decision process and solve it with a deep Q-network. The framework targets modern DNNs (ResNet, MobileNet, Transformers) and generalizable accelerator architectures. Experiments on CIFAR-10 with ResNet-18 show that the RL agent learns non-obvious trade-offs, achieving 42.9% energy reduction and 45.9% latency reduction compared to a dense baseline, with only 0.7% accuracy loss (from 80.1% to 79.4%). The learned policy outperforms a rule-based heuristic (43.6% energy reduction but 1% accuracy drop) and a

	random search baseline (40.4% energy reduction with a 1% accuracy drop). The agent autonomously discovers effective strategies such as block-sparse pruning with row-stationary dataflow at 50% sparsity, a combination that human designers might overlook. This work provides a foundation for autonomous, accuracy-aware DNN accelerator optimization.
11:25	#171 **CNN-Based Many-Core Single-Image Super-Resolution** Pei Yun Chen and Bevan Baas (University of California, Davis) Single-image super-resolution (SISR) using convolutional neural networks (CNNs) has become important for modern image enhancement but remains computationally challenging for real-time deployment on edge devices. This work presents an implementation of the Fast Super-Resolution Convolutional Neural Network (FSRCNN-s) able to perform 2×, 3×, and 4× upscaling on the KiloCore many-core platform using 133 fine-grain programmable processors. The C++ design achieves a scaled throughput-per-area improvement of 29-100×, 19-39×, and 7-8× higher than the Intel 11800H, Intel 12600K, and NVIDIA RTX 3050 GPU, respectively. Its scaled energy efficiency (images/J) exceeds the commercial processors' by 25-87×, 43-87× and 6-10× respectively. For the combined performance and efficiency metric of throughput/area × images/J, the KiloCore exceeds the commercial chips by 200-2420×, 225-934×, and 11-15× respectively.
11:40	#313 **Accelerating Sparse Matrix-Vector Multiplication on Ryzen XDNA2 NPU** Hitoshi Komurasaki (The University of Tokyo); Masayuki Usui (Tokyo University); Takatsugu Ono (Kyushu University); Shinya Takamaeda-Yamazaki (The University of Tokyo & RIKEN) This paper presents a sparse matrix-vector multiplication (SpMV) method for the XDNA2 neural processing unit (NPU) in Ryzen AI processors. NPUs are increasingly used to accelerate local large language model (LLM) inference, in which matrix-vector multiplication accounts for a substantial portion of the workload during autoregressive decoding and is often limited by the bandwidth between main memory and the NPU. Unstructured pruning can substantially reduce the weight footprint of LLMs and potentially accelerate autoregressive decoding, but exploiting sparse weights requires fast SpMV for XDNA2 NPU. The sparse matrix format is central to this challenge because it determines how nonzero elements are organized for vector computation and data transfer. CSR retains compact storage, but its variable row lengths complicate fixed-width vector processing and fixed-size block transfers on XDNA2 NPU. ELL regularizes both by padding rows to a uniform width, potentially introducing substantial zero-padding overhead. We present Blocked SELL-C-σ, a sparse matrix format and SpMV implementation for XDNA2 NPU. The proposed design reduces padding while preserving fixed-length transfers and enabling efficient vector computation. We evaluate the method using five representative weight matrices from a Llama-2-7B model pruned with ELSA to a 10% nonzero fraction. Compared with dense GEMV applied to the same pruned weights, our SpMV implementation achieves a 2.23-4.29x speedup.

B5: S-Session 1: Performance Optimization and Auto-Tuning of Software on Multicore/Manycore Systems (POAT2026)

10:40–12:10 · Room B · Chair: Siva Kumar Chintham (LTM) · Co-chair: TBC

10:40	#81 **Neural-Guided Warm Starts for Digital Annealing of Large-Scale QUBO Problems** Yu Ting Chen, Yu-Ru Pai and Yu-Chen Shu (National Cheng Kung University) Quadratic Unconstrained Binary Optimization(QUBO) is a fundamental NP-hard problem with broad applications in combinatorial decision-making. Although digital annealers can produce high-quality solutions, their runtime remains a bottleneck in repeated-query and large-scale settings. In this work, we propose a neural warm-start strategy for accelerating QUBO optimization. Rather than replacing the annealing process, we use a neural network to generate initial solutions for further QUBO procedure using the Fujitsu Digital Annealer. Numerical experiments suggest the effectiveness of our method by showing a 35.19% improvement.
10:55	#172 **Benchmark-Guided Selection of Managed, Prefetched, and Manual GPU-to-GPU Migration** Masatoshi Kawai (Tohoku University); Satoshi Ohshima (Kyushu University); Hiroyuki Takizawa (Tohoku University) GPU-to-GPU migration is important for multi-GPU applications, but the best implementation method depends on the migration pattern. Managed Memory improves programmability, while prefetching and manual transfer can reduce migration overhead. This paper proposes a benchmark-guided selection model for choosing among managed, proactive, and manual migration. The model estimates migration time using measured latency indexed by array size and the number of arrays. Microbenchmark results on NVLink-connected H100 GPUs show that the appropriate method depends on both parameters. The proposed model enables a staged policy that keeps data managed when possible and applies prefetching or manual transfer only to performance-critical migrations.
11:10	#220 **A Branch-Free General Renormalization Scheme for Pair Arithmetic and Its Performance Evaluation** Tomonori Kouya (Otemon Gakuin University) Pair arithmetic omits the renormalization stage performed at the end of each operation of multi-component multiple-precision arithmetic, reducing the operation count and, by eliminating conditional branches, easing SIMD vectorization. Renormalization cannot be omitted entirely, however, because iterative solvers may then fail to converge. We propose RenormBF-pair, a branch-free fixed-trip-count renormalizer parameterized by the word count K and the input length n: a VecSum sweep, a tail fold, and r rounds of a FastTwoSum chain, costing 6(n-1)+(n-K)+3r(K-1) flops. Over 200,000 trials per condition, the smallest round count producing no non-overlap violation was r=1 for K ≤ 3 and r=2 for K=4, whereas the same cost spent on a VecSumK-style TwoSum chain fails for K=4. We delimit with explicit counterexamples both the input condition under which the sum relation holds and the fact that the scheme does not, in general, guarantee a correctly rounded lowest component. Integrating it into CG and BiCGStab over pair arithmetic in binary64 and binary32, we sweep 240 paired runs including real matrices screened from SuiteSparse. The proposed scheme and VecSumK returned bitwise identical words wherever both succeeded, so they are indistinguishable for CG in binary64; differences appear only where VecSumK breaks, namely binary32 and BiCGStab. There the 22–16 win–loss record is not significant (sign test, p=0.42); the clear bias is structural, non-overlap violations dropping from 66 rows to 37. The cost is +3.4% for the whole CG solver.
11:25	#282 **Degrees of Freedom in Submit-Time HPC Job Power Prediction: A Measurement Study**

	Haoran Ding, Sichen Tao, Masatoshi Kawai and Hiroyuki Takizawa (Tohoku University) Submit-time job power prediction is judged on small margins: recent work reports gains of a few tenths of a watt. We ask what a result of that size depends on, in a measurement study of two Fugaku windows under one pipeline, reporting replicate distributions rather than point estimates. Three degrees of freedom decide the outcome before any method choice does. The first is measurement: training-row order alone moves the pooled coefficient of determination by up to 0.08 across otherwise identical replicates, and 0.21 once the clustering seed varies with it; resampling jobs rather than days narrows a confidence interval thirtyfold. The second is the data condition: a cohort reporting per-node power below the machine's idle draw falls in the test half of one window and the training half of the other, so the two are not comparable populations; removing it moves the numeric baseline's R^2 from 0.811 to 0.269. The third is the window itself, which decides which cluster count is best (k = 100 ranks first of seven on one window and sixth on the other), whether a clustering prior helps at all, and whether the per-day gain is predictable (three submit-time day descriptors survive multiple-testing correction in one of the four cells, none in the rest). Against these, the method's own configuration moves the result by about as much as re-running it does, while equal-budget tuning moves the arms unequally, by more than they differ from each other in half the cells. We close with a reporting checklist.
11:40	#294 **Virtual Threads on a Heterogeneous Multicore SoC: An Empirical Study of Java Thread-per-Connection Servers under I/O, CPU and Synchronization Load** Udaya Bhasker Mattupalli (Accenture); Ayaskant Pattnaik and ShravanKumar Arram; Krishnaprasad Chouty (Sophlogic Global LLC); Dinesh Babu Mannu (Fidelity Investments); Balamurugan Chelladurai (Wipro Technology); Naresh Kumar Sankar Kumar Java 21 introduced virtual threads, which the JVM multiplexes onto a small pool of carrier threads sized by the JVM's processor count. We measure when they help a thread-per-connection server on a heterogeneous multicore SoC (Apple M2, 4 performance + 4 efficiency cores), using an open-loop load generator that avoids coordinated omission and 5 independent repetitions of every load experiment. For blocking I/O, a 200-thread platform pool was capped by Little's law at 3,789 ± 88 requests/s, and a 1,000-thread pool failed in 3 of 5 repetitions at 15,000 requests/s, while virtual threads sustained the full load in every repetition. On JDK 21, blocking inside synchronized pinned carriers and cut throughput to 74.3 requests/s, at least 27 times below the unpinned variants; on JDK 27 (with JEP 491) the same code sustained the full offered 2,000 requests/s. For CPU-bound work, virtual threads delivered 29% (one carrier) down to 7% (four carriers) more capacity than a platform pool with one thread per carrier, because such a pool strands keep-alive connections; at eight carriers the difference was not significant. A control shows that -XX: ActiveProcessorCount sizes the JVM but does not confine platform threads to cores. Virtual threads also removed the OS-thread ceiling that limited a platform server to about 2,000 idle connections. Carrier placement on performance or efficiency cores was not controlled, and the load generator shares the SoC, so high-load tail latencies are upper bounds. The benchmark, scripts and raw data will be released upon acceptance.
11:55	#297 **Accuracy at a Cost: Efficiency-Performance Trade-offs in Recurrent and Attention-Based Code Models** Muhammad Jawad Arshad, Anoop Yadav, Timur Jaganov and Yutaka Watanobe (University of Aizu) The rapid development of Artificial Intelligence (AI) has led to the widespread adoption of diverse neural architectures for complex prediction tasks. However, comparing architectures solely on predictive performance provides an incomplete view of their practical effectiveness, particularly when computational cost, model capacity, and resource requirements are considered. This paper presents a controlled empirical study of two widely used neural architectures, Bidirectional Gated Recurrent Units (BiGRU) and Transformer-based encoders, under both single-task and multi-task learning settings. To ensure a fair comparison, experiments are conducted using two configurations: default hyperparameters and a parameter-matched setting in which the architectures are adjusted to provide comparable model capacity. The models are evaluated across multiple random seeds using predictive performance and computational efficiency measures, including accuracy, parameter count, floating-point operations (FLOPs), inference latency, peak memory consumption, and energy usage. A TF-IDF with Logistic Regression model is additionally employed as a lightweight non-neural baseline. Experiments are conducted on a subset of the Project CodeNet dataset. Furthermore, sequence-length scaling experiments are performed to examine the trade-offs between predictive performance and computational cost. The results reveal a clear accuracy-efficiency trade-off: BiGRU achieves higher predictive accuracy, whereas the Transformer requires substantially fewer FLOPs and lower inference latency and estimated GPU energy consumption, but uses substantially more peak GPU memory. These findings show that architecture selection should consider both predictive performance and computational resource constraints rather than accuracy alone.

A6: Architectural Innovations in Embedded Multicore/Manycore SoCs – I

14:40–15:40 · Room A · Chair: Mushtaq Ahmed (Malaviya National Institute of Technology Jaipur) · Co-chair: Siva Prasad Nandi (Oracle America Inc.)

14:40	#114 **IAIP: Interrupt-Aware Instruction Prefetcher** Tim Twardzik and Lars Nolte (Technische Universität München); Thomas Wild and Andreas Herkersdorf (Technical University of Munich) Quick execution of an interrupt service routine is important to reduce event-processing latency, which is critical for networking, packet processing, and applications that rely on inter-process communication. However, instruction cache hit rate during interrupt processing can be low, leading to high instruction fetch latencies that stall the CPU pipeline and increase interrupt handling time. Many state-of-the-art prefetchers that generate prefetches based on previously accessed memory regions struggle with asynchronous events, such as interrupts, which disrupt the program's/CPU's execution flow. To address this issue, we propose IAIP, a novel concept that provides interrupt-aware instruction prefetching in hardware for either L1 or L2 caches to accelerate interrupt handling. For interrupts generated by inter-process communication, we observe reductions in interrupt processing latency of up to 45% for L2 prefetching and up to 50% for L1 prefetching.
14:55	#113 **HENDAL: Hardware-Assisted Early Notification for Event-Driven Applications in Linux** Lars Nolte and Tim Twardzik (Technische Universität München); Marco Liess, Thomas Wild and Andreas Herkersdorf (Technical University of Munich) When developing applications that interact with asynchronous event sources, there are two main approaches for waiting for event readiness: polling and blocking. Polling often achieves low latency by continuously checking for new events, but it consumes CPU resources. Blocking conserves CPU cycles by relying on Operating System notifications but introduces significant latency due to thread wake-ups on the critical path. We present HENDAL, a novel hardware-assisted early notification mechanism for Linux that addresses this trade-off in the context of network packet reception. Rather than relying on notifications by the kernel after packet

	inspection, HENDAL extends a Smart Network Interface Card (NIC) to detect incoming packets and wake up sleeping threads. Evaluation on an ARM-FPGA prototype shows that HENDAL reduces the latency (NIC to application thread) by up to 35% compared to traditional blocking mechanisms while remaining only 11% slower than polling for individual events at low packet rates. Crucially, HENDAL maintains the CPU efficiency of blocking approaches, providing an improved balance between performance and resource utilization. A practical evaluation with Memcached demonstrates end-to-end improvements of up to 10% on our prototype platform.
15:10	#167 **Buffer Designs for Dynamic Scheduling of Mixed-Latency Operations** Demyana Emil Mekhail Selim and Klaus Schneider (RPTU Kaiserslautern-Landau) Most processor architectures have operations with low execution latency like addition and high execution latency such as division and load/store instructions. Exploiting instruction-level parallelism in the presence of such mixed-latency operations is a challenging task because high-latency operations can cause stalling that also blocks low-latency operations. This paper presents lightweight hardware designs that use special buffers to enable dedicated processing units to efficiently execute high- and low-latency operations in parallel. Specifically, we examine two designs: The reorder buffer design collects results of both types of operations in a single reorder buffer, while the side buffer design stores only the results of low-latency operations until results of the preceding high-latency operations leave the processing unit. Our experiments demonstrate that both designs reduce average latency by effectively exploiting available parallelism in the instruction stream. Furthermore, the side buffer design reduces circuit size and power consumption while enabling higher clock frequencies.
15:25	#315 **Memory Admission and Result-Retention Sizing in an RV32IMC FPGA Soft Processor** Yihang Dai, Shuyu Xu, Yinfeng Zhou and Chenglei Peng (Nanjing University) In short FPGA pipelines the frontend and the load/store unit share block RAM, and memory admission can delay younger instructions whose arithmetic is already prepared. How long should those results remain usable? We study this question in H3, our board-tested five-stage RV32IMC soft processor, by varying admission and result lifetime on fixed binaries while tracking the same dynamic instructions. Offering data requests one edge earlier cuts geometric-mean Embench cycles by 18.2% and reduces opportunities for longer retention from 375,063 to 4,373. Admission and retention cycle gains are conditional and non-additive. Under the same slower-read service, a 99% capture target selects 18 edges with registered admission and 28 with early admission, even though none of the matched measured owners waits longer. A sizing rule declared before profiling predicts recovered opportunities in all finite tests, including on kernels outside model development, and owner-level analysis closely estimates execution-cycle savings. Memory admission thus shapes both the usefulness of retention and the lifetime selected from its residual demand.

B6: Design Automation and Low-Power Techniques for Embedded Manycore SoCs

14:40–15:40 · Room B · Chair: Jani Babu Shaik (Shanghai Jiao Tong University) · Co-chair: Sandeep Shivam (Independent Researcher)

14:40	#198 **Thermal-Aware Floor Planning Under Process Variation Using Reinforcement Learning for 3D-IC Microelectronics** Md. Ariful Islam Bhuiyan (California State University, Northridge); Dipta Roy, Kazi Mushfiq Rafid, Raj Kumar and Somnath Chattopadhyay (California State University Northridge) Heterogeneous multicore and many-core SoC integration via 3D-IC and chiplet packaging introduces severe thermal gradients and process-variation sensitivity that existing floorplanning frameworks address separately or incompletely. This paper presents ThermVar-Floor, a reinforcement learning-based 3D-IC floorplanning framework that jointly minimises thermal hotspot severity and process-variation-induced timing degradation within a unified proximal policy optimisation (PPO) agent. The reward function incorporates a compact hotspot-calibrated thermal model and an ellipsoidal process-variation uncertainty set to penalise floor-plan configurations that are thermally unsafe and variation-fragile. A PV-descending placement curriculum exploits thermal headroom before high-power blocks constrain the temperature map. Evaluated on GSRC and HotFloorplan benchmarks with TSMC 7 nm technology parameters, ThermVar-Floor reduces peak junction temperature by 12-19% and worst-case PV-induced slack degradation by 23-38% relative to simulated annealing baselines at a runtime overhead below 15% versus single-objective RL. A direct comparison with RLPlanner confirms 18-22% additional variation-induced slack improvement. The framework provides a practical co-design tool for heterogeneous 3D-IC manycore SoC systems targeting CHIPS Act-aligned manufacturing reliability requirements.
14:55	#177 **CARDAMOM: Cycle-Accurate RISC-V Dynamic Method for Power Prediction using Attention Mechanisms** Ameya Ramteke (Birla Institute of Technology and Science); Paresh Saxena (BITS Pilani); Soumya J (BITS Pilani, Hyderabad Campus); Ankit Agrawal and Shankar Ramakrishnan (Google LLC) Cycle-accurate RTL dynamic power prediction is valuable for early-stage design-space exploration, but existing machine-learning predictors are typically optimized only for accuracy and tell us little about which features actually drive power under different execution contexts. We address this gap with an attention-enhanced prediction framework for cycle-accurate dynamic power modeling in a RI5CY RISC-V core. We propose CARDAMOM, a lightweight Dynamic Feature Attention (DFA) mechanism integrated with a Multilayer Perceptron (MLP) regressor, which reweights activity features according to pipeline block and instruction context on a per-cycle basis. Evaluated against decision tree, linear SVM, random forest, and plain MLP baselines, CARDAMOM matches the best neural baseline in accuracy (RMSE of 0.0495 mW, MAE of 0.0395 mW, R2 = 0.9963) while adding only modest overhead in parameter count and inference latency. The learned attention maps reveal patterns that line up well with processor microarchitecture, including stronger switching-related emphasis in Execute and the Load Store Unit (LSU), and a greater reliance on instruction context in the Control and Status Register (CSR) block. These findings suggest that the real need in RTL dynamic power prediction is not just accurate estimation, but validated, context-aware feature prioritization, which the proposed framework provides without materially changing the accuracy and complexity tradeoff.
15:10	#268 **Low-Power FPGA Implementation of Adaptive Line Enhancers Using Reduced-Precision Arithmetic for SoC-Based IoT and Neural-Computing Applications** Hen-Geul Yeh (California State University Long Beach); Brian Pham and Ryan Chang (California State University, Long Beach) The rapid growth of Internet of Things (IoT), system-on-chip (SoC), and neural-computing real-time applications has increased the demand for energy-efficient computing platforms. Field-programmable gate array (FPGA) implementations provide a flexible approach to developing and evaluating hardware architectures by tailoring computational resources and numerical precision to application requirements. This paper

	investigates a low-power FPGA implementation strategy using an adaptive line enhancer (ALE) as a representative signal-processing front end for energy-efficient neural-computing and future SoC-based systems. The ALE is a finite impulse response (FIR) adaptive filter that employs the least mean squares (LMS) algorithm to dynamically update its filter coefficients for noise reduction. Both conventional nonlinear-phase (NLP) and linear-phase (LP) ALE architectures are implemented using double-precision (64-bit) floating-point, 16-bit Q15, and 12-bit Q11 fixed-point arithmetic. The implementations are evaluated in terms of power consumption and numerical representation, while a convolutional neural network (CNN) is trained to classify LP and NLP ALE outputs at signal-to-noise ratios (SNRs) of 10 and 20 dB. A dataset of 20,000 filtered-output samples is divided into 80% training and 20% testing sets. Experimental results show that the 16-bit fixed-point representation requires approximately 75% less storage than the 64-bit floating-point representation. Moreover, the 16-bit and 12-bit fixed-point LP implementations reduce power consumption by approximately 40% and 44%, respectively, compared with the 64-bit floating-point implementation. The CNN achieves comparable accuracy, precision, recall, and F1-score with fixed-point data while exhibiting faster convergence. These results demonstrate the potential of reduced-precision FPGA implementations as energy-efficient signal-processing accelerators for neural-computing and SoC-based IoT applications.
15:25	#239 **A Capacitance-Aware Training Framework for Decision Tree Inference Circuits** Woramet Janchum and Thanasin Bunnam (Rajamangala University of Technology Thanyaburi); Pitchayapatchaya Srikram (RMUTT) Decision trees (DTs) are widely used in machine learning models because of their interpretability and relatively low computational complexity. However, conventional DTs have limited access to related memory values. This research proposes a circuit-aware DT training method where the physical threshold constraints of capacitive decision trees are directly integrated into the DT learning process to alleviate circuit-level constraints. We simulated the proposed DT inference circuits in Cadence Spectre using a UMC 65-nm CMOS technology at a 1.0 V supply voltage. The proposed work achieves 75.21 nW power consumption and 4.481 fJ per decision. This solution overcomes memory access limitations and reduces design complexity for edge AI applications.

A7: Architectural Innovations in Embedded Multicore/Manycore SoCs – II

16:00–17:00 · Room A · Chair: TBC · Co-chair: Rashmi Prabhakar Sarode (Indian Institute of Technology Madras)

16:00	#158 **An On-Chip Supervisor Architecture for the Configuration, Diagnosis, and Debug of High-Speed Communication Datapaths** Federico I. Villar (Fundación Fulgor); Ariel Pola (Fundación Fulgor); Pablo S. Mandolesi (Universidad Nacional del Sur) High-speed transceivers embed digital signal processing (DSP) datapaths of increasing complexity, while the digital subsystem responsible for the configuration, observation, and debug is typically treated as a black box. This work presents an on-chip supervisor architecture for communication SoCs that unifies configuration, diagnosis, and debug of the datapath in a single scalable subsystem. The design separates traffic into three isolated AMBA planes (control, diagnostic data, and storage) and combines a distributed network of Diagnostic Units feeding a Central Control Unit, three Configuration Units that expose the datapath parameters as memory-mapped registers, and a four-core RISC-V supervisor with functional task partitioning. A specification-driven generation flow produces the RTL, the JSON register-map descriptors, and the generic testbenches from a single source. Instantiated as the supervisor of a 56 Gb/s PAM-4 SerDes transceiver, the design passes roughly 930 functional tests and synthesizes to about 335k standard cells on the 130 nm SkyWater PDK.
16:15	#285 **COMET: A Lightweight Rate-Portable Ethernet Remote-Memory Engine for Heterogeneous FPGA Systems** Keisuke Takano (Japan Advanced Institute of Science and Technology); Akira Uejima (Okayama University of Science); Yasushi Inoguchi (Japan Advance Institute of Science and Technology) Expanding a heterogeneous FPGA cluster can require repeated adaptation and verification of its network circuits as board and link configurations change. This paper presents COMET, a reusable communication component intended to simplify that integration by separating common PUT/GET and AXI-memory semantics from MII-, GMII-, and XGMII-specific framing. A canonical octet stream, length-driven control, and parameterized width adaptation preserve the command and address contract from 100Mb/s to 10Gb/s; a 25Gb/s configuration is also implemented as a reference. Exhaustive simulation of all 256 transfer-length codes completed with no mismatch or timeout. Hardware tests passed on all three interfaces and on the MII-GMII, GMII-XGMII, and MII-XGMII switched pairings. For 8192-byte payloads, measured goodput was 99.40Mb/s, 993.81Mb/s, and 9.83Gb/s, or 98.85-99.98% of the corresponding wire-format bounds. Four post-route configurations occupied 6,669-7,295 LUTs and 20,283-21,327 flip-flops. The results support a common communication building block whose behavior is preserved and whose performance follows the selected Ethernet rate over a 100x range, as a first step toward a reusable FPGA-cluster ecosystem.
16:30	#211 **Dual-Anchor Calibration and Design Space Exploration of the Open-Source RISC-V Core C910 for Embedded IoT Workloads** Ze Hong (The 58th Research Institute of China Electronics Technology Group); Zilin Wang, Huiwang Zhou and Zhenjiao Chen (The 58th Research Institute of China Electronics Technology Group) Architecture-level simulators are indispensable for microarchitecture design space exploration (DSE), yet their fidelity hinges on rigorous calibration against real silicon. This paper presents an RTL-anchored calibration methodology for the open-source RISC-V C910 core, targeting embedded and IoT workloads. We map all major performance-model parameters directly from the RTL source, and perform workload equivalence verification to rule out errors caused by runtime library or code-generation discrepancies. After calibration, the gem5 model achieves a mean cycle error of 9.2% across seven real benchmarks; using microbenchmarks to attribute major structural mismatches further reduces the error to 4.1%. For power, we train an XGBoost model on publicly reported measured data, attaining a 12.7% leave-one-out cross-validation error. Armed with this calibrated pair (performance + power), we conduct a DSE over 358 design points in an 8-dimensional space, using a Gaussian process surrogate (CV $R^2 > 0.987$) to efficiently navigate the trade-offs. Interestingly, the original C910 baseline is not Pareto-optimal. At equal power, performance can be improved by up to 11.6%; at equal performance, power can be reduced by 27% (energy efficiency +47%). Sensitivity analysis identifies the load queue, reorder buffer, branch target buffer, and issue width as the dominant knobs. A data-footprint experiment further clarifies the generalisability of these conclusions, showing that the observed trends hold across a broad range of working-set sizes but may saturate for extremely small or large footprints.
16:45	#224 **A Tri-mode Drop-K Sequential Backward Selection Architecture and Implementation** Hsuan-Fu Liu, Tao Jung Wang, Lan-Da Van, Hui-chun Tseng and Chen Wei Liao (National Yang Ming Chiao Tung University); Chun-Ming Huang (Taiwan Semiconductor Research Institute, National Institutes of Applied Research); Tzyy-ping Jung (UCSD)

	This paper proposes a new flexible drop-K sequential backward selection (SBS) architecture and implementation, which we refer to as a tri-mode drop-K SBS architecture, to speed up the feature selection process. The proposed architecture supports K=1, 2, and 4. Simulation results demonstrate that drop-2 and drop-4 reduce execution time by 49.5% and 74.24%, respectively, compared to the conventional drop-1 SBS algorithm. We implemented the tri-mode SBS hardware using TSMC 40nm CMOS technology process.

B7: Multicore SoCs for Cyber-Physical and Autonomous Systems – II

16:00–17:00 · Room B · Chair: TBC · Co-chair: Abdul Khadir Mahamed (New York University)

16:00	#159 **Hardware-Assisted Event-Driven Temperature Monitoring on a LiteX-Based RISC-V SoC** Daisuke Suzuki (The University of Aizu); Takahiro Hanyu (Tohoku University) The growing deployment of Internet-of-Things (IoT) devices has increased the demand for energy-efficient sensing systems. Conventional sensor nodes typically rely on periodic software polling, where processors repeatedly execute sensing tasks regardless of environmental changes, resulting in unnecessary processor activity. This paper presents a hardware-assisted event-driven sensing platform implemented on a LiteX-based RISC-V system-on-chip (SoC). The proposed platform integrates a custom TMP2 temperature sensor controller IP and programmable event detection logic into a VexRiscv-based SoC implemented on a Digilent Cmod S7 FPGA board. Four event detection modes are supported, including threshold-above, threshold-below, temperature-change-above, and temperature-change-below detection. When a selected event condition is satisfied, an interrupt request is generated to activate software processing; otherwise, processor execution is skipped. The sensor controller is integrated into the LiteX framework through the CSR and interrupt infrastructures, enabling flexible hardware/software co-design. Experimental evaluations were conducted using a physical FPGA prototype equipped with a Digilent Pmod TMP2 temperature sensor and an external power-monitoring system. The results demonstrate correct event-driven operation and successful interrupt-based processor activation. Furthermore, the effects of sampling interval on event occurrence rate and event detection latency were investigated. FPGA implementation results show that the complete platform occupies 18.1% of LUT resources and 7.6% of registers on the target device. The proposed platform provides a flexible framework for evaluating event-driven sensing techniques and serves as a foundation for future energy-efficient IoT systems.
16:15	#274 **Delaunay-Seeded Column Generation for Coverage-Constrained Sensor Augmentation** Ali Suliman Alsalehy (Jubail Industrial College) Restricting candidate sensor locations to Delaunay circumcenters cuts the discrete search sharply, but the restriction can cost physical hardware. In the ten layouts studied here, a repaired Delaunay seed needs 4.2 added sensors on average, while A dense grid exact model needs 3.1. This paper closes that gap with LP-guided column generation. Delaunay circumcenters first seed sparse regions, and a deterministic repair step keeps every sampled uncovered point feasible. The LP relaxation then prices a finite reserve grid; sites with negative reduced cost are admitted before the final integer solve. The refined master contains 49.5 active candidates on average instead of 285.1 for the full grid, an 82.6% reduction, while matching the full-grid integer sensor count in all ten layouts. The same objective match holds when the reserve grid is increased from 21 ×21 to 25 ×25 and 29 ×29. Grid-greedy needs 4.0 sensors on average, and matched-budget random placement covers 89.67% of sampled points, reaching full sampled coverage in only 13 of 2,000 trials. The compression does not translate into a runtime win at this scale: the complete refinement pipeline averages 99.95 ms versus 37.51 ms for the direct grid MILP. The measured benefit is a smaller final integer model with no observed loss in sensor count, not faster solution time on small instances.
16:30	#303 **Branch Pruning in Graph-Based Video Summarization: Fewer Branches, Lower Latency on Edge Devices** Yu Wang, Tom White and Dakota Sterling (Gonzaga University); Yujie Li (Guilin University of Electronic Technology); Michael C Meyer (Eastern Washington University) The cost of graph-based video summarizers that model relations between frames is rarely studied on embedded hardware. In our previous work, DSTG-VS processes three frame graphs (forward, backward, and omni-directional) in parallel graph-convolution branches and sums their outputs. We show that the omnidirectional graph is the result of adding the forward and backward graphs, so the number of inter-frame connections used by the three branches is no more than that of the forward and backward branches, but an extra full pass of graph convolution is required. To test whether the extra branches are worth the overhead, we remove the branches in different combinations from a simplified reimplementation of DSTG-VS and compare the accuracy and latency of these combinations. We run this ablation on TVSum, selecting checkpoints on held-out training videos instead of test videos and scoring per-annotator rank correlation, and we measure scoring-stage latency and memory on a Raspberry Pi~5 mounted on a mobile robot. For a 1000-frame input, latency is reduced 30.6% by cutting the omni-directional branch, from 412, ms to 286, ms, while mean Kendall's tau stays within 0.001 of the three-branch reference with both validation-selected and final-epoch checkpoints. Keeping only the omni-directional branch cuts latency by 43.5% but trails the reference by 0.017 in tau at the final epoch. F-score can also mislead: in our run, a widely used baseline evaluated at its final checkpoint scored below random in F-score while its rankings correlated positively with human annotations.
16:45	#86 **NAS-DNM: Neural Architecture Search for Dendritic Neuron Modeling in VA-ECMO Mortality Prediction** Bingchao Shi (Hirosaki University); Sichen Tao (Tohoku University); Ruihan Zhao (Tongji University); Hanyu Hu (China University of Petroleum); Sicheng Liu (Yantai Vocational College); Yifei Yang (Hirosaki University) Accurate risk prediction for patients receiving VA-ECMO is critical for clinical decision-making but remains challenging due to small sample size, high feature heterogeneity, and complex nonlinear relationships in medical tabular data. In this work, we propose NAS-DNM, a neural architecture search framework based on the dendritic neuron model, for VA-ECMO mortality risk prediction using 25 readily available clinical features. The proposed method treats dendritic model design as a structured architecture search problem and automatically identifies suitable dendritic configurations for the target data, thereby reducing reliance on manually designed model structures. By combining the hierarchical representation ability of dendritic computation with architecture-level adaptation, NAS-DNM is designed to capture nonlinear feature interactions in small-sample medical tabular learning. Experimental results on a clinical cohort demonstrate that NAS-DNM achieves competitive performance compared with multiple baseline models. Further analysis using decision curve analysis and SHAP interpretability indicates that the proposed model shows potential clinical net benefit within a meaningful threshold range and relies on a compact subset of dominant features for prediction. These findings suggest that NAS-DNM provides a promising neural architecture search framework for VA-ECMO mortality prediction. However, due to the limited sample size, particularly in the validation cohort, the reported

	performance may be optimistic and requires further validation on larger independent datasets.

A8: Hardware Acceleration of AI on Embedded Edge SoCs – II

17:00–18:00 · Room A · Chair: TBC · Co-chair: Somesh Rahul (Independent Researcher)

17:00	#98 **LiteBear: A Robust and Efficient FPGA-Based Bearing Fault Diagnosis System** Jianglan Zhu, Yulin Fu, Jiale Li, Longyu Ma and Chiu Wing Sham (The University of Auckland) Bearing fault diagnosis is essential for predictive maintenance in industrial rotating machinery. This paper presents LiteBear, an FPGA-based bearing fault diagnosis system that combines lightweight hardware-aware neural architecture search (NAS) with a fully pipelined streaming dataflow accelerator. The search evaluates over 20,000 candidates under explicit resource budgets and identifies a 1,574-parameter depthwise-separable CNN that operates directly on raw vibration signals. On the CWRU bearing dataset across four operating loads, LiteBear achieves 98.90% macro-F1 under clean conditions and 96.80% at -4 dB SNR. Deployed on a low-cost Zynq-7020, the accelerator completes inference in 7.23 us at 3.72 uJ per inference – 8.3x faster and 10.9x more energy-efficient than BearingPGA-Net on a larger Kintex-7, with only 0.72% F1 loss. These results demonstrate that real-time, robust bearing diagnosis is achievable on low-cost edge hardware, enabling scalable predictive maintenance for both industrial machinery and consumer-facing systems such as smart appliances and home equipment.
17:15	#108 **A Robust PYNQ-Based Runtime for Partially Reconfigurable CNN Inference on Zynq MPSoC** Yuhui Jiang, Hitesh B, Sheral Crescent Tissera and Rajesh C Panicker (National University of Singapore) Convolutional neural networks (CNNs) are increasingly used for remote-sensing image analysis and on-board satellite applications because of their strong feature extraction and target recognition capability. However, deploying CNN models on resource-constrained embedded platforms remains challenging due to limited hardware resources, accelerator management overhead, and runtime synchronization requirements. This work presents a runtime-robust partially reconfigurable CNN inference system on a Xilinx Kria KV260 Zynq UltraScale MPSoC using a PYNQ-based runtime framework. The proposed design decomposes a MobileNet-style network into layer-wise hardware modules and dynamically loads them into a shared reconfigurable region. A layer-major scheduling strategy is adopted to amortize partial reconfiguration (PR) overhead by reusing each configured module across multiple input samples. To improve execution stability, a fast-polling Direct Memory Access (DMA) synchronization strategy is introduced. Instead of relying on blocking DMA waits, the PYNQ runtime monitors DMA status registers and uses receive-side DMA completion as the indication that valid output availability. Timeout or error conditions trigger status inspection, DMA reset, and safe continuation. Experimental results on the KV260 platform show that the optimized implementation achieves an accelerator execution time of approximately 0.097 s per image and a PR-related runtime of about 0.50–0.56 s across different batch sizes. The FPGA implementation achieves 93.80% classification accuracy and a 93.06% F1-score. A 22.8-minute long-run experiment shows an average board-level power of 3.77W without an increasing power trend. These results demonstrate that PYNQ-based software-controlled PR and runtime-aware DMA management provide a practical hardware–software co-design approach for runtime-robust CNN inference on partially reconfigurable MPSoC systems.
17:30	#112 **Accelerating CNN on FPGA for Real-time Object Detection in Defense Applications** Adithya R Prabhu, Kevin Jose and Kala S (Indian Institute of Information Technology Kottayam) Edge deployment of deep learning models for real-time object detection presents significant challenges in balancing computational efficiency, model accuracy, and hardware resource constraints. This is particularly critical in defense and surveillance applications, where low latency, operational security, and independence from network infrastructure are essential requirements. This paper presents a real-time object detection system that addresses these challenges through a hardware-software co-design approach, using FPGA and ARM processor. The system employs a custom-trained YOLOv5n convolutional neural network (CNN) model to detect defense-relevant classes-person, vehicle, drone, and weapon. INT8 quantization is applied to reduce computational and memory requirements by converting weights and activations to 8-bit integers while maintaining accuracy. The model is partitioned by offloading the first two convolutional layers to a custom INT8 FPGA accelerator using High-Level Synthesis, while the remaining layers run on the ARM processor. The FPGA-ARM co-design achieves nearly twice the frame rate of the ARM-only implementation, at an operating frequency of 100 MHz. Inference is performed entirely on-board with no dependency on external compute resources or network connectivity, making the system well-suited for secure, low-latency deployment in real-time edge applications.
17:45	#202 **An FPGA-Based CNN Accelerator Architecture Using Line-Buffer Layer Fusion for ADAS Applications** Jeonghun Im (Santa Clara University); Radhika Grover (Santa Clara University & Member, Frugal Innovation Lab, Santa Clara University) This paper presents a line-buffer-based layer-fusion architecture for accelerating LeNet-5-style convolutional neural networks on FPGA platforms for Advanced Driver Assistance System (ADAS) workloads. The proposed design fuses each convolution layer with its following max-pooling layer by generating convolution outputs in the order required by the pooling operation, thereby reducing intermediate storage and avoiding separate feature-map transfers between layers, without altering the number of arithmetic operations performed. The architecture is evaluated for traffic sign recognition using a PYNQ-Z2 FPGA implementation and compared against conventional FPGA-based LeNet-5 accelerator approaches. By combining line buffering, window reuse with layer fusion, quantization, and HLS optimization pragmas, the design targets lower latency, reduced memory traffic, and improved hardware efficiency while preserving classification functionality. Our results show that fusion reduces the initiation interval from 13,096 to 3,978 cycles, a 3.29× improvement over an otherwise identical line-buffer design without fusion, and 2.66× over the fastest non-fused design that fits the device, while reducing LUT usage by 2.53× and leaving block RAM unchanged. On the board the fused accelerator classifies one image in 0.401 ms and sustains 1,959 images per second. A previously published PYNQ-Z2 LeNet-5 traffic-sign-recognition accelerator reports 20 ms per image, though its network is larger and takes a color input, so that figure is not a like-for-like comparison. The complete hardware source code, testbench, and evaluation script for this design are publicly available online.

B8: Application Domains for Embedded Multicore/Manycore SoCs

17:00–18:15 · Room B · Chair: Zhishang Wang (The University of Aizu) · Co-chair: Abdul Khadir Mahamed (New York University)

17:00	#210 **Metadata-Aware Policy Enforcement for Secure Enterprise Data Lakes** Murali Shankar Dulam (JPMorgan Chase); Darshan Mohan Bidkar (IEEE); Mythili Annamalai Sekar (Amazon Inc); Siva Kumar Chintham (LTM); Naga Surya Kamala Dattatreya Pasupuleti (JPMorgan Chase); Shiva Kumar Devalokam Subrayappa (Evernorth Health Services); Narender Reddy Bitla (JPMorgan Chase); Bikesh Kumar (Apple Inc.) Enterprise data lakes now serve engineering, analytics, machine learning, and regulatory workloads from a single substrate, so authorization must remain consistent across heterogeneous query engines, accounts, datasets, columns, and rows. Bucket policies and engine-local roles address this only partially. This paper presents a policy-centric security architecture for AWS data lakes that uses Lake Formation as a governance control plane over Amazon S3 and the Glue Data Catalog. The design separates identity, governance, metadata, and data planes; combines IAM authentication with tag-based (LF-Tag) permissions for scalable administration; and applies fine-grained row, column, and cell filters to sensitive datasets. We contribute an implementation and cross-account sharing workflow, a policy-evaluation abstraction, and four automated validation gates suitable for policy-as-code pipelines. A threat-driven analysis addresses storage-layer bypass, privilege accumulation, policy drift, confused-deputy paths, and cross-account misconfiguration. A scenario-based analytical evaluation introduces entitlement-delta analysis, which detects authorization changes whose effective blast radius exceeds their textual diff, and shows that metadata-driven policy reduces 24,000 role-table relationships to 160 top-level expressions under a representative enterprise model.
17:15	#325 **Adaptive Push-Down Query Optimization Under Multi-Cloud DoS Attacks** Nachiappan Chockalingam (Meta); Shomik Basu (Workday Inc); Sai Srinivas Choudary Potla (Innovaccer); Sumit Saha (East West Bank); Ashok Gadi Parthi (None); Mayilsamy Palanigounder (NTT Data); Akash K Agarwal (Albertsons Companies Inc); Atul Prasad (University of San Diego) Modern enterprises increasingly rely on distributed multi-cloud data integration pipelines to consolidate and process large-scale analytical workloads across heterogeneous database engines. Push-down optimization, which delegates transformation logic to native database engines rather than external extract-transform-load (ETL) orchestrators, substantially reduces data movement and query latency in single-cloud deployments. However, extending push-down optimization across multi-cloud boundaries introduces new failure modes, particularly when adversarial or degraded network conditions disrupt inter-cloud communication channels. This paper proposes a resilience-aware push-down query optimization framework, termed RA-PDO, that jointly minimizes query execution latency and data transfer cost while providing provable robustness guarantees against packet-loss and denial-of-service (DoS) interference on wide-area inter-cloud links. RA-PDO models the multi-cloud topology as a directed weighted graph, formulates push-down site selection as a mixed-integer linear program (MILP), and embeds an adaptive retry and rerouting layer inspired by resilient wide-area control theory. Simulation experiments conducted over a synthetic three-cloud topology with parameterized DoS injection rates demonstrate that RA-PDO reduces end-to-end query latency by up to 38% and cross-cloud data egress by up to 44% compared to a centralized ETL baseline, while maintaining pipeline availability above 99.2% even at attack intensities that cause a 40% packet-drop rate on primary inter-cloud links. These results establish RA-PDO as a practical foundation for secure, high-performance multi-cloud data integration in critical enterprise and infrastructure environments.
17:30	#280 **Multicore Processing for Scalable Execution Layer in Secure Online Examination Systems** Rashmi P Sarode (Indian Institute of Technology, Madras); Yutaka Watanobe (University of Aizu); Subhash Bhalla (IIT Delhi) Online examination platforms must support increasing numbers of concurrent users while ensuring reliable processing, low response time, and secure handling of examination records. Existing decentralized examination systems improve transparency and integrity through distributed ledger technology and smart contracts, but their scalability under multicore execution has received limited attention. This work extends a decentralized online examination framework by introducing a multicore processing layer for parallel examination grading and cryptographic integrity validation. The proposed approach distributes examination workloads across multiple worker processes and evaluates system performance under varying concurrency levels and worker configurations. Rather than proposing a new parallelization mechanism, this study characterizes application-level worker allocation for concurrent examination submissions that require grading and cryptographic integrity processing before trusted record commitment. The experimental analysis considers response time, tail latency, throughput, and request reliability. The results indicate that multicore parallelism can substantially improve performance under moderate and high examination workloads, while also revealing workload-dependent performance differences across the evaluated worker configurations. These findings characterize the scalability of the multicore examination-processing stage in a secure online examination architecture. The distributed-ledger record-management layer remains part of the earlier framework. However, its performance is not evaluated in the present study.
17:45	#331 **Adaptive Walsh-Hadamard Precoding and Opportunistic Reception for Reliable Uplink OFDM-NOMA** Ravinder Singh (Malaviya National Institute of Technology Jaipur); Mushtaq Ahmed (Malaviya National Institute of Technology Jaipur INDIA); Mehwish Qureshi and Aryan Pingoliya (MNIT Jaipur); Rajendra Prasad Yadav (Malaviya National Institute of Technology Jaipur) Non-Orthogonal Multiple Access (NOMA) is one of the promising multiple access schemes for 5G and beyond communications due to its high spectral efficiency. However, inter-user interference in uplink Orthogonal Frequency Division Multiplexing (OFDM)-NOMA systems can significantly affect the bit-error rate (BER), particularly for low-power-allocation users after Successive Interference Cancellation (SIC). In this paper, an opportunistic power selection combining (PSC) ap proach, together with a Walsh-Hadamard (WH) spreading-based precoding scheme, is proposed. The precoding scheme modifies the patterns of constructive and destructive interference across OFDM subcarriers, while a zero-feedback receiver selects the stronger signal between the precoded and non-precoded signals for each individual subcarrier. Four additional improvements are proposed: adaptive coding, hybrid linear combining, channel aware power allocation, and SIC with confidence measures. Simulations are performed for a system with K = 3 users and N = 64 subcarriers, employing Quadrature Phase Shift Keying (QPSK) modulation over independent and identically distributed (i.i.d.) Rayleigh fading channels. The experiments demonstrate reliable BER reductions of 0.44% to 0.61% over an SNR range of 0 to 12 dB, with a maximum per-subcarrier SNR improvement of 0.88 dB and 56.2% of the subcarriers exhibiting a positive spreading gain. The proposed technique requires no transmitter feedback and can be readily integrated into the OFDM-NOMA processing chain, demonstrating the potential of jointly exploiting spreading, opportunistic power selection, adaptive coding, channel-aware power allocation, and confidence assisted SIC to improve the reliability of uplink OFDM-NOMA systems.

18:00	#322 **Hardware-Acceleration of Advanced Precoding for mMIMO 5G/6G Systems on RISC-V vDSPs** Javier Acevedo (Technical Univierstiy of Dresden); Patrick Seeling (Central Michigan University); Frank H.P. Fitzek (Technische Universität Dresden & ComNets - Communication Networks Group) The arrival of massive Multiple Input Multiple Output (mMIMO) systems in Long Term Evolution (LTE)-A and 5th Generation Cellular Networks (5G), enabled by advanced digital and hybrid beamforming techniques, has significantly improved spectral and power efficiency by directing narrow yet precise energy beams toward individual User Equipment (UE). Nevertheless, the constantly increasing number of antenna elements in mid-range bands for commercial base stations, with configurations of 64 × 64 antenna elements, substantially augments the computing requirements to serve multiple terminals per cell simultaneously, ensuring at the same time high data rates and very low latency for current 5G and future 6th Generation Cellular Networks (6G) applications. Therefore, in this paper we study how the Reduced Instruction Set Computer (RISC)- V-based vector processors can exploit the hardware lanes to efficiently compute the state-of-the-art precoding algorithms in such a manner that they can be effectively be considered for their massive adoption in wireless modems and base stations. Concretely, we implemented the main linear precoding kernels such as the Zero-Forcing (ZF), the gold-standard Minimum Mean Square Error (MMSE) and the Large-Scale Fading (LSF), and compare it against its non-linear counterpart such as the Dirty Paper Coding (DPC). In order to validate our findings, we provide metrics about the processor clock cycle, power consumption and throughput based on a software-programmable implementation. Our results demonstrate that computing the precoding algorithms throughout matrix operation optimization turns out in gains of 21.44, 23.21, and 16.58 in comparison with the scalar processors.

Wednesday, December 16, 2026

A9: Design Methodologies for Embedded Multicore/Manycore SoCs – II

10:15–11:15 · Room A · Chair: TBC · Co-chair: Youngsoo Kim (University of Minnesota Duluth)

10:15	#135 **End-to-End Memory Design Automation: From Behavioral Specifications to Platform-Specific Implementation** Mohamed Badawy (Infineon Technologies AG & Technical University of Munich); Hari Sankar Sahasranaman (Infineon Technologies AG, Technical University of Dresden); Wolfgang Ecker (Infineon Technologies AG) Memory subsystems are critical components in modern digital designs, yet their mapping from behavioral specifications to technology-specific implementations remains largely manual and error-prone. This paper presents an end-to-end memory design automation framework that automates the complete flow from behavioral specifications to technology-aware RTL, supporting both FPGA and ASIC platforms without design rework. The framework automatically explores feasible configurations, validates implementation feasibility, and generates RTL through four sequential stages while maintaining a unified interface across all designs. We demonstrate integration with commercial EDA tools for synthesis and place-and-route. Comprehensive evaluation on Artix-7, Zynq, and 22nm ASIC validates cross-platform applicability and demonstrates that automated design space exploration reveals non-obvious Pareto-optimal trade-offs-on 22nm ASIC, 4-bank configurations achieve 30% higher frequency than single-bank designs, yet optimal column mux ratios shift across banking levels (16:1 dominates frequency, power, and congestion at low banking, while at high banking all mux ratios converge in frequency but diverge sharply in area and congestion), demonstrating that no single configuration optimizes all objectives. Our results validate that the framework enables rapid technology migration and data-driven design decisions through systematic multi-dimensional exploration.
10:30	#121 **Exploring Energy-Efficient Accelerator for Pyramidal Lucas-Kanade Optical Flow Algorithm via Processing Element Optimization** Zhifei Liu and Shogo Semba (The University of Aizu); Yoichi Tomioka (University of Aizu); Hiroshi Saito (The University of Aizu) This paper presents an energy-efficient FPGA-based acceleration framework for the Pyramidal Lucas–Kanade (PLK) optical-flow algorithm. Within a streaming processing framework, multiple RTL-level Processing Element (PE) architectures, including baseline, high-throughput, area-efficient, and hybrid variants, are developed to investigate the trade-offs among throughput, hardware resource utilization, dynamic power, and energy consumption. The proposed architectures are implemented on the PYNQ-Z2 and ZCU104 FPGA platforms under different pyramid configurations. Experimental results show that the four-layer high-throughput architecture on ZCU104 achieves 651 FPS for 640×480 images at 200 MHz, while the lightweight area-efficient architecture on PYNQ-Z2 consumes only 0.060 W of dynamic power. On the Middlebury dataset, the four-layer 32-bit ZCU104 implementation attains an Average Endpoint Error (AEE) of 0.97 pixels and an Average Angular Error (AAE) of 13.72 degrees on the Grove2 sequence, indicating slight accuracy degradation from the proposed hardware simplifications. Notably, the baseline architecture achieves the lowest total dynamic energy among the evaluated variants, whereas the area-efficient architecture minimizes instantaneous dynamic power.
10:45	#225 **Performance Estimation and Algorithmic Transformation for Filtering Accelerator using FPGAs** Youngsoo Kim (University of Minnesota at Duluth) Most image processing applications such as 2D Discrete Wavelet Transform (DWT) implementation requires a two-dimensional filtering algorithm with a flexible block-parallel architecture. This paper focuses on developing an analytical dataflow based design methodology of digital signal processing algorithms as a means to extract performance parameters in the early stages of a design. This allows designers to quickly explore the design space based on performance parameters that are continuously assessed by the impact of design parameter changes and classical transformations such as the block data parallel concept, pipelining and architectural transformations. As a result, short system development time greatly reduces overall design time which is chiefly consumed by algorithm refinement steps.
11:00	#133 **PIMScope: Augmenting Ramulator 2.0 with Command-Level LPDDR-PIM for Transformer Inference** Ting Lin and Runxi Wang (Shanghai Jiao Tong University); Jani Babu Shaik (Shanghai jiao tong university); Xinfei Guo (Shanghai Jiao Tong University) Processing-in-memory (PIM) addresses the memory-bandwidth bottleneck of autoregressive transformer decoding by placing compute inside the DRAM banks. On mobile and edge SoCs, LPDDR-PIM is a natural candidate for on-device inference. Yet no open-source command-level LPDDR-PIM backend that simulates full transformer command traces has been released. We present PIMScope, an open-source command-level

	LPDDR-PIM backend built on Ramulator 2.0, a widely-used cycle-accurate DRAM simulator. PIMScope adds native PIM opcodes, a shared-PIM-block model central to commercial LPDDR-PIM architecture, and a two-level timing model with a structural energy interface that separates JEDEC-grounded LPDDR costs from PIM-specific coefficients. A three-stage trace-lowering pipeline converts decoder-only transformer specifications into backend-executable LPDDR-PIM command streams. We evaluate PIMScope on 15 decoder-only transformer models from five families. The backend completes all traces with command matching, quantifies a 1.17-1.80x slowdown for the two-bank shared-PIM-block design over a single-bank design, and measures weight-preload overhead at 18-38% of steady-state cycles in decode versus 2-3% in prefill, pointing that weight placement and residency are critical for performance optimization.

B9: S-Session 24: Human-Centric AI and Intelligent Systems for Sustainable Industry – I

10:15–11:15 · Room B · Chair: Siva Kumar Chintham (LTM) · Co-chair: Aswathnarayan Muthukrishnan Kirubakaran (Meta Inc.)

10:15	#48 **Event Contract Drift Detection and Impact Analysis in Enterprise Event-Driven Integration Systems** Siva Rama Krishna Varma Bayyavarapu (Senior IEEE Member); Srinivasateja Songa (The Home Depot); Siva Prasad Nandi (Oracle); Krishna Suman Dara (None); Raghuram Malpe Pai (HPE Networking) Schema drift in event-driven systems occurs when event schemas evolve in ways that break downstream consumers. We address the problem of detecting such drift and its impact before production failures. We propose the Enterprise Event Contract Drift Detection Framework (ECDDF), which monitors schema registries, evaluates compatibility via a structured algorithm, and builds a consumer dependency graph to estimate which services are at risk. Evaluation uses a Python-based simulation of 20 microservices, 12 event types, and 40 consumers, with event streams generated in Python and schema evolution injected at runtime. Results show that the framework detects all tested schema change types correctly in the executed runs, identifies affected consumers with 96.2% accuracy (results averaged across 20 runs per scenario), and yields an average detection latency of 2.02 s (including a simulated registry poll delay). The framework provides dependency-aware impact visibility that schema registries alone do not offer; production monitoring provides visibility only after consumer failures occur.
10:30	#259 **Binding Drift: Latent Conflicts in Intent-Based Networks** Krishna Suman Dara, Raghuram Malpe Pai, Somesh Rahul, Mohammed Kamran Syed, Amarnath G S and Shivaprasad Gali (Independent Researcher) Intent-based networking admits a new intent only after verifying that it is jointly satisfiable with those already deployed. The verification is sound relative to a binding - the mapping from an intent's declarative scope to concrete header space - and relative to nothing else. That binding is mutable network state. The check runs once, at admission, and is never re-evaluated when the binding moves. We name the resulting failure binding drift and define it so that capacity failures are excluded by construction: a pair conflicts by binding drift only when each intent remains individually satisfiable and the pair does not. We prove two things about this class. Every such conflict is irreconcilable, so re-pushing configuration cannot resolve one. Correspondence-based assurance, meanwhile, cannot detect one at all, because it excludes from both intents' checks exactly the class on which the conflict lives. Across four topologies and 1800 trials the per-pair incidence is flat in intent-set size at roughly 0.5%, but varies fifteen-fold across substrates. At the deployment level the share holding a latent conflict rises from 5.0% at five intents to 60.0% at eighty. Only events that rewrite the attribute binding create these conflicts. An event-triggered re-admission checker costs 2.6 to 112 ms per event; because a language model only proposes scope predicates while a decision procedure issues every verdict with a witness class, extraction error costs coverage and precision rather than soundness.
10:45	#260 **CertiSense: Certified Telemetry Acquisition for Learned Anomaly Detection under a Quality Budget** Raghuram Malpe Pai, Krishna Suman Dara, Somesh Rahul, Amarnath G S, Anudeep Rentala and Shivaprasad Gali (Independent Researcher) Learned anomaly detectors increasingly consume the telemetry network devices export. Devices reduce this stream for reconstruction error, a fixed interval or a compression ratio, none of which is the detector's goal. Existing guarantees apply after a model has seen the full data, so the device still collects everything. CertiSense constrains collection itself. The operator sets a tolerable loss in detector quality, measured without labels as the change in detector output between reduced and full streams. Full-fidelity audit windows calibrate an upper confidence bound. The controller picks the coarsest sampling interval whose bound stays within budget, which holds the mean loss at 90% confidence. On 24 entities of a public server-telemetry dataset, this cuts export volume 3.6x at an AUPRC loss of 0.042. On held-out machines, where the certified interval and direct tuning differ, tuning exceeds the budget on some device in seven of eight cases and the certificate in one. Our second finding is negative. Giving each metric its own sampling rate loses more quality than uniform reduction at the same byte budget, for four allocators and four detectors, one an autoencoder. On the server data, uniform leads on the mean in all 48 cells, 36 of them significant after Holm correction. The result repeats on an industrial dataset where the test has power and holds for a detector that scores metrics separately. Bytes grow linearly with sampling rate but loss is convex in it, so an even spread wins unless a few metrics carry most of the signal.
11:00	#84 **AIQEA: An Autonomous AI-Driven Quality Embedding Architecture** Vinil Pasupuleti (IBM); Koushik K Ganeeb (Salesforce); Shyalendar Reddy Allala (Global Atlantic Financial Company); Siva Rama Krishna Varma Bayyavarapu (Senior IEEE Member); Srinivasateja Songa (The Home Depot); Shrey Tyagi (Salesforce Inc) Late-stage defect discovery continues to impose unsustainable quality costs on software-intensive organisations. Traditional QA treats quality as a terminal-phase activity, resulting in defect escape rates averaging 36.8% and fix costs escalating up to 100× relative to design-stage remediation. This paper introduces AIQEA, the Autonomous AI-Driven Quality Embedding Architecture, a domain-agnostic three-pillar framework that repositions quality assurance as a continuous, structurally-embedded property of enterprise software systems. AIQEA integrates: (i) AI-Driven Continuous Analytics for real-time defect-pattern detection, (ii) an Intelligent Data Platform for proactive data-integrity validation, and (iii) an Autonomous Testing Engine that self-generates and executes regression tests on every code commit. We evaluate the framework through five controlled simulation experiments on a 1,200-module synthetic dataset and validate on the real NASA KC1 PROMISE benchmark. Simulation results indicate AIQEA achieves defect detection recall of ~99%, reduces MTTD by ~79%, cuts production defect escape from 36.8% to 7.0%, grows test coverage to 94% and delivers ~74% cost-of-quality savings per release cycle. These results represent a simulation upper bound; real-world validation is identified as essential future work.

A10: Hardware Acceleration of AI and Neuromorphic Edge Systems

11:15–12:30 · Room A · Chair: Tee Hui Teo (Singapore University of Technology and Design) · Co-chair: Ananda Biswas (Iowa State University)

11:15	#219 **A Zero-DSP Binarized Convolutional Neural Network on MNIST Digit Recognition on FPGA** Yuxuan Jiang (University of Electronic Science and Technology of China) Convolutional neural networks (CNNs) have been dominating the field of image recognition recently, yet their enormous amount of multiply-accumulate (MAC) operations and intensive usage of multiplier hardware make them difficult to be deployed on low-cost edge devices. In this paper, we propose a completely hardware-implemented binarized CNN system for MNIST handwritten digit recognition, which runs on a Xilinx Artix-7 FPGA without using any DSP resource. We binarize all the network weights to +1 or −1 so that the conventional multiplications are converted into sign-controlled additions and subtractions, thus completely removing the dependence on DSP48E1 hard cores. We develop a hardware-aware training framework based on the straight-through estimator (STE) to bridge the gap between software training and hardware quantization constraints to reduce the accuracy degradation after deployment. Our system adopts a serial pixel-by-pixel computing architecture with on-chip Block RAM storage. We also implement a standard two-stage FSM for timing control. The system is connected to a host application on the PC through UART, which can be used to write images manually, echo inference results and output debug feature maps, as an end-to-end recognition system. The experimental data indicates that the design has a high degree of numerical similarity between hardware and software, and it is able to run at millisecond level, and its entire deployment may be carried out in entry-level FPGA chips. The low cost solution to edge AI image classification problems offered by the proposed design is low resource and low cost.
11:30	#261 **A Linux-Capable FPGA Heterogeneous SoC with an Out-of-Order LoongArch Core and NPU for Multimodel Edge Inference** Zigian Chen (Beihang University) During a local interaction session, users can switch between tasks such as handwritten-digit recognition and image-based gesture detection. Supporting this workflow on a customizable FPGA calls for efficient allocation of finite logic and memory between general-purpose processing and neural computation. We design and implement an Artix-7 heterogeneous system comprising our own multi-issue out-of-order LoongArch core, our configurable INT8 tensor accelerator, and a persistent inference runtime under Linux on the soft core. Lane-aware deployment supports 8- and 16-lane implementations, while prepared model contexts and registered constants reduce recurring request work. Across twenty Loongson-provided benchmarks, the standalone core achieves a geometric-mean openLA500-to-design cycle-count ratio of 1.97. Doubling accelerator lanes reduces gesture accelerator cycles to approximately 1/1.96 of the 8-lane count, adding 1,244 LUTs and eight DSP blocks to the complete system. On the integrated 16-lane platform, tensor offload provides 30.72× inference-stage and 1.80× application-processing speedups over the CPU backend. Registering constants yields a 3.21× speedup in median digit-model request latency relative to per-request copying.
11:45	#299 **FLEXN: M-Flexible Low-cost EXecution of N: M Sparsity for DNNs** Ozgur Ozerdem and Alex Orailoglu (UC San Diego, Computer Science and Engineering) Semi-structured sparsity has gained recognition because of its ability to preserve accuracy while remaining compatible with hardware-efficient execution. However, leveraging semistructured sparsity in practice requires a cohesive methodology that jointly accounts for hardware cost, the expressivity of the sparsity configurations, and flexibility across layers and workloads, which is yet to be established. In this work, we provide a comprehensive approach that addresses all three aspects in unison. First, we introduce Optimal N: M Routing, a design methodology that minimizes the hardware cost of supporting N: M sparsity. Second, we present Balanced N: M Sparsity, a novel sparsity scheme that maximizes expressivity. Finally, we demonstrate how these two methods can be integrated within a weight-stationary systolic array to enable flexible execution of sparsity configurations. In unison, these contributions offer a comprehensive framework for efficient and flexible semistructured sparsity in hardware accelerators. Our evaluations demonstrate up to 10.64% area and 14.14% power improvements gained by Optimal N: M Routing and up to 7.9% post-pruning accuracy improvement or 20.7% sparsity improvement gained by the flexibility and expressivity boost we deliver.
12:00	#104 **Reducing Instruction-Fetch Energy in RISC-V for Embedded AI Processing via Dynamic and Static Loop Caching** Wiebren Wijnstra, Sameed Sohail, Berend Jan van der Zwaag, Sabih H. Gerez and Amirreza Yousefzadeh (University of Twente) Embedded RISC-V processors are increasingly deployed for on-device AI inference at the edge, where energy efficiency is a primary design constraint. Instruction fetching from SRAM-based memory is a dominant source of energy consumption in these cores, accounting for over 40% of total energy in our baseline measurements. This paper presents two loop cache architectures integrated into the datapath of a RISC-V processor: a dynamic loop cache that automatically detects and caches short backward-branch loops at runtime, and a static loop cache that functions as a software-managed hot-code instruction buffer, allowing preloading of arbitrary instruction blocks during the boot sequence. Both designs are implemented in the open-source NEORV32 RISC-V processor and evaluated on a LeNet-5 convolutional neural network inference workload, synthesized on GlobalFoundries 22nm FDX+ technology at 0.5V and 250MHz. The dynamic cache reduces instruction fetches by 48.3% and total energy by 21.5%, while the static cache achieves an 83.3% fetch reduction and 35.5% total energy savings. The area overhead of both designs remains below 0.2% of the full SoC area. The complete implementation is open source.
12:15	#176 **A Low-Power Spiking Neural Network on FPGA for EEG Motor Imagery Classification** Tahir Towsif and Arash Ahmadi (Carleton University) This work presents an FPGA implementation of a spiking neural network (SNN) for four-class EEG motor imagery classification using the BCI Competition IV Dataset 2a. The primary contribution is a recurrent leaky integrate-and-fire (RLIF) model that achieves low resource utilization per inference (i.e., 13.3% LUTs, 19% BRAMs, 0.035 W dynamic power, and 612 µJ per inference) while maintaining moderate hardware accuracy (65.86%) across 9 subjects. The design was synthesized and implemented on a Basys3 Artix-7 FPGA development board, demonstrating a viable balance between accuracy and resource utilization for resource-constrained brain-computer interface applications.

B10: S-Session 24: Human-Centric AI and Intelligent Systems for Sustainable Industry – II

11:15–12:15 · Room B · Chair: Zhishang Wang (The University of Aizu) · Co-chair: Aswathnarayan Muthukrishnan Kirubakaran (Meta Inc.)

11:15	#47 **Federated NLP with Cross-Attention Transformers for Climate-Regulatory Risk Analytics in Banking** Rohit Nimmala (Bank of America); Madhushree Kumari (IEEE Senior Member); Jagrut Nimmala (Lowes Companies) When one jurisdiction changes its climate policy, the ripple effects can reach banking capital requirements in other jurisdictions. These regulatory spillovers are difficult to track. Manual document comparison does not scale to the 75 active carbon pricing programs and 20+ Basel Committee on Banking Supervision (BCBS) member jurisdictions that exist today, and simple regression models miss textual nuances. To date, no one has combined cross-attention transformers with federated learning to detect these spillovers while respecting institutional data privacy. Banks are left without a scalable, privacy-compliant way to measure cross-border climate-regulatory risk. We address this with a three-part framework. A hierarchical cross-attention encoder reads regulatory texts from multiple jurisdictions at once. A federated aggregation module built on FedProx lets banks train together without exchanging raw regulatory filings. A spillover quantification head then estimates a capital impact score for every jurisdiction pair. We test the framework on Basel III implementation documents from 15 BCBS jurisdictions paired with EU Emissions Trading System carbon pricing data from 2019 to 2025. The model reaches a 91.4% spillover detection F1-score, beating BERT-base by 11.6 percentage points and the domain-adapted FinBERT by 8.3 percentage points. Under differential privacy at ε = 8.0, the federated variant keeps 97.2% of the centralized model's accuracy. The work brings together established NLP and federated learning techniques for a new application: climate-aware banking regulation with privacy-preserving cross-border collaboration.
11:30	#316 **MAESTRO: A Multi-Agent, Evolving-memory System with Tool-grounded Reasoning and Orchestration Harness for PLC Code Generation** Mudar Sarem (Manara University); Tarek Jurdi (Tishreen University); Laya Albshlawy (Gen. Org of Remote Sensing); Ebrahim Massrie (Osmania University) Programmable logic controllers (PLCs) operate safety-critical industrial processes, so a Structured Text (ST) program drafted by a large language model (LLM) is useful only if it compiles and provably satisfies its formal requirements. Existing generate-and-repair frameworks pair LLMs with compilers and model checkers, but they let a single model write, judge, and repair its own code, and they discard what was learned once each task ends. To address this challenge, inspired by Agents4PLC, a state-of-the-art framework that automates PLC code generation without any human interaction. we have developed MAESTRO, a multi-agent framework in which only the LLM agents propose the code, and the formal tools alone decide the correctness. We evaluate MAESTRO on the verifiable ST benchmark released with Agents4PLC, counting a task as passed only when its program compiles and every formal requirement is proven. The experimental results show that on the 23-task version, MAESTRO with GPT-4o passes 22 tasks (95.7%), against 43.5% for Agents4PLC and 0.0% for LLM4PLC with the same base model. Also, on the extended 92-task version, it passes 64.1% of tasks with GPT-4o and 70.7% with DeepSeek V3.1. A paired ablation further shows that the knowledge in the playbook of the Evolving Engine raises the share of proven requirements from 0.29 to 0.83 with GPT-4o and from 0.22 to 0.68 with DeepSeek V3.1. These results suggest that MAESTRO is a promising framework that, on the 23-task benchmark (v1), outperforms Agents4PLC and LLM4PLC under the same base model. MAESTRO is available at https://github.com/TarekRidaJurdi/MAESTRO.
11:45	#251 **Anchored Chromatic Normalisation for Camera-Based Virtual Try-On: A Spectral Study of White Balance and Colour Fidelity Across Skin Tones** Somesh Rahul, Krishna Suman Dara, Raghuram Malpe Pai, Mohammed Kamran Syed and Anudeep Rentala (IEEE Independent Researcher) Virtual try-on paints a cosmetic shade onto a shopper's live camera feed, and the colour it shows has to match the product on that person. Before the shade is drawn the camera white-balances the frame, and scene-statistic white balance assumes the scene averages to grey while a try-on frame is mostly face. We built a spectral evaluation from public measured data (CIE illuminants, NPL camera sensitivities, ColorChecker reflectances and a two-layer skin model fitted to measured skin) and ran every method through one camera pipeline. On 4,608 held-out renderings per scene across six Individual Typology Angle groups, four illuminants, twelve shades and two sensors, gray-world reaches a mean CIEDE2000 error of 10.63, worse than no correction (9.01) on an average that includes two D65-based conditions. Its error grows with the skin share of the frame, and its gap between skin-tone groups is 4.05, a ratio of 1.50 [1.41, 1.61] that peaks at intermediate tones. Anchored Chromatic Normalisation (ACN) estimates the illuminant from the sclera, a landmark of known reflectance. With the sclera an exact neutral, as our model assumes, ACN comes within 0.19 of an illuminant oracle and is flat across groups. White-patch nearly matches it but fails on the lightest skin. When the sclera's colour is allowed to vary, ACN stays ahead of gray-world up to 6.3% per-channel variation and falls behind no correction beyond 4.7%. A disparity penalty adds nothing for ACN.
12:00	#277 **Adaptive Multimodal Fusion for Robust Hazard Detection on a Low-Cost Embedded Platform** Bo Wen Heng and Teo Tee Hui (Singapore University of Technology and Design); I-Chyn Wey (Chang Gung University); Chiang Liang Kok (The University of Newcastle Australia) Hazard detection systems relying on a single sensing modality are vulnerable to conditions where that modality becomes unreliable, such as poor lighting for vision or ambient noise for audio, limiting their reliability for real-world embedded deployment. This paper presents a multimodal hazard detection system combining a visual hazard-sign recognition branch and an audio hazard-sound recognition branch, both based on MobileNetV2 and statically quantized to INT8, deployed end-to-end on a low-cost Xilinx Zynq-7020 embedded platform. Rather than a fixed-weight decision-level fusion rule, which cannot distinguish a currently reliable modality from a currently degraded one, we propose a compact, learned feature-level fusion network trained on a mixture of clean and synthetically degraded visual and audio embeddings, enabling adaptive, per-instance modality weighting. The training corpus was audited for exact-content duplication prior to all experiments, and all reported results are averaged across five independent training runs to avoid overstating reliability from a single run. Single-modality ablation confirms the fusion network genuinely leverages both modalities rather than defaulting to whichever is easier, with full fusion outperforming either ablation in every one of the five runs. Across the same five runs, the learned fusion sustains 87.5–99.0% hazard-category accuracy under real adverse conditions, including background noise, occlusion, and blur, compared to a fixed-weight baseline that collapses to 17.5–53.75% under the same conditions. The complete pipeline, including on-device sensor capture, inference, and fusion, is validated on real embedded hardware, with a measured peak memory footprint of 41.15 MiB and full-pipeline latency in the hundreds-of-milliseconds range per branch.

A11: Testing, Security, and Trust in Embedded Multicore/Manycore Systems – I

13:30–14:30 · Room A · Chair: Jani Babu Shaik (Shanghai Jiao Tong University) · Co-chair: Murali Shankar Dulam (JPMorgan Chase & Co.)

13:30	#163 **Stealthy Black Hole Router Attacks in Network-on-Chip: Autoencoder-Based Detection with Adaptive DyXY Routing Mitigation** Dilshad Mohammed, Bhavya R and Adersh A (NSS College of Engineering, Palakkad); Adeela Farshana A and Syam Sankar (NSS College of Engineering, Palakkad) Network-on-Chip (NoC) architectures serve as the communication backbone of modern multi-core systems, where reliable packet delivery is critical for correct system operation. The increasing integration of third-party intellectual property (IP) components has significantly elevated the risk of hardware Trojans (HTs) within NoC routers. In this work, we investigate a stealthy Black Hole Router (BHR) attack in which malicious logic embedded within a router manipulates the credit-based flow-control mechanism to induce covert packet loss. By spoofing credit signals and overwriting buffered flits, the compromised router silently drops packets while continuing to participate in normal routing decisions, making the attack difficult to detect using the conventional rule-based monitoring techniques.To address this challenge, we propose a lightweight unsupervised anomaly detection framework based on an autoencoder trained exclusively on normal router behaviour. The model captures correlations among key router-level metrics such as virtual channel buffer occupancy, credit exchange patterns, and buffer waiting time, and identifies anomalies using a reconstruction-error threshold. Experimental evaluation on a 4×4 mesh NoC using the gem5 simulator demonstrates accurate spatial localization of compromised routers with minimal hardware and communication overhead. The proposed method can reduce the packet loss by 55%.
13:45	#51 **A-DSRM-SEC: An Agile-Infused Design Science Research Methodology for Security Artifact Development in Embedded Multicore/Manycore Systems** Jovita Nsoh (University of Houston & Google) The proliferation of embedded multicore and manycore Systems-on-Chip (SoCs) in safety-critical domains-including energy infrastructure, autonomous vehicles, and industrial control systems-has created an urgent demand for systematic, repeatable, and adversarial-aware security design methodologies. Existing approaches to hardware security artifact development remain fragmented: threat modeling, design-for-security, and verification are treated as disconnected activities, and no unified methodology accounts for the iterative, evolving nature of adversarial threats against embedded multicore platforms. This paper introduces A-DSRM-SEC (Agile-Infused Design Science Research Methodology for Security), a novel methodology that extends Peffers et al.'s Design Science Research Methodology with agile iteration cycles, adversarial threat co-evolution, and embedded-systems-specific evaluation criteria. A-DSRM-SEC structures security artifact development through five iterative phases: Adversarial Problem Definition, Security Proposition Formulation, Artifact Construction, Adversarial Evaluation, and Deployment Implication Analysis. We instantiate the methodology through the design of TrustWeave, a hardware-rooted trust verification artifact for Network-on-Chip (NoC) fabrics in manycore processors. Analytical evaluation demonstrates that A-DSRM-SEC reduces design-to-verification latency by formalizing adversarial feedback loops, and simulation-based threat modeling confirms the artifact's resilience against data snooping, hardware Trojan injection, and side-channel leakage across shared NoC resources. The methodology positions security artifact development as a first-class design science discipline within the embedded multicore domain, addressing a critical gap at the intersection of design methodology, hardware trust, and system resilience.
14:00	#288 **GovAgent: Stateful, Auditable Policy Enforcement for Enterprise AI Agents on Multicore Processors** Udaya Bhasker Mattupalli (Accenture); Ayaskant Pattnaik, Nagaraju Akkuledu Uppara, Gangadhara Raghu Kopparapu and ShravanKumar Arram; Krishnaprasad Chouty (Sophlogic Global LLC) AI agents that call enterprise tools can issue unauthorized refunds, leak personal data, or be steered by injected instructions. GovAgent is a policy decision point (PDP) that sits between agent planners and tools. It checks every proposed action against four layers: role scope, declarative rules, sliding-window state, and a risk score. Each action is then allowed, denied, or escalated to a human, and every decision is recorded in a hashchained audit log. We report a feasibility study on a synthetic, labeled workload with seven violation classes. On the workload the policy was written for, GovAgent intercepted every violating incident, against 70.5% for static rules, while 90.6% of actions needed no human. However, only 53.2% of violating actions were intercepted, because bulk scraping is caught only after about 19.0 reads. On a held-out workload with wider attack parameters, incident recall fell to 85.5%: a slower or daytime scraper evaded the risk score entirely. A per-session volume budget raised it to 97.1% and capped leakage at 20 reads. Spreading structured refunds over parallel sessions defeated session-affinity routing, and routing entity-keyed tools by entity restored detection. On the systems side, the audit log, not the policy logic, limits multicore throughput. On an 8-core SoC, removing per-record hashing from the shared critical section raised Java throughput from 3.6M to 10.6M governed actions/s. Per-core chains and a single chain of batch digests both achieved this. A decision took 0.08 μs at the median.
14:15	#320 **GHOST: GNN-powered Attacks on Hardware Accelerator Obfuscation at RTL** Srinivasa Rao Dara (IDRBT & NITW); Dipanjan Roy (IDRBT Hyderabad); Ilaiah Kavati (National Institute of Technology Warangal) Hardware obfuscation is a widely adopted defense mechanism for protecting hardware intellectual property against reverse-engineering (RE) attacks. For Hardware Accelerators (HAs), obfuscation is increasingly performed at higher abstraction levels, such as High-Level Synthesis (HLS) and Register Transfer Level (RTL), to accommodate complex datapath-intensive designs and modern design flows. Despite the growing adoption of HLS/RTL obfuscation techniques, their resilience against oracle-less (OL) attacks remains largely unexplored. This paper presents GHOST, a Graph Neural Network (GNN)-powered OL attack engine that targets the synthesized gate-level netlists of HLS/RTL-obfuscated HAs. The proposed engine represents the netlist as a graph and learns structural patterns around the key-controlled gates to predict their key values without requiring oracle access, a functional circuit, a golden reference, or the RTL description. Experimental evaluation on hardware accelerator benchmarks protected using ASSURE, KOIL, and PROTECTS demonstrates an average Key Prediction Accuracy (KPA) of approximately 71.0%. Across the evaluated benchmarks, GHOST achieves higher average KPA than the MLP baseline and OMLA, although the comparison with OMLA is subject to differences in attack settings. To the best of our knowledge, this is the first OL attack on the synthesized gate-level netlists of HLS/RTL-obfuscated hardware designs, and the first work to demonstrate that key values embedded at the RTL level remain predictable after logic synthesis.

B11: Real-Time Systems for Embedded Multicore/Manycore SoCs

13:30–14:15 · Room B · Chair: TBC · Co-chair: Raghuram Malpe Pai (HPE Networking)

13:30	#110 **Compilation of Neural Networks for Time-Predictable Multicore Processors with Local Memories** Maximilian Kirschner and Marius Kreutzer (FZI Research Center for Information Technology); Juergen Becker (Karlsruhe Institute of Technology) With the increasing use of machine learning in safety-critical, real-time domains, there is a growing need for efficient, time-predictable execution of neural networks on embedded hardware. One solution is the use of multicore processors with local memories, which can provide high performance and low latency while ensuring time-predictability. However, compiling neural networks for such architectures poses several challenges, including the static scheduling of data transfers and computations and the estimation of the overall worst-case execution time. In this paper, we present a compilation approach for neural networks targeting time-predictable multicore processors. The compiler separates the computational workload into tiles, which are computed in parallel. The computation schedule is optimized to increase parallelism and to reduce the amount of data transferred between the main memory and the local memories. We implement the presented concepts in an MLIR-based compiler that is compatible with upstream dialects and optimizations and demonstrate it in two case studies: We evaluate the resulting schedule for selected convolution layers and for a convolutional neural network (CNN). The presented compilation approach achieves a speedup of up to 8.72x for individual layers and 2.48x for the CNN with a four worker core configuration compared to a single-core vector processor.
13:45	#245 **ePMD-TT: Placeholder-based Mixed-Criticality Deployment of Event-triggered Anytime Neural Networks into Time-triggered Periodic Schedules** Matthias Stammler (Karlsruhe Institute of Technology); Felix Frombach (Karlsruher Institut für Technologie); Patrick Schmidt (Karlsruhe Institute of Technology); Federico Gavioli (Universita di Modena e Reggio Emilia); Paolo Burgio (University of Modena and Reggio Emilia); Jürgen Becker (Karlsruher Institut für Technologie) Fulfilling timing guarantees for machine learning-based tasks in centralized automotive execution units necessitates eliminating or bounding interference. The main reason for this interference lies in shared resource contention. When tasks are executed in an event-triggered manner, either a careful scheduling approach, or a pessimistic estimation of the runtime is necessary, making a schedulability analysis for event-triggered tasks necessary. Phased time-triggered execution can omit these tests by generating a fixed schedule, allowing only one task to be in a memory access phase at the same time but they lack the ability to react to spikes of released sporadic jobs. Our work combines event-triggered ML-based tasks and time-triggered periodic tasks. First, a time-triggered schedule is created under the assumption that the event-triggered tasks are released periodically, allowing for at least one execution of the event-triggered task per hyperperiod. During runtime, the periodic part of the schedule is executed and the event-triggered tasks are executed not accessing shared resources blocked by periodic tasks, making use of early exits to adapt the runtime. We evaluate our framework with three periodic task sets with increasing utilization and three release characteristics for the ML-based tasks. The evaluation shows that an increase in ML-based task releases and a higher utilization leads to an acceptable drop in accuracy and earlier chosen exits, as expected. A case study completes the validation.
14:00	#302 **WARP - A Runtime Wrapper Architecture for Task and Data Parallel Workloads** Ananda Biswas, Manish Manepalli and Akhilesh Tyagi (Iowa State University) Edge computing platforms are typically limited by cost, energy, silicon area, and payload constraints. The demand for Artificial Intelligence (AI) on the edge further worsens the issue. AI workloads often require substantial computational throughput, memory bandwidth, and data movement. The efficient utilization of resources and careful balancing of computational and memory resources is therefore required to support these workloads alongside conventional applications. Chiplet architectures have gained popularity because they improve manufacturing yields and enable highly modular designs. Die-to-die interface has been standardized by Universal Chiplet Interconnect Express (UCIe) which has promoted interoperability between chiplets and design scalability. This combination presents an opportunity to design specialized edge computing platforms by combining multiple small, low-power chiplets optimized for efficiency and targeting intended applications. In this work, we propose a UCIe connected chiplet-based System-on-a-Chip (SoC) architecture implemented as low-power RISC-V cores with private caches and organized as nodes of k compute cores with a central wrapper core that wraps the computation of the compute cores by mediating the program state. We evaluate the performance and scaling capabilities of a single node of connected chiplets on lightweight AI applications executed in a distributed fashion. We demonstrate that the compute core count on each node can be scaled to achieve up to a 7x performance gain, beyond which further scaling yields diminishing returns.

A12: Testing, Security, and Trust in Embedded Multicore/Manycore Systems – II

14:30–15:30 · Room A · Chair: TBC · Co-chair: Murali Shankar Dulam (JPMorgan Chase & Co.)

14:30	#222 **Fortified-Edge 6.0: Adversarial Analysis and Defense of Machine-Learning-Assisted PUF Authentication Systems** Jayaditya Vetsa and Sai Venkat Saketh Veerapaneni (University of Texas at Austin); Seema Gangaiah Aarella (Austin College & University of North Texas); Naga Surya Kamala Dattatreya Pasupuleti, Narender Reddy Bitla and Murali Shankar Dulam (JPMorgan Chase) Physical Unclonable Functions (PUFs) provide lightweight hardware security for authentication, but environmental noise in challenge-response pairs (CRPs) requires error correction to maintain reliability. Fortified-Edge addresses this challenge through an integrated security framework that uses machine learning-based response recovery to improve authentication reliability. This work evaluates the security implications of such correction mechanisms under an active response-manipulation attacker with partial challenge and response visibility. A ridge linear surrogate guides targeted response modifications against visible and unseen challenge classes, achieving a maximum targeted forgery success rate of 20.03% and substantially outperforming random modifications. To mitigate these attacks, we introduce a lightweight rejection gate that distinguishes legitimate responses from adversarial forgeries. The defense reduces targeted forgery success to below 1% while maintaining over 99% legitimate acceptance. These results demonstrate vulnerabilities in machine-learning-assisted PUF correction and show that lightweight rejection mechanisms can effectively mitigate targeted forgery attacks.
14:45	#203 **An ESP32 SoftAP Testbed for Authorized Wi-Fi Security Evaluation Under Single-Radio Constraints**

	Unais Ali and Jared Oluoch (Eastern Michigan University) Rogue access points and Evil Twin attacks exploit user trust in familiar Wi-Fi network names and remain relevant threats to IEEE 802.11 environments. This paper presents the design and implementation of an ESP32-based soft access point (SoftAP) testbed for authorized Wi-Fi security evaluation under single-radio constraints. The platform integrates network reconnaissance, Controlled Service Set Identifier (SSID) impersonation scenarios, local captive-portal delivery using synthetic test inputs, and browser-based administration on 2.4 GHz hardware. A Free Real-Time Operating System (FreeRTOS)-based control plane uses asynchronous commands, task separation, and temporary off-channel scheduling with management-SoftAP restoration logic intended to support operator access during controlled radio experiments. The design is explicitly limited to owned or authorized test networks and does not claim reliable operation against mandatory Protected Management Frames (PMF), WPA3-only deployments, or native 5 GHz targets. We report release-build resource occupancy, configured firmware limits, and exploratory authorized laboratory observations from selected configurations of scenarios S1-S6, with Advanced Twin observations reported separately in Table VII. These observations demonstrate task execution and firmware-counter availability; they do not establish statistical performance, tested operating limits, or detector accuracy. Dynamic power measurements and at least ten independent trials per configuration remain for follow-up work. The resulting platform provides a low-cost embedded testbed for security education and for controlled evaluation of rogue-access-point and Evil Twin detection mechanisms. Index Terms-ESP32, embedded IoT security, Wi-Fi SoftAP, FreeRTOS, rogue access point detection, authorized security testing
15:00	#319 **Large-Language-Model-Aided Intent Verification via Grounded Obligation Models** Surinder Sood (University of Manchester & ARM); John Goodacre (The university of manchester); Lucas Cordeiro (University of Manchester) Engineering specifications remain largely prose, making operator intent difficult to verify and enforce. We present SECCOM, which extracts feature catalogs and labelled transition systems (LTSs), decomposes natural-language intent into source-grounded obligations, and uses NuSMV to check evidence on a neutral Spec LTS and satisfaction on a restricted Intent LTS. Its grounded-obligation layer ties each temporal atom to source evidence and reports absent catalog, model, or runtime support as a named GAP rather than a vacuous success. From four hardware standards, SECCOM extracts 8,173 features and 7,023 state machines. Audit-guided repair raises reference coverage from 55.3/65.2/42.4% to 97.4/87.0/100.0% for Transmission Control Protocol (TCP)/Universal Serial Bus (USB) 2.0/Compute Express Link (CXL) 4.0, respectively. The repaired TCP catalog achieves 91.7% strict and 95.5% lenient precision; normal-BM25 development-slice audits give 95.0/97.5% for USB and 87.5/100.0% for CXL, while their coverage values remain intent-scoped. Two intents exercise A3 decomposition in depth; a ten-intent, reference-grounded verification benchmark over five specifications grounds and verifies 30/34 atoms and reports the remaining four plus one clause-level coverage gap. A diagnostic comparison with nl2spec and NL2LTL shows that proposition grounding and parser coverage remain substantial limitations for the baseline outputs. Eight QEMU scenarios exercise USB, PCI Express (PCIe), and CXL actions; the expected verdicts are obtained after three runtime-model faults are repaired. The results establish traceable failure reporting for safety and reachability properties in heterogeneous platform specifications and an emulated Linux runtime
15:15	#52 **IGPO-SoC: Identity-Governed Privacy Orchestration for AI Acceleration on Embedded Multicore Systems-on-Chip** Jovita Nsoh (University of Houston & Google) Privacy-preserving AI inference on embedded multicore Systems-on-Chip (SoCs) increasingly relies on techniques such as differential privacy (DP), trusted execution environments (TEEs), and homomorphic encryption (HE). This paper introduces IGPO-SoC (Identity-Governed Privacy Orchestration for SoC), a hardware-software co-design architecture that elevates identity to the control plane of privacy enforcement on embedded multicore platforms. IGPO-SoC extends Zero Trust principles to the SoC fabric by binding cryptographic workload identities to dynamically selected privacy-preserving techniques through a hardware-enforced policy engine. The architecture comprises three novel components: (1) a lightweight hardware Identity Attestation Unit (IAU) that verifies workload identity at core-assignment time, (2) a Privacy Policy Engine (PPE) that maps identity-trust tuples to privacy technique configurations in real time, and (3) an Adaptive Technique Multiplexer (ATM) that routes inference data through DP noise injection, TEE enclaves, or HE accelerator pipelines based on PPE directives. We formalize IGPO-SoC using A-DSRM-SEC (Agile-Infused Design Science Research Methodology for Security) and evaluate the architecture through analytical modeling and cycle-accurate simulation on a 16-core RISC-V SoC model. Results demonstrate that identity-governed adaptive privacy selection reduces inference latency by 31-47% compared to static privacy enforcement while maintaining equivalent privacy guarantees (ε-differential privacy bounds and TEE attestation integrity). The architecture incurs less than 3.2% area overhead and achieves 2.1× energy efficiency improvement over static HE-only configurations. IGPO-SoC establishes identity-as-control-plane as a foundational principle for privacy-preserving AI acceleration, bridging a critical gap between Zero Trust governance and embedded hardware privacy enforcement.

B12: Multicore SoCs for Cyber-Physical and Autonomous Systems – I

14:30–15:30 · Room B · Chair: Jani Babu Shaik (Shanghai Jiao Tong University) · Co-chair: Nachiappan Chockalingam (Meta)

14:30	#308 **Lightweight Visual SLAM for Nano-Drones with a Multi-Core RISC-V SoC** Hiroto Sugai (Tsukuba University); Keisuke Sugiura (University of Tsukuba); Yoshiki Yamaguchi (University of Tsukuba & Kumamoto University); Kentaro Tanaka (NTT Corporation); Toshihito Fujiwara (NTT & Access Network Service Systems Laboratories); Tatsuya Fukui (NTT Corporation & Access Network Service Systems Laboratories); Keita Takahashi (NTT) Nano-drones such as the Bitcraze Crazyflie localize themselves with a downward optical-flow sensor and a single-point time-of-flight (ToF) rangefinder; without loop closure, the position error of this estimator accumulates with the distance traveled. We propose a localization method for nano-drones with limited computing resources that integrates a single downward-facing monocular camera for two purposes. The method closes loops by matching image features against stored keyframes and correcting the trajectory on a pose graph, and it estimates the horizontal velocity from images of the floor; both results are fed back to an extended Kalman filter (EKF). In simulated environments without yaw drift, the method reduced the median final drift of a square path by 67% compared with the same EKF without the camera, a reduction that comes from loop closure; with a yaw drift of 1.4 degrees/min, only the configurations with loop closure completed the flights. We implement the feature extraction, matching, loop detection, and pose-graph optimization on the GAP8 system-on-chip of the AI-deck, an expansion board of the Crazyflie, distributing the image processing over the eight-core RISC-V cluster of the GAP8 (5.3 times faster than one core) and keeping the keyframes in external memory, because the on-chip memory of the GAP8 has only 512 KB. In five real-world flights, the implementation ran at 7.3-9.3 fps and accepted 3-4 of 18-39 loop candidates per flight.

14:45	#275 **Real-Time Vision Guided Walking Control for Humanoid Robots Using Detection and Tracking** Zhishang Wang, Hiromu Nagai and Khanh N. Dang (University of Aizu) Humanoid robots require continuous visual feedback to adjust their motion during walking. This ability is essential for path planning, since a planned path must be continuously corrected based on the target position and humanoid robot motion. However, frequent deep object detection introduces high computational cost on resource-constrained onboard platforms. This makes real-time visual feedback difficult during physical motion. To address this limitation, this paper presents a realtime vision framework for humanoid robot walking using object detection and lightweight tracking. YOLO is used for target detection and periodic re-detection, while MOSSE tracks the target between detector executions. Using this visual feedback, the humanoid robot performs head alignment, turning, and forward walking toward the target. Evaluation is conducted on a physical humanoid robot equipped with a Raspberry Pi 5 through experiments covering perception workload, head alignment, body alignment, and closed-loop target approach. In the stationary workload experiment, the proposed YOLO-MOSSE framework achieves about 22.7 FPS, approximately 4.6 times that of continuous YOLO, with a tracking latency of only 0.37-0.42 ms. During closed-loop target approach, it maintains 21.25-22.59 FPS with approximately 22% system CPU utilization, compared with approximately 78% for continuous YOLO and 70-81% for YOLO-KCF. These results demonstrate that the proposed detector-tracker framework provides frequent visual feedback with low processing overhead for vision-guided humanoid walking. This serves as a practical basis for future path planning and autonomous navigation.
15:00	#129 **LAS-LiDAR: A Modular Linear LiDAR Architecture for Embedded Passage Monitoring** Wided Hammedi (Orange); Franck Roudet (Orange Labs); David Excoffier (Orange Research) This paper presents LAS-LiDAR (Linear Array Sensing LiDAR), a modular linear sensing architecture for real-time embedded passage monitoring. The system relies on configurable sensor strips composed of independently operating distance sensors. In the evaluated configuration, three strips with 12 TFMini modules each are arranged at 45° and connected through a distributed multi-MCU architecture combining STM32F103 acquisition nodes, centralized processing on an STM32WB55 MCU, and CAN-based communication. To reduce inter-sensor interference and power consumption, only three sensors are activated simultaneously, with adaptive sampling between 10 Hz and 100 Hz depending on scene activity. The embedded pipeline performs calibration, thresholding, temporal-depth analysis, multi-object delimitation, and point cloud generation. Experimental results show an average inference time of 15 ms per object, an end-to-end latency of 50 ms, and an energy consumption of 167 Wh over 24 hours. Analytical comparisons indicate potential energy savings of 59% to 83% relative to conventional LiDAR systems. These results support the relevance of LAS-LiDAR for modular, low-power, and privacy-preserving embedded sensing.
15:15	#281 **A Minimal Fixed-Function Triangle Rasterization Pipeline: Edge-Function Rasterization and Incremental Depth Testing on FPGA** Jiaqi Chen and Teo Tee Hui (Singapore University of Technology and Design); Chung-Yi Li (Chang Gung University); Chiang Liang Kok (The University of Newcastle Australia) Design guide for hardware implementation of triangle rasterization will be presented in the paper. The rasterization stage is introduced by building it from the basic formulas, so that the reader can follow the path from theory to the circuit. This paper starts with the general algorithms, the edge function, the top-left fill rule, and the barycentric depth interpolation. These formulas are then rewritten into fixed-point and hardware-friendly forms. After this algorithmic understanding, the design is mapped onto RTL modules, in which a setup module performs all the per-triangle computation and the rasterizer becomes an incremental scanner with no per-pixel multiplication or division. To implement it, the fixed-point formats, the packed depth buffer, and the depth test that does not interrupt the scan are introduced. Finally, the design is verified against a software reference model and evaluated on a Zynq-7020 device.

A13: S-Session 2: Emerging Machine Learning and Deep Learning Models: Theory and Applications

15:50–16:50 · Room A · Chair: TBC · Co-chair: Md Farhadur Reza (Eastern Illinois University)

15:50	#122 **Personalized Ontology-Driven Reasoning for Depression Detection on Social Media via Large Language Models** Zheqi Shen (The university of Aizu); Incheon Paik (The University of Aizu) Depression is a common mental disorder, yet timely clinical screening remains constrained by stigma and limited access to mental health resources. In our previous work, we proposed an ontology-driven framework integrating a formal depression ontology with Large Language Models (LLMs), achieving 88.0% accuracy on a 200-user Twitter dataset. However, that framework relied on globally fixed behavioral thresholds and produced only binary classifications, leaving individual behavioral differences and clinical severity gradations unaddressed. This paper proposes a personalized ontology-driven reasoning framework that extends our prior ontology with three targeted enhancements: (1) a user-level behavioral baseline mechanism that replaces fixed indicator thresholds with personalized z-score deviation detection; (2) a continuous risk score R(u) ∈ [0, 100] with three-tier risk stratification; and (3) a dynamic three-phase sliding-window temporal verification strategy that distinguishes clinically persistent symptoms from transient emotional disturbances. On an extended dataset of 500 Twitter users, the proposed framework achieves 91.2% accuracy and 93.8% precision; under identical experimental conditions, this corresponds to gains of +3.6 and +4.5 percentage points respectively over the reproduced original framework. The framework provides fully traceable reasoning chains linking digital behavioral signals to formal clinical symptom criteria, supporting explainable AI-assisted mental health monitoring.
16:05	#127 **Formalizing Ethical Adaptation: A PEFT Framework with Rubric-Based Evaluation for LLM** Kumuduni Ranasinghe and Rumesh Weerakoon (University of Aizu); Incheon Paik (The University of Aizu) With the rapid evolution of artificial intelligence (AI), Large Language Models (LLM) have become integral to numerous daily applications. Despite their impressive performance, LLMs often exhibit ethical inconsistencies and bias, largely influenced by the nature of their training data and the trained methods. This research investigates the need for improved ethical alignment in LLMs through a novel approach. A primary data set is systematically identified through multi corpus density evaluation. Subsequently, this study evaluates how the additive based, reparameterization based and specification based methods applied via a layer wise approach affect the ethical inferences of the model. Additionally, a weighted rubric based evaluation framework is proposed for assessing ethical reasoning quality. The evaluations results highlight the performance of LoRA, Adapter tuning, and Masking method are 83.39%, 82.66% and 65.87% respectively. The findings suggest that ethical inference could be enhanced using the use of LoRA and the implemented rubric provides a reproducible, robust framework for prompt based evaluations of the ethical language model.

16:20	#165 **A Diffusion-Augmented 3D Convolutional Framework for Multi-Class Dementia Screening** Md Farhadur Reza (Eastern Illinois University) Dementia classification from structural MRI remains challenging due to severe class imbalance, with dementia cases comprising as few as 1.5% of clinical datasets. While generative adversarial network (GAN)-augmented convolutional neural networks (CNNs) have shown promise, achieving 100% dementia recall in prior work, they exhibit critical limitations: poor mild cognitive impairment (MCI) recall (11.1%), low dementia precision (25%), and mode collapse in synthetic minority-class generation. This paper proposes a fundamentally different generative framework that replaces GANs with a conditional latent diffusion model (LDM) for high-fidelity synthetic MRI synthesis, integrates a lightweight 3D CNN to capture true volumetric biomarkers, and employs focal loss to focus on hard-to-classify minority examples, particularly MCI and dementia. Evaluated on the Mayo Clinic Study of Aging dataset with severe minority-class scarcity (86.6% healthy, 11.9% MCI, 1.5% dementia), the proposed framework achieves 50.3% MCI recall (353% relative improvement over GAN-CNN baseline), 61.6% dementia recall, and 58.7% dementia precision, while maintaining 87.95% best test accuracy and healthy recall above 90%. The latent diffusion model generates diverse, anatomically consistent 3D volumes for minority classes without mode collapse, the 3D CNN captures hippocampal and ventricular shape features inaccessible to 2D architectures, and focal loss effectively up-weights minority samples, directing gradient updates toward misclassified MCI and dementia cases. This work demonstrates that diffusion-based augmentation fundamentally outperforms GAN-based methods for severely imbalanced medical imaging tasks, establishing a new state-of-the-art for early dementia screening.
16:35	#205 **Dependency-Aware Thread Optimization in Multicore Systems** Ravi Meena (Malaviya National Institute of Technology Jaipur); Arjun Chaurasiya (Malaviya National Institute of Technology, Jaipur); Mushtaq Ahmed (Malaviya National Institute of Technology Jaipur INDIA) Selecting an appropriate number of threads is an important performance optimization problem in multicore systems because execution time depends on workload characteristics, thread configuration, and the available degree of parallelism. Inter-iteration data dependencies can further affect how workloads respond to increasing thread counts. This paper presents a dependency-aware machine learning framework for predicting execution time and selecting OpenMP thread configurations for parallel C++ workloads. A parameterized dataset containing both dependent and non-dependent workloads is generated by varying operation type, datatype, loop iterations, input size, number of variables and arrays, and candidate thread count. Execution time is used as the regression target, while dependency status and thread count are included as input features. For a given workload, execution time is predicted for multiple candidate thread configurations, and the configuration with the minimum predicted execution time is selected. Random Forest, Gradient Boosting, and XGBoost regression models are evaluated using four train–test splits. XGBoost achieves the strongest overall performance, with an RMSE of 0.03819, a MAPE of 12.28%, and an R^2 value of 0.9697 for the 90:10 split. The results demonstrate that dependency-aware machine learning can effectively model OpenMP execution performance and support automated thread configuration selection.

B13: EmbodiCore Global Summit on Embodied AI, Collaborative Humanoid, and Distributed Robotic Computing

15:50–16:50 · Room B · Chair: TBC · Co-chair: Tejas Pravinbhai Patel (Amazon)

15:50	#290 **Design and Pipelined Implementation of an Artificial Neural Network for XOR Logic** Yihang Liu and Teo Tee Hui (Singapore University of Technology and Design) Although Artificial Neural Networks (ANNs) have revolutionized high-level machine intelligence, bridging abstract numerical algorithms with efficient, low-power, and high frequency Integrated Circuit (IC) implementations remains a critical barrier in both edge-AI engineering and digital IC design education. This paper presents an end-to-end, education-oriented, and fully explainable hardware-software co-design methodology for a Multi-Layer Perceptron (MLP) accelerating the non-linearly separable Exclusive-OR (XOR) logic function. Leveraging software AI models in Python to guide and automate weight and bias convergence via Backpropagation, we map continuous parameters into an optimized fixed-point two's complement arithmetic system (Q3.2 representation for parameters, with bounded intermediate word lengths from Q5.2 to Q8.7/Q3.10 to guarantee overflow immunity). To eliminate transcendental calculation overheads, a Piecewise Linear (PWL) Sigmoid approximation logic is synthesized. To resolve the combinational propagation delay bottleneck inherent in unrolled deep logic (6.0 ns critical path), we propose a 2-stage pipelined microarchitecture. Cycle accurate simulation and synthesis via Icarus Verilog demonstrate that inserting inter-stage D-Flip-Flops balances the execution path into two 3.0 ns stages, doubling the maximum operational clock frequency (Fmax) from 166.67 MHz to 333.33 MHz with a deterministic latency of 2 clock cycles. Complete waveforms, error bound evaluations, and gate-level trade-offs are presented, establishing an accessible and reproducible educational blueprint for beginners' neural silicon realization.
16:05	#247 **A Parameter Sensitivity-Based Secure Audio Encryption Framework for Multimedia Communications** Adel Nadhem Naeem (Shatt Al-Arab University College); Ali Abdulqader Bin-Salem (Universiti Malaysia Sabah); Mahmood A. Al-Shareeda (Southern Technical University) In the field of encryption for networked multimedia audio communications, maintaining a balance between transmission security and computational efficiency in high-concurrency scenarios is a common research need across the field. The parameter-sensitivity-based selective encryption schemes proposed in previous studies reduce computational overhead by encrypting only sensitive audio frames, but these schemes have core flaws: weak resistance to integrity attacks and insufficient defense against replay threats. This paper puts forward a parameter-sensitive audio encryption framework that integrates AES-256 encryption logic, adding three new mechanisms: dynamic session key generation, SHA-256 integrity verification, and timestamp-sequence-number-based replay protection. Verified through controlled experiments, the new framework only causes a small increase in encryption and decryption time, retains the original high audio quality, and simultaneously achieves three core security improvements: confidentiality, integrity, and authenticity.
16:20	#266 **DT-NHURO: A Real-Time Digital Twin Framework for Bidirectional Control and Actuator-State Monitoring of a Neuromorphic Humanoid Robot** Yerima Williams (The University of Aizu & Allegis Group/Woven by Toyota); Zhishang Wang and Khanh N. Dang (University of Aizu) Digital twins are increasingly used in robotics for real-time monitoring and control. However, high-level command and status information provide limited insight into actuator activity during humanoid motion. This paper presents DT-NHURO, a real-time digital twin framework for bidirectional control, state feedback, and actuator-state monitoring, implemented and evaluated on the Neuromorphic Humanoid Robot (NHURO) platform. The framework integrates a monitoring interface, Raspberry Pi control bridge, and NHURO SDK through TCP/IP. A Bidirectional Command-Feedback (BCF) algorithm coordinates motion execution and state

	feedback, while an Actuator-State Mapping (ASM) algorithm maps reported servo activity to monitored actuator states. Evaluation of NHURO through DT-NHURO over five sessions comprising 81 physical motion executions achieved 100% observed execution success, with mean execution times of 2.913 s for WALK and 6.802 s for WAVE motions. During WAVE, reported servo activity was consistently mapped to the corresponding digital actuator states.
16:35	#298 **An Event-Driven Agentic AI Architecture for Autonomous Error Recovery and Self-Evolution in Goal-Oriented Embodied Warehouse Robots** Nagaraju Akkuledu Uppara and Gangadhara Raghu Kopparapu; Udaya Bhasker Mattupalli (Accenture); Ayaskant Pattnaik and ShravanKumar Arram; Pralav Prasun (N2 Services Inc) Automated embodied systems rely on deterministic control software, which is fast and predictable but fails on situations its designers did not enumerate, while generative models such as large language models (LLMs) can handle such situations but cost seconds of latency per decision. This paper describes an event-driven agentic AI architecture for autonomous error recovery in warehouse mobile robots. It is a hybrid of analytical and generative reasoning: a deterministic rule-based loop acts, and a generative engine is consulted only as an exception path. Uncovered events are published as schema-validated fault messages to an asynchronous Generative Optimization Engine, which proposes a remediation, has it evaluated, and installs an accepted one in the running rule table, so the rule-based system evolves at runtime. The architecture composes two established patterns under an explicit authority contract: the engine only selects and parameterizes pre-verified maneuvers, and a deterministic protective stop always precedes it. A discrete-event simulation with a mock generative engine and sampled latencies (200 seeds of 500 events) illustrates the control flow. The architecture needed 40 generative calls per 500 events, against 218 without rule installation and 730 when the engine was consulted on every event. Rule installation cut the fault rate from 13.0% in the first 100 events to 4.1% in the last 100. When 10% of generated fixes were wrong, 15% of runs ended with a wrong rule installed; an on-robot monitor with a 90% catch rate reduced this to 1%. These figures describe queuing behavior under stated assumptions, not real-model accuracy.

Thursday, December 17, 2026

A14: Testing, Security, and Trust in Embedded Multicore/Manycore Systems – III

09:00–10:00 · Room A · Chair: Siva Kumar Chintham (LTM) · Co-chair: TBC

09:00	#311 **A Framework for Test Program Automatic Generation Based on Large Language Model** Yang Yue, Heshan Yu, Zhenyu Wang and Zouqin Wang (Tianjin University); Zhaohui Guo (School of Microelectronics, Tianjin University); Qiang Liu (Tianjin University) The increasing integration of multiple processing cores and heterogeneous functional blocks in modern system-on-chips (SoCs) leads to more diverse test requirements and increasingly complex test program development for automatic test equipment (ATE), thereby placing greater demands on development efficiency. To address this bottleneck, this paper proposes a framework for automatic test program generation based on large language model (LLM) designed to significantly enhance test efficiency. The framework is composed of two modules. The information retrieval module parses the test plan and consolidates the parameter information required by each test step from explicit specifications, distributed test plan content, and tester-specific knowledge to construct generation instructions. The test program generation module uses the instruction-tuned LLM and a merger to generate correct and complete ATE test programs. The efficacy of instruction tuning is evaluated by measuring the LLM's performance on the task of test program generation. The outputs generated by the instruction-tuned WizardCoder-15B achieve scores of 0.989, 0.987, 0.988, 0.981, and 0.964 on metrics ROUGE-1, ROUGE-2, ROUGE-L, token F1, and Ratcliff-Obershelp similarity, respectively. The result demonstrates the effectiveness of the proposed framework.
09:15	#116 **Custom RISCV Instructions for Reducing Size of Software-Based Self-Test** Jad Al Halabi (Infineon Technologies AG); Endri Kaja (Infineon Technologies AG & Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau); Ares Tahiraga, Florian Thienel and Wolfgang Ecker (Infineon Technologies AG) The increasing adoption of microchips in safety-critical applications, coupled with aggressive miniaturization, has exacerbated manufacturing defects, including transient bitflips and permanent stuck-at faults. Software-Based Self-Testing (SBST) offers a cost-effective solution for CPU-based chips by leveraging on-chip resources to execute instruction-based programs for fault detection. Applicable across the chip's lifecycle from production to in-field testing, SBST requires minimal Design-for-Test (DFT) infrastructure. However, the large size of test programs incurs significant costs: in-field applications suffer from limited memory capacity, while production testing faces prolonged application times due to the repeated loading of large downloadable tests. To address these limitations, this paper proposes a custom RISC-V Instruction Set Architecture (ISA) extension featuring six multi-cycle instructions for efficient bulk access to General Purpose Registers (GPRs), Control and Status Registers (CSRs), and memory. By enabling the access of multiple registers through a single instruction, the proposed extension drastically reduces the instruction count and memory footprint required for SBST. The approach is validated through a comprehensive evaluation of the RISC-V register file for stuck-at-0 and stuck-at-1 faults. Results demonstrate that the proposed method achieves high fault coverage while reducing memory usage up to 97% and incurring an area overhead equivalent to only 775 NAND2 gates.
09:30	#120 **Selectable PE-Level Safety Mechanisms for an Open-Source Chisel AI Accelerator** Felix Frombach (Karlsruher Institut für Technologie); Georgios Sotiropoulos, Tanja Harbaum and Juergen Becker (Karlsruhe Institute of Technology) Artificial Intelligence (AI) accelerators are increasingly integrated in safety-critical embedded systems such as autonomous vehicles, medical devices, and aerospace platforms. This hardware must tolerate permanent faults from manufacturing and aging defects, which can corrupt inference outputs without an externally visible signal and cause functional-safety violations under standards such as ISO 26262. Many open-source AI accelerators are written in Chisel, which generates parameterized hardware from a single description and lets the designer swap in alternative protection mechanisms at synthesis time. Yet, current Chisel-based accelerator generators lack fault-detection and fault-correction primitives at the Processing Element (PE) level, and the coverage-versus-overhead trade-off for candidate mechanisms is not characterized. In this work, we make PE-level safety a selectable property of Pelta, an open-source weight-stationary Chisel systolic array that we develop. A single Multiply-Accumulate (MAC) PE source elaborates, through one configuration

	choice, into a plain baseline PE, a parity-protected PE, or a triple-modular-redundancy PE. ParityPE adds single-bit parity on the accumulator, and the Triple Modular Redundancy PE (TMRPE) triplicates the PE with a bitwise majority voter. Both mechanisms detect every effective injection by construction in an exhaustive single-bit stuck-at campaign on the accumulator. On a Xilinx Artix-7 Field-Programmable Gate Array (FPGA) target, ParityPE adds 6% LUTs and reduces fmax by 25%, while TMRPE adds 271% LUTs at baseline fmax. ParityPE provides fail-safe detection at low area cost and TMRPE provides fail-operational masking at baseline fmax, establishing a quantitative baseline for selectable PE-level fault detection and correction in Chisel-based AI accelerator generators.
09:45	#321 **Structural-Guided LLM-Based Reverse Engineering Attack for HLS/RTL Obfuscation** Srinivasa Rao Dara (IDRBT & NITW); Dipanjan Roy (IDRBT Hyderabad); Ilaiah Kavati (National Institute of Technology Warangal) Hardware obfuscation applied at the high-level synthesis (HLS) or register-transfer level (RTL) protects hardware accelerators against intellectual property theft and reverse engineering, and its resilience must be assessed against emerging attack techniques. Large Language Models (LLMs) can now reason over code and structured hardware representations, which raises the question of whether an LLM can recover the secret key of an HLS/RTL-obfuscated design without a functional oracle. This work presents an oracle-less LLM-based attack that operates on the synthesized gate-level netlist, for an adversary that holds only the netlist. A connectivity-aware preprocessing stage traces each key input to its key-controlled gates, extracts their two-hop structural neighborhoods, and serializes them into compact Verilog windows. Each window is combined with a small offline reference library and supplied to the LLM, which infers the corresponding key bit; the predicted bits are then assembled into the complete key. No fine-tuning, activated chip, or input-output pairs are required. The attack is evaluated with three LLMs on four hardware accelerator benchmarks protected by ASSURE, KOIL, and PROTECTS, and compared with the OMLA and lightweight multilayer perceptron (MLP) oracle-less baselines. Taking the best-performing LLM for each benchmark, the attack achieves an overall average key prediction accuracy of 56.1% on gate-level netlists containing 20k-50k gates. These results provide a first assessment of LLM-based oracle-less attacks on HLS/RTL obfuscation after logic synthesis.

B14: Machine Learning for Energy-Efficient and Reliable Manycore Systems

09:00–10:00 · Room B · Chair: Lichen Weng (Tenstorrent) · Co-chair: Md Farhadur Reza (Eastern Illinois University)

09:00	#255 **FIRE: Enhanced Fine-Tuning for Efficient Self-Repair of Permanent Hardware Failures in AI Systems** Yixuan Luo (University of Chicago); Yanjing Li (Northeastern University) Hardware failures are a top concern for AI systems. We focus on self-repair in AI systems, which allows faulty hardware to continue operating even in the presence of permanent hardware failures that persistently affect computations. General-purpose self-repair strategies typically rely on (1) incorporating spare hardware, which is costly, or (2) disabling faulty components, which increases latency. For DL workloads, previous studies exploit the inherent fault tolerance of neural networks, but require a detailed fault map that is difficult to obtain; otherwise, they significantly degrade model accuracy. To overcome these limitations, we devise FIRE, a new self-repair approach for AI workloads. Our key insight is that permanent hardware failures induce deterministic, learnable shifts in layer output values-making fine-tuning a natural and highly effective repair mechanism. Enhanced with a multi-decoder majority voting strategy and a synthetic dataset generation method, FIRE achieves accuracy improvements of 16.3%-21.0% on average (up to 78.7%) over uncorrected faulty baselines across representative AI workloads, restoring accuracy close to the fault-free level with minimal latency impact, and without requiring hardware modifications, a detailed fault map, or access to the original training dataset.
09:15	#276 **A Spike-Domain RRAM-CIM Drift Monitoring and Calibration Framework with Dual-Stage Neural Adaptation for Cough Detection** Keyi Yang and Zixuan Shen (Huazhong University of Science and Technology); Yingchen Ji and Haoran Wang (Peking University); Sunbin Deng (Huazhong University of Science and Technology); Yimao Cai and Zongwei Wang (Peking University); Wang Chao (Huazhong University of Science and Technology) Continuous cough monitoring is vital for the early diagnosis of respiratory diseases, while Resistive Random-Access Memory (RRAM)-based Computing-In-Memory (CIM) architectures enable low-power local processing for battery-limited edge devices. To address the two major issues of high-complexity Deep Neural Network (DNN) cough detection algorithm and drift-induced detection accuracy loss in RRAM-based CIM, this paper presents a lightweight Spiking Neural Network (SNN) cough detection algorithm for low-power edge inference and a spiking-domain drift monitoring and calibration framework for reliable RRAM-CIM systems. Benefiting from a depthwise separable architecture, the proposed lightweight model features only 0.009 M parameters and 0.48 M Multiply-Accumulates (MACs) per frame. To overcome RRAM retention drift, the framework monitors neuron firing rates, performing layer-wise Vth tuning of neurons followed by neuron-wise residual bias calibration without reprogramming weights. Simulation results show that the proposed framework recovers cough detection accuracy from 83.02% to 86.52% under retention drift, which is close to the 86.60% fresh baseline, providing a robust calibration solution for RRAM-CIM cough-detection systems.
09:30	#301 **Transfer Limits of Static and Causal Readout Calibration in Memristor Crossbars** Jin Shi (University of Nottingham Malaysia); Nandha Kumar Thulasiraman (University of Nottingham); Haider Abbas F. Almurib (The University of Nottingham Malaysia Campus) Memristor crossbar currents are distorted by wire resistance and parasitic capacitance, which can change the first output to spike. Static gain restores settled amplitude but leaves temporal distortion unresolved. Bounded causal lead correction for each column is evaluated against scalar gains fitted to DC current, charge, and waveform error in a distributed RC model of a 16 × 16 array. Conductances and neuron parameters remain fixed, and calibration and evaluation use disjoint row combinations. On 32 independent conductance maps, lead correction preserves 73.25% of 17,958 eligible first-spike decisions, compared with 67.26% for DC gain. The net gain of 5.99 percentage points comprises 2,035 recoveries and 959 regressions, with a 95% bootstrap interval of 4.91-7.06 points obtained by resampling maps. Independent SPICE checks support the numerical results. The benefit varies with activity and capacitance: sparse inputs can regress, static gain is nearly sufficient at small capacitance, and neither correction is reliable at the largest capacitance. These results identify the operating conditions and transfer limits of low-order temporal readout calibration.
09:45	#286 **A Trainable and Resource-Efficient Quantum Support Vector Machine for NISQ Computers**

	Sohrab Sajadimanesh, Saad Bin Ahmed and Ehsan Atoofian (Lakehead University) Quantum machine learning has emerged as a promising approach for enhancing classical learning models by exploiting the representational capabilities of quantum systems. In this work, we develop a quantum support vector machine (QSVM) framework based on a trainable quantum kernel, where classical data are encoded into parametrized quantum circuits (PQC) using angle encoding. Building upon the principles of support vector machines (SVMs), the proposed QSVM replaces classical kernel evaluations with quantum state overlaps, enabling implicit mapping into high-dimensional Hilbert spaces. Unlike conventional fixed quantum feature maps, our approach introduces trainable parameters within the embedding circuit, allowing the feature space to adapt to the underlying data distribution. These parameters are optimized via a kernel alignment strategy, improving class separability without requiring repeated retraining of the classifier. To ensure practical applicability on noisy intermediate-scale quantum (NISQ) devices, particularly superconducting qubit platforms, we incorporate a circuit pruning mechanism that dynamically reduces circuit depth by removing less significant operations during training. This approach mitigates the impact of noise, decoherence, and gate imperfections while preserving the most informative components of the model. Experimental results on classification tasks show that the proposed trainable QSVM improves classification accuracy by incorporating the proposed pruning technique on a quantum backend. These results suggest that adaptive and resource-efficient quantum kernels provide a viable pathway toward implementing quantum machine learning algorithms on real noisy quantum hardware.

A15: Hardware Acceleration of AI on Embedded Edge SoCs – III

10:20–11:20 · Room A · Chair: TBC · Co-chair: Ehsan Atoofian (Lakehead University)

10:20	#317 **A Parameterized Decoder Architecture with One-hot Regime Extraction for Posit Arithmetic Units** Diksha Shekhawat (CSIR-Central Electronics Engineering Research Institute); Jugal Gandhi (Academy of Scientific and Innovative Research (AcSIR) & CSIR-Central Electronics Engineering Research Institute); Jai Gopal Pandey (CSIR-Central Electronics Engineering Research Institute, Pilani); Chandra Shekhar (Birla Institute of Technology and Science) Posit numbers emerged as an advanced approach to numerical representation, offering several advantages over the conventional IEEE-754 standard. This research presents an efficient hardware architecture for the posit data decoder that utilizes a multiplexer-based leading-one detector (LOD). The proposed architecture has been implemented on FPGA platforms for 8, 16, and 32-bit posits with different exponent sizes (es) to validate the flexibility and performance of the design. The ASIC implementation of the architecture has been done in a 32 nm CMOS technology node. Compared to existing leading difference detector (LDD)/LOD-based decoders, the proposed design provides an average improvement of 32.01/25.73% in dynamic power, 14.24/58.39% in delay, 16.78/27.44% in areapower product (APP), 19.79/62.82% in area-delay product (ADP), and 46.14/67.19% in power-delay product (PDP).
10:35	#212 **Shared-Exponent Cross-Group Accumulation for Efficient Group-Wise Quantized LLM Inference** Zhengpan Fei and Satoshi Kawakami (Kyushu University); Takumi Honda and Naoto Fukumoto (Fujitsu Limited); Koji Inoue (Kyushu University) Fine-grained group-wise quantization is widely used to preserve the accuracy of aggressively quantized large language models. However, independent scales across groups break continuous integer accumulation and require repeated dequantization at group boundaries, introducing additional floating-point (FP) hardware and energy overhead that grows as the group size decreases. Existing approaches reduce this cost primarily by constraining scale generation or transforming data distributions to make scale handling more hardware-friendly, potentially limiting quantization flexibility or introducing additional algorithmic overhead. We propose a novel approach that decouples group-scale selection from cross-group accumulation. Our key observation is that exact FP dequantization preserves more numerical resolution than is necessary during cross-group accumulation. We therefore retain independently selected group scales while approximately aligning their partial sums in a shared-exponent integer domain. This design enables low-cost cross-group accumulation entirely in the integer domain, while preserving the flexibility of independently selected group scales. Experiments on LLaMA-2-7B, Mistral-7B, and Qwen2.5-1.5B under group-wise W4A4 quantization (4-bit weights and 4-bit activations) show that the proposed design closely matches exact FP dequantization, with less than 1% perplexity degradation across all evaluated configurations on WikiText-2. An RTL implementation using the Nangate45 technology library and an 8x8 INT4 compute core with eight dequantization lanes reduces area by 33.0%, power by up to 17.6%, and energy by up to 20.8% compared with conventional FP dequantization.
10:50	#289 **Adaptive Resource Allocation for Low-Latency LLM Serving in Dynamic Environments** Masayuki Usui (Tokyo University); Shinya Takamaeda-Yamazaki (The University of Tokyo & RIKEN) Large language models face challenges in achieving low-latency serving because the effectiveness of latency reduction techniques, such as speculative decoding and chunked-prefills, is highly sensitive to hyperparameter choices, whose optimal values vary with dynamic system conditions, including request rates and input distributions. To address this challenge, we propose a dynamic hyperparameter optimization framework that adaptively selects hyperparameter values, minimizing TPOT for speculative decoding and balancing the TTFT-TPOT trade-off for chunked-prefills. However, even with dynamically optimized hyperparameters, existing speculative decoding algorithms exhibit diminished latency gains at high request rates due to a small number of accepted tokens. To overcome this limitation, we introduce a batch-aware speculative decoding algorithm that selects candidate tokens with higher acceptance probabilities across the entire batch, thereby increasing the number of accepted tokens under high request rates. Experiments across multiple models and datasets demonstrate that our dynamic hyperparameter optimization consistently achieves low latency across diverse workloads, while the proposed batch-aware speculative decoding provides further latency reductions specifically under high request rates.
11:05	#305 **Data Movement and Memory Visibility in Coupling a CNN Accelerator to a RISC-V Host** Foivos Paraskevas, Iuliia Topko and Jan-Luca Ganz (Karlsruher Institut für Technologie); Georgios Sotiropoulos (Karlsruhe Institute of Technology); Sakir Sezer (Queen's University Belfast & Nvidia); Jürgen Becker (Karlsruher Institut für Technologie) The end-to-end speedup of a Convolutional Neural Network accelerator depends on how quickly its host can configure it and move its operands, yet integrations usually modify the control path and the data path together and report only the combined result. We couple the FleXNNgine accelerator to a Linux-capable CVA6 RISC-V core in two stages at the register-transfer level and measure each stage running Linux on an AMD

	VCU118 FPGA board. The first stage changes only the control path. Register accesses become custom instructions over the CORE-V eXtension Interface, which leaves the CVA6 pipeline unchanged. Profiling this stage shows that register setup takes at most 5.1 percent of the traced time, while copying operands into the scratchpad with single-beat processor stores takes up to 81.0 percent. The second stage changes only the data path and adds a descriptor-driven Direct Memory Access engine that the same instructions program. When tensors are made visible to it by flushing the processor's data cache, this data mover roughly halves transport time and shortens end-to-end time by a factor of 1.08 to 1.79 across three workloads. Making tensors visible through an uncached window instead yields the fastest transfers, up to 4.9 times faster than processor copies, yet no end-to-end gain, because the processor must first copy the tensors into it. Once transport is offloaded, the choice of control interface changes runtime by an estimated 4 percent at most. For hosts that reach the accelerator's scratchpad only through single-beat accesses, offloading transport should therefore come first.

B15: Heterogeneous Integration and Advanced On-Chip Interconnects

10:20–11:20 · Room B · Chair: Siddhesh Shirodker (Qualcomm Technologies) · Co-chair: Abdul Khadir Mahamed (New York University)

10:20	#296 **Open Chiplet Atlas® Architecture: A Layered Chiplet Interoperability Definition with an Adoption-Ready Reference Design** Lichen Weng, Zhimin Chen and Wei-Han Lien (Tenstorrent) Chiplet-based System-in-Package (SiP) integration promises modularity and faster time-to-market, but multi-vendor chiplets rarely share a common architecture for boot, system management, security, interrupts, or debug. This paper presents the layered Open Chiplet Atlas (OCA) architecture, which standardizes chiplet interoperability across the physical, transport, protocol, system, and software layers without constraining chiplet-internal implementation choices. We describe the OCA plug-talk-play interoperability model and its ecosystem of components (architecture, reference design, models, chiplets, and compliance tests). We then state the specific rules the OCA architecture defines for five example topics within its Functional Domains – boot, address domains, fabric deadlock avoidance, security domains with secure lifecycle control, and time synchronization – and why each is necessary for interoperability across independently designed chiplets. We report early FPGA synthesis results for the OCA Harness die-to-die (D2D) reference implementation as initial feasibility evidence, with system-management-unit results forthcoming and further supported by an independent full-chip implementation and consortium-level adoption, and relate the OCA architecture to existing link- and physical-layer-focused interconnect standards.
10:35	#312 **A Low-Latency BCH-FEC Architecture for XSR SerDes-Based Chiplet D2D Interfaces** Wang Wenyi, Jiaqing Xu, Ouyang Kuo, Qiang Wang and Lai Mingche (National University of Defense Technology) Hard-decision BCH decoding is attractive for XSR SerDes-based chiplet die-to-die (D2D) interfaces because it provides strong error-correction capability with relatively low latency and implementation complexity. In latency-sensitive chiplet systems, the latency introduced by FEC decoding becomes a non-negligible component of the overall D2D communication latency budget, especially for high-bandwidth interconnects. However, in simplified inversion-less Berlekamp-Massey (SiBM)-based decoders, the key-equation solver (KES) still suffers from strict inter-iteration dependency, which limits the achievable decoding latency. This paper presents a two-step look-ahead SiBM KES that composes two consecutive SiBM recursions into one outer iteration. Unlike direct two-step unfolding, the proposed algorithm is algebraically factorized into a branch-dependent single-multiplier-stage pipeline, thereby avoiding critical-path growth caused by cascaded Galois-field (GF) multipliers. The proposed algorithm maintains the correction capability of the conventional SiBM-based decoder while enabling a more timing-efficient hardware realization. A BCH(320, 260, 6) decoder is implemented to validate the proposed architecture. At the KES level, the proposed design reduces the average iteration count from 6 to approximately 3.12 in a 28-nm CMOS process, while delivering 1.55-ns latency and 21.08-pJ PDP at the practical 2-GHz operating point. At the decoder level, the proposed architecture reduces the end-to-end decoding latency from 3.92 ns to 2.52 ns with only an 8.2% area increase compared with the conventional SiBM-based decoder. These results indicate that the proposed KES is an efficient low-latency BCH solution for next-generation XSR SerDes-based chiplet D2D interfaces.
10:50	#291 **Quality-Adaptive Region Selection with Admission Pruning for Heterogeneous Partially Reconfigurable NoCs** Nguyen Van Cuong (Industrial University of Ho Chi Minh City) On an SoC-FPGA, a run-time manager must place each arriving streaming application in a region of a network-on-chip (NoC) whose tiles are soft processors and partially reconfigurable regions (RRs) of different sizes. Logic is scarce, every region change costs a partial reconfiguration, and the power budget varies, so the manager must trade quality for admission at a low decision cost. We study how to make this adaptation cheap. An O(N) admission test on necessary resource counts prunes infeasible quality levels before any mapping is tried, and the APS-RS-QE region selector grows regions over Manhattan rings around a free-neighbour centre, keeping large RRs for large accelerators. We also release a benchmark with quality ladders from real applications, accelerator areas and reconfiguration costs calibrated on published FPGA designs and the 7-series frame structure, and Zynq-7000-sized platforms. In 71,160 simulated runs, quality adaptation is the dominant effect, removing 26-57% of rejections. The admission test changes no decision of deterministic selectors and removes 9-69% of the mapping attempts of all eight selectors. When every selector adapts, APS-RS-QE adds a small benefit on heterogeneous RRs: on a synthetic platform and five layouts of a Zynq-7100-sized platform it is 0.6-5.8% above five of seven alternatives and statistically tied with the other two. Its regions lose fewer frames to reconfiguration than our earlier angular sweep (1.2 vs. 1.8 per run of about 17,600 frames), though several baselines lose as few or fewer. It does not lead on equal RRs or the smallest device.
11:05	#324 **Predictable Fixed-Hop Routing in Rectangular Mesh NoCs** C V Sesha Sai Kumar (BITS Pilani, Hyderabad Campus & Optima Design Automation Ltd); Soumya J (BITS Pilani, Hyderabad Campus); Sri Parameswaran (University of Sydney) Fixed-hop communication is valuable in predictable NoCs, but most mesh routing schemes optimize shortest paths or congestion rather than explicit hop count. This paper presents corridor-constrained exact-hop compilation for rectangular mesh NoCs. A source-destination pair is assigned a rectangular corridor, a base route is selected, and local cell-boundary expansions add two hops at a time while preserving route simplicity. For the full rectangular mesh, the paper also establishes a guaranteed long-route result: every source-destination pair admits a simple path of length D or D − 1, where D = (m − 1) + (n − 1). The result supports time-triggered, statically scheduled, and security oriented communication, where hop count is a design parameter rather than a side effect. Fixed-hop routing is also useful when multiple flows must be matched in path depth, when route latency must be reproducible across runs, and when communication paths must be pre-validated for timing or isolation.

A16: Design Methodologies for Embedded Multicore/Manycore SoCs – I

11:20–12:20 · Room A · Chair: Mohamed Badawy (Infineon Technologies AG & Technical University of Munich) · Co-chair: Ehsan Atoofian (Lakehead University)

11:20	#117 **End-to-End Design and Evaluation of an IOMMU-Enabled RISC-V Heterogeneous SoC Platform for TEE Secure I/O** Yanqing Liu, Haocheng Xu, Hidetoshi Uranami, Akihiro Saiki and Keiji Kimura (Waseda University) Trusted Execution Environments (TEEs) must protect memory against both CPU-initiated and device-initiated accesses. In heterogeneous SoCs, DMA-capable accelerators can bypass CPU-side protection and directly access system memory, making IOMMU-enforced DMA isolation essential for secure I/O. There is ample room to explore hardware and software for TEE with IOMMU, as their importance is increasing. This paper presents an end-to-end, Linux-enabled RISC-V heterogeneous SoC platform that integrates an open RISC-V IOMMU and an NVDLA accelerator in Chipyard and deploys the complete system on an FPGA. The platform connects real accelerator DMA traffic to the Linux IOMMU/DMA subsystem, supports IOVA-based mappings, and exposes fault containment for invalid device accesses. Unlike an isolated RTL-level IOMMU experiment, the prototype evaluates the complete path from Linux DMA mapping and driver-managed buffers to accelerator-issued memory transactions, hardware translation, and fault reporting. Evaluation with NVDLA workloads shows correct DMA isolation and suggests that the observed overhead is not determined solely by IOTLB misses: superpages reduce translation pressure, while the integrated IOMMU IP's single-request translation behavior, bridge behavior, and request-ordering constraints also contribute to the remaining end-to-end cost. The resulting platform provides a practical basis for RISC-V TEE-oriented secure-I/O research.
11:35	#217 **Design and FPGA Implementation of the RISC-V Advanced Interrupt Architecture for a Linux-Capable Multicore CVA6 SoC** Xinyuan Fu, Akihiro Saiki and Keiji Kimura (Waseda University) Interrupt handling is a critical bottleneck in multicore systems-on-chip. The legacy RISC-V Platform-Level Interrupt Controller routes all external interrupts through a single centralized claim-and-complete mechanism, which serializes servicing across cores. The RISC-V Advanced Interrupt Architecture (AIA) replaces it with modern per-hart Incoming MSI Controllers (IMSICs) and an Advanced PLIC (APLIC) that converts wired interrupts into message-signaled interrupts. Although there is a software emulation environment for RISC-V AIA on QEMU, there is no cycle-accurate evaluation platform. In this paper, we present a Linux-capable multicore evaluation platform that integrates an existing open-source AIA IP into an open-source manycore framework on an FPGA board. The evaluation result on our platform with four RISC-V cores shows that MSI delivery through the IMSIC is 2.03x faster than a CLINT software interrupt (62 versus 126 cycles), and at least 1.88x faster once the firmware round-trip supervisor software must pay to reach the CLINT is included. Besides, an end-to-end cross-hart wakeup evaluation reveals that interrupt delivery accounts for only 0.3% of application-visible latency, and hardware plus kernel entry for 2.9%.
11:50	#314 **MOSAIC-SoC: A Configuration-Driven Generator for Heterogeneous Multi-Core RISC-V Systems-on-Chip** Miloudi Adel Hani (Université de Blida); Taouinar Abdelbasset (University of Blida 1 & CDTA); Kheir Eddine Yassine Hamzi (Université de Blida); Nouma Izeboudjen (CDTA); Hamida Bougherira (Université de Blida I & Laboratoire LATSI); Nadjia Benblidia (Blida 1 university) Edge devices run three kinds of work: control, signal processing, and always-on sensing, and each needs a different amount of processing. Homogeneous multi-core microcontrollers still give all three the same core. We propose MOSAIC-SoC, a new configuration-driven generator for heterogeneous multi-core RISC-V systems-on-chip (SoCs) from a single YAML (YAML Ain't Markup Language) file, a plain-text configuration format. The file picks cores from a catalog of fourteen designs, from a bit-serial core of about 200 gates to an out-of-order 64-bit core, places each in one of three tiers, and chooses the interconnect, the memory, and the peripherals. Our main contributions are threefold. First, a three-tier core catalog and a strict configuration format whose validator checks each file before any hardware is generated and names the rule that a faulty file breaks. Second, the Standard Core Interface (SCI), which wraps every core so that it presents the same Open Bus Interface (OBI 1.3), whatever its native protocol: nine wrappers cover seven non-OBI interfaces. Third, the Task Dispatch Unit (TDU), which queues tasks and, on each push, wakes at most one core, the one the task names, without cache coherence. We evaluate integration effort and configuration coverage with a 33-step regression sweep and a 248-configuration pairwise covering array. Ten of the fourteen cores pass full-SoC simulation in the standing regression, and an eleventh passed in a covering-array run.
12:05	#254 **SimRV: An Explainable RISC-V Full-System Simulator for Architecture Education and Hardware/Software Codesign** Lennart Trunk and Kenji Kise (Institute of Science Tokyo) Computer architecture education and hardware/software codesign are hindered by fragmented tools: introductory visualizers cannot run realistic system software, while full-system emulators hide microarchitectural dynamics behind internal execution models. This paper presents SimRV, an opensource RISC-V simulator in C++23 that bridges this continuum from single-cycle pipeline hazard inspection to full Linux boot within a single unified environment. SimRV couples a compiletime parametric RV32 or RV64 core with two distinct engines: a fast functional interpreter with direct-lookup decode caching and an in-order five-stage cycle-oriented coroutine timing model with dynamic hazard detection. A split-screen terminal user interface, or TUI, pairs live register, pipeline, and cache inspection with an active guest console over a pseudo-terminal bridge, or PTY, providing zero-trace causal stall explanations across local shells, including Windows Subsystem for Linux (WSL), and remote SSH sessions without graphical forwarding overhead. On an AMD Ryzen 9 7950X3D host, SimRV sustains 265.51 MIPS on a 4- billion-instruction RV64 CoreMark run, within 1.25× of Spike, and reaches the first Linux shell prompt in 37.84 s in TUI mode and 37.73 s in CLI mode in single-run observations, alongside an analysis of dynamic fast functional (IA) and cycle-oriented (CA) runtime switching and cache limitations. Educational case studies demonstrate how SimRV's causal attribution quantifies the impact of operand forwarding and cache conflict thrashing, turning abstract microarchitectural trade-offs into visible, measurable concepts.

B16: Hardware Acceleration of AI on Embedded Edge SoCs – IV

11:20–12:20 · Room B · Chair: Lichen Weng (Tenstorrent) · Co-chair: Krishna Suman Dara (HPE)

11:20	#124 **Compact RISC-V Packed-Vector and Matrix Custom Instructions for edge-AI Kernels** Kehan Peng, Rajesh C Panicker and Enming Xu (National University of Singapore)

	Edge-AI kernels on compact RISC-V cores spend substantial instruction bandwidth on byte extraction, signed-lane conversion, mask construction, short dot products, and small matrix-tile movement. This paper presents a single-core RISC-V custom-instruction extension for fixed-shape int8 edge-AI kernels. The extension combines CUSTOM-0 packed-vector instructions for four-lane byte operations and short dot products with CUSTOM-1 matrix instructions for explicit 4×4 tile movement and multiply-accumulate operations. The design is checked through a C reference model, patched QEMU 6.2.0 custom-opcode execution, directed Verilator RTL tests, and Vivado FPGA implementation. Across five int8 kernels, the custom path reduces software-visible instruction count by 3.47×-4.93× over optimized scalar baselines. Post-route characterization on an Artix-7 device further shows that the wide single-cycle matrix datapath, rather than the packed-vector logic, forms the dominant timing-limiting structure. These results connect ISA-level instruction compression with RTL validation and physical implementation feedback, providing an implementation case study for compact custom instructions in edge-SoC datapaths.
11:35	#318 **Lightweight 3×3 Depthwise Acceleration with Sliding-Window Reuse in a RISC-V Custom Function Unit for INT8 MobileNetV2** Kasunya Himani, Mindi Rebeka, Sasmitha Jayasinghe and Stephan Balasuriya (University of Moratuwa); Chathuranga Hettiarachchi (University of Moratuwa & Nanyang Technological University) Depthwise convolution remains a significant computational cost in MobileNetV2 inference on resource-constrained RISC-V processors, even when 1×1 pointwise convolutions are hardware-accelerated. This paper presents a dedicated 3×3 depthwise convolution custom function unit (CFU) integrated into CFU Playground alongside the existing 1×1 datapath. The design combines a pipelined INT8 multiply-accumulate (MAC) datapath with sliding-window caching to reuse overlapping input values and reduce redundant CFU LOAD instructions. Evaluated using TensorFlow Lite Micro on a VexRiscv-based LiteX SoC running on a Xilinx Nexys A7-100T at 75 MHz, the depthwise convolution CFU reduces end-to-end inference cycles by 25.8%, achieving a 1.35× speedup over the existing CFU configuration. Sliding-window caching provides a further 6.85% cycle reduction, yielding a 30.8% overall reduction and a 1.45× speedup. It also reduces CFU LOAD instructions by 71.3%. The design introduces modest LUT and FF overhead without additional BRAM or DSP usage, demonstrating a lightweight approach to depthwise convolution acceleration on resource constrained FPGAs.
11:50	#213 **MCU-Compatible Recurrent Neural Inference for Embedded AHRS** Qingshan Shan (Ningbo Intelligent Equipment Research Institute Co. Ltd. & School of Artificial Intelligence, Ningbo University of Finance & Economics) Recurrent neural inference provides compact temporal modeling for embedded sensor applications. Its deployment on microcontrollers requires a graph representation compatible with bounded resources and deterministic execution. This paper presents a deployment-oriented methodology for recurrent neural inference in a fixed-horizon attitude and heading reference system (AHRS). A compatibility analysis established the recurrent graph representation required by the evaluated ST Edge AI toolchain. The recurrent computation is represented by 30 explicit calls to a shared 24-unit GRUCell. The resulting static TensorFlow Lite graph contains no WHILE operator and supports both ST Edge AI Analyze and Generate. An interface-binding contract preserves the 30-sample window, stride-5 schedule, train-only nine-channel normalization, tensor signatures, quaternion convention, and end-anchor reconstruction across model conversion and firmware integration. The deployed 6,316-parameter network requires estimated resources of 117,692 B RAM, 162,435 B Flash, and 185,066 MACC per inference. On the STM32F767ZI, the measured mean network and complete-pipeline latencies are 14.938150 and 15.000898 ms, respectively. Across 7,560 synchronized runtime pairs, the mean EKF-relative angular difference is 0.588°, with a P95 of 2.144°. These results demonstrate MCU toolchain compatibility, resource feasibility, real-time execution, and end-to-end deployment integrity.
12:05	#157 **An Efficient QML Simulator with Programmable Interconnection and Unified Operator Fusion** Zhengke Li (School of Integrated Circuits, Sun Yat-Sen University); Qiwei Dong (Nanjing University); Haolu Xie (School of Integrated Circuits, Sun Yat-Sen University); Wendong Mao and Zhongfeng Wang (Sun Yat-sen University) Variational Quantum Algorithms (VQAs) are pivotal for Quantum Machine Learning (QML) in the Noisy Intermediate-Scale Quantum (NISQ) era. Classical state-vector simulation remains essential for VQA validation but faces severe microarchitectural bottlenecks, including host-device communication walls, inefficient mapping of logical operators, and cross-boundary memory congestion. This paper proposes a highly integrated, 32-bit fixed-point on-chip QML simulator. By internalizing state preparation and CORDIC-based parameter generation, the architecture eliminates hardware-host interaction overheads. A spatiotemporal scheduler masks parameter generation latency by overlapping it with state updates. Furthermore, a frontend instruction fusion mechanism coupled with an operator-classification strategy routes logic gates to a DSP-free Boolean bypass, avoiding redundant mathematical executions. To accommodate stride-variant memory patterns, we deploy a pipelined Benes network co-optimized with an adaptive memory splicing mechanism, resolving cross-boundary routing bottlenecks with minimal area footprint.

Venue, Travel and Accommodation

Conference venue

Shanghai Jiao Tong University (Xuhui Campus)
KoGuan School of Law
1954 Huashan Road, Xuhui District, Shanghai
上海市徐汇区华山路 1954 号（上海交通大学徐汇校区）
Entrance: Gate 1 (1 号门)
Google Maps: maps.app.goo.gl/aPN72B4rCP9pksZUA
Baidu Maps: j.map.baidu.com/20/CA

Nearest metro station

Jiao Tong University Station (交通大学站), Exit 2 — Metro Lines 10 and 11
From Exit 2: turn left onto Huaihai West Road, then right onto Huashan Road. Gate 1 is about a 5-minute walk.
By taxi / DiDi: set the destination to “Shanghai Jiao Tong University Xuhui Campus – Gate 1” (上海交通大学徐汇校区–1 号门).

From Shanghai Hongqiao Airport (SHA) / Hongqiao Railway Station

Take Metro Line 10 directly to Jiao Tong University Station, Exit 2.
No transfer needed.

From Shanghai Pudong International Airport (PVG)

Take Metro Line 2 to Jiangsu Road Station, transfer to Line 11, and get off at Jiao Tong University Station, Exit 2.
A taxi or DiDi is also available from the airport taxi stand.

Accommodation

Recommended hotel: Crowne Plaza Shanghai
400 Panyu Road, Changning District, Shanghai
About 500 m from Metro Lines 10 and 11.
Group rates may be available; please contact the organizers.

Visa and invitation letters

Please check the visa requirements for your trip as early as possible; they depend on your nationality, length of stay and purpose of travel.
Invitation letters are issued only to participants who have completed full registration and payment. Contact: mcsoc.forum@gmail.com
An invitation letter does not guarantee visa approval.

More information: mcsoc-forum.org → Venue Access Map, Accommodation and Visa · Contact: mcsoc.forum@gmail.com