2025 IEEE 33rd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM 2025)

Fayetteville, Arkansas, USA
4-7 May 2025

IEEE

IEEE Catalog Number: CFP25054-POD
ISBN: 979-8-3315-0282-9

Copyright © 2025 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP25054-POD
ISBN (Print-On-Demand):	979-8-3315-0282-9
ISBN (Online):	979-8-3315-0281-2
ISSN:	2576-2613

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

2025 IEEE 33rd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)

FCCM 2025

Table of Contents

Message from the General and Program Chairs ..... **xiv**
Message from the Chair of the Special Session: Workshop on Mixed Analog-Digital Configurable Adaptable Processing (MADCAP) ..... **xvi**
Organizing Committee ..... **xvii**
Program Committee ..... **xviii**
Additional Reviewers ..... **xx**
Artifact Evaluators ..... **xxi**
Sponsors ..... **xxii**

Abstractions, Programming Models, and Tools 1

AutoNTT: Automatic Architecture Design and Exploration for Number Theoretic Transform Acceleration on FPGAs 1
Dilshan Kumarathunga (Simon Fraser University, Canada), Qilin Hu (Hunan University, China), and Zhenman Fang (Simon Fraser University, Canada)

RealProbe: An Automated and Lightweight Performance Profiler for In-FPGA Execution of High-Level Synthesis Designs 10
Jiho Kim (Georgia Institute of Technology) and Cong Hao (Georgia Institute of Technology)

HP-FFT: A General High-Performance FFT Generator using High-Level Synthesis 19
Chengyue Wang (University of California, Los Angeles), Jiaohao Zhang (University of California, Los Angeles), Yingquan Wu (Tenafe, Inc.), and Jason Cong (University of California, Los Angeles)

FREEDOM: FPGA-Based Hardware Redaction Emulator 24
Chaitali G. Sathe (The University of Texas at Dallas, USA), Yiorgos Makris (The University of Texas at Dallas, USA), and Benjamin Carrion Schafer (The University of Texas at Dallas, USA)

Memories

HBMex: An Attachment for Nonbursting Accelerators to Enhance HBM Performance 29
Canberk Sönmez (École Polytechnique Fédérale de Lausanne, Switzerland), Mohamed Shahawy (École Polytechnique Fédérale de Lausanne, Switzerland), and Paolo Ienne (École Polytechnique Fédérale de Lausanne, Switzerland)

High Throughput Matrix Transposition on HBM-Enabled FPGAs 38
Yang Yang (University of Southern California), Rajgopal Kannan (DEVCOM ARL Army Research Office, CA), and Viktor K. Prasanna (University of Southern California)

Banked Memories for Soft SIMT Processors 47
Martin Langhammer (Altera Corporation and Imperial College London, UK) and George A. Constantinides (Imperial College London, UK)

Applications I

Soaring with TRILLI: An HW/SW Heterogeneous Accelerator for Multi-Modal Image Registration.. 56
Giuseppe Sorrentino (Politecnico di Milano; University of Edinburgh), Paolo S. Galfano (Politecnico di Milano; University of Edinburgh), Eleonora D'Arnese (University of Edinburgh), and Davide Conficconi (Politecnico di Milano)

HighWave: Large-Scale High-Bandwidth Wave Simulations on FPGAs 66
Dimitrios Gourounas (The University of Texas at Austin, USA), Austin G. James (The University of Texas at Austin, USA), Bagus Hanindhito (The University of Texas at Austin, USA), Arash Fathi (ExxonMobil Technology and Engineering Company, USA), Lizy K. John (The University of Texas at Austin, USA), and Andreas Gerstlauer (The University of Texas at Austin, USA)

SMART: High-Performance SAR ATR through Model-Architecture Co-Design on FPGA 75
Sachini Wickramasinghe (University of Southern California, USA), Yi-Chien Lin (University of Southern California, USA), Cauligi Raghavendra (University of Southern California, USA), and Viktor Prasanna (University of Southern California, USA)

Abstractions, Programming Models, and Tools 2

Efficiency, Expressivity, and Extensibility in a Close-to-Metal NPU Programming Interface 85
Erika Hunhoff (University of Colorado at Boulder, USA; Work performed at AMD), Joseph Melber (AMD), Kristof Denolf (AMD), Andra Bisca (AMD), Samuel Bayliss (AMD), Stephen Neuendorffer (AMD), Jeff Fifield (AMD), Jack Lo (AMD), Pranathi Vasireddy (AMD), Phil James-Roxby (AMD), and Eric Keller (University of Colorado at Boulder, USA)

Efficient and Distributed Computation of Electron Repulsion Integrals on AMD AI Engines 95
Johannes Menzel (Paderborn University, Germany) and Christian Plessl (Paderborn University, Germany)

Chronbench: An Incremental HDL Benchmark Suite 105
Zakary Nafziger (University of British Columbia, Canada) and Steven J.E. Wilton (University of British Columbia, Canada)

Abstractions, Programming Models, and Tools 3

ITERA-LLM: Boosting Sub-8-Bit Large Language Model Inference via Iterative Tensor Decomposition 114
Yinting Huang (Imperial College London, United Kingdom), Keran Zheng (Imperial College London, United Kingdom), Zhewen Yu (Imperial College London, United Kingdom), and Christos-Savvas Bouganis (Imperial College London, United Kingdom)

InTAR: Inter-Task Auto-Reconfigurable Accelerator Design for High Data Volume Variation in DNNs 123
Zifan He (University of California, Los Angeles), Anderson Truong (University of California, Los Angeles), Yingqi Cao (University of California, San Diego), and Jason Cong (University of California, Los Angeles)

LLM4DV: using Large Language Models for Hardware Test Stimuli Generation 133
Zixi Zhang (University of Cambridge, UK), Balint Szekely (Imperial College London, UK), Pedro Gimenes (Imperial College London, UK), Greg Chadwick (lowRISC, UK), Hugo McNally (lowRISC, UK), Jianyi Cheng (University of Edinburgh, UK), Robert Mullins (University of Cambridge, UK), and Yiren Zhao (Imperial College London, UK)

SoftCUDA: Running CUDA on Softcore GPU 138
Chihyo Ahn (Georgia Institute of Technology, USA), Ruobing Han (Georgia Institute of Technology, USA), Udit Subramanya (Georgia Institute of Technology, USA), Jisheng Zhao (Georgia Institute of Technology, USA), Blaise Tine (University of California, USA), and Hyesoon Kim (Georgia Institute of Technology, USA)

CAD/NOC

Guaranteed Yet Hard to Find: Uncovering FPGA Routing Convergence Paradox 143
Shashwat Shrivastava (EPFL), Stefan Nikolić (EPFL), Sun Tanaka (EPFL), Chirag Ravishankar (AMD), Dinesh Gaitonde (AMD), and Mirjana Stojilović (EPFL)

N-TORC: Native Tensor Optimizer for Real-Time Constraints 152
Suyash Vardhan Singh (University of South Carolina, USA), Iftakhar Ahmad (University of South Carolina, USA), David Andrews (University of Arkansas, USA), Miaoqing Huang (University of Arkansas, USA), Austin R.J. Downey (University of South Carolina, USA), and Jason D. Bakos (University of South Carolina, USA)

NoH: NoC Compilation in High Level Synthesis 162
Huifeng Ke (University of California, Los Angeles), Sihao Liu (University of California, Los Angeles), Licheng Guo (RapidStream Design Automation, Inc.), Zifan He (University of California, Los Angeles), Linghao Song (University of California, Los Angeles), Suhail Basalama (University of California, Los Angeles), Yuze Chi (RapidStream Design Automation, Inc.), Tony Nowatzki (University of California, Los Angeles), and Jason Cong (University of California, Los Angeles)

A Partitioning-Based CAD Flow for Interposer-Based Multi-Die FPGAs 172
Mahesh A. Iyer (Altera), Andrew B. Kahng (University of California San Diego, USA), Jason Luu (Intel Corporation), Bodhisatta Pramanik (University of California San Diego, USA), Kristofer Vorwerk (Altera), and Grace Zgheib (Altera)

Applications 2 -Embedded and IoT Edge

Transfer Learning on the Edge for a Wireless Application using an SoC Platform 181
Yiyue Jiang (Northeastern University, US), John Dooley (Maynooth University, Ireland), Aidan Edward Colgan (Maynooth University, Ireland), Jonathan Guimaraes Ribeiro (Maynooth University, Ireland), Zhilin Ren (Northeastern University, US), and Miriam Leeser (Northeastern University, US)

Moyogi: A Memory-Centric Accelerator for Low-Latency Random Forest Inference on Embedded Devices 189
Alessandro Verosimile (Politecnico di Milano, Italy), Francesco Peverelli (Politecnico di Milano, Italy), and Marco D. Santambrogio (Politecnico di Milano, Italy)

IceSpy: Reconfigurable Edge Accelerator for Scalable and Private Structural Health Monitoring 198
Alexandra Zhang Jiang (University of California, California), Jonathan Ta (University of California, California), Yuqiao Li (University of California, California), Zhou Li (University of California, California), Nalini Venkatasubramanian (University of California, California), Monica D. Kohler (California Institute of Technology, California), and Sang-Woo Jun (University of California, California)

ML Architectures

NeuraLUT-Assemble: Hardware-Aware Assembling of Sub-Neural Networks for Efficient LUT Inference 208
Marta Andronic (Imperial College London, UK) and George A. Constantinides (Imperial College London, UK)

An Efficient FPGA-Based Hardware Accelerator of Fully Quantized Mamba-2 217
Kailing Zhou (Sun Yat-sen University, China), Han Jiao (Sun Yat-sen University, China), Wenjin Huang (Sun Yat-sen University, China), and Yihua Huang (Sun Yat-sen University, China)

FPGA-Based Approximate Multiplier for FP8 227
Ruiqi Chen (Vrije Universiteit Brussel), Yangxintong Lyu (Vrije Universiteit Brussel), Han Bao (Vrije Universiteit Brussel), Jiayu Liu (University College London), Yanxiang Zhu (VeriMake Innovation Lab), Shidi Tang (Southeast University), Ming Ling (Southeast University), and Bruno da Silva (Vrije Universiteit Brussel)

Special Session: Workshop on Mixed Analog-Digital Configurable Adaptable Processing (MADCAP)

A FeFET-Based Compute-in-Memory Architecture on FPGA for Neural Network Inference 236
Minghan Jiang (Zhejiang University, China), Yonggen Li (Zhejiang University, China), Rui Xiao (Zhejiang University, China), Haibin Shen (Zhejiang University, China), and Kejie Huang (Zhejiang University, China)

Compute-In-Memory on FPGAs for Deep Learning: A Review 243
Aman Arora (Arizona State University)

Reconfigurable Retina-Inspired Looming Detection 254
Jason Sinaga (University of Wisconsin—Madison, USA), Shay Snyder (George Mason University, USA), Md Abdullah-Al Kaiser (University of Wisconsin—Madison, USA), Dan Jinoy (University of Wisconsin—Madison, USA), Gregory Schwartz (Northwestern University, USA), Maryam Parsa (George Mason University, USA), and Akhilesh Jaiswal (University of Wisconsin—Madison, USA)

Toward Reconfigurable In-Pixel Computing: A Fault-Tolerant Design Flow for Machine Learning Accelerators 261
Houxuan Guo (Northwestern University, USA), Manuel Blanco Valentín (Northwestern University, USA), Xiuyuan He (Northwestern University, USA), and Seda Ogrenci (Northwestern University, USA)

Poster Session 1

Defending Side-Channel Attacks in Convolutional Neural Networks with Channel-Level Parallelization 268
Yankun Zhu (ShanghaiTech University, China), Ranxi Lin (ShanghaiTech University, China), and Pingqiang Zhou (ShanghaiTech University, China)

TrackGNN: A Highly Parallelized and Self-Adaptive GNN Accelerator for Track Reconstruction on FPGAs 269
Shuyang Li (Fudan University), Hanqing Zhang (Zhejiang University), Ruiqi Chen (Vrije Universiteit Brussel), Bruno da Silva (Vrije Universiteit Brussel), Giorgian Borca-Tasciuc (Rensselaer Polytechnic Institute), Dantong Yu (New Jersey Institute of Technology), and Cong Callie Hao (Georgia Institute of Technology)

SparseLUT: Sparse Connectivity Optimization for Lookup Table-Based Deep Neural Networks 270
Binglei Lou (The University of Sydney, Australia), Ruilin Wu (The University of Sydney, Australia), and Philip Leong (The University of Sydney, Australia)

Unlocking the AMD Neural Processing Unit for ML Training on the Client using Bare-Metal-Programming Tools 271
André Rösti (University of California, Irvine) and Michael Franz (University of California, Irvine)

An Energy-Efficient FPGA-Based Vision Transformer Accelerator via Software-Hardware Co-Design 272
Jiacheng Cao (Fudan University, China), Jiaqi Guo (Fudan University, China), Wei Xiong (Fudan University, China), Huanlin Luo (China Aerospace Corporation Organization, China), Jian Wang (Fudan University, China), and Jinmei Lai (Fudan University, China)

C2OPU: Hybrid Compute-in-Memory and Coarse-Grained Reconfigurable Architecture for Overlay Processing of Transformers 273
Siyuan Miao (University of California, United States), Lingkang Zhu (University of Nottingham Ningbo China), Chen Wu (Eastern Institute of Technology, China; Engineering Research Center of Chiplet Design and Manufacturing of Zhejiang Province, China), Shaoqiang Lu (Eastern Institute of Technology, China), Jinming Lyu (Eastern Institute of Technology, China; Engineering Research Center of Chiplet Design and Manufacturing of Zhejiang Province, China), and Lei He (University of California, United States)

UltraFormer: An Efficient Transformer for FPGAs 274
Victor Agostinelli (Oregon State University, Oregon; Pacific Northwest National Laboratory, Washington), Nicolas Bohm Agostini (Pacific Northwest National Laboratory, Washington), and Antonino Tumeo (Pacific Northwest National Laboratory, Washington)

Microscaling Vision Transformers on FPGAs 275
Can Xiao (Imperial College London), Jianyi Cheng (University of Edinburgh), and Aaron Zhao (Imperial College London)

BiKA: Binarized KAN-Inspired Neural Network for Efficient Hardware Accelerator Designs 276
Yuhao Liu (Dresden University of Technology, Germany; Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI Dresden/Leipzig), Germany), Salim Ullah (Ruhr University Bochum, Germany), and Akash Kumar (Ruhr University Bochum, Germany; Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI Dresden/Leipzig), Germany)

Multi-FPGA Synchronization and Data Communication for Quantum Control and Measurement 277
Yilun Xu (Lawrence Berkeley National Laboratory, USA), Abhi D. Rajagopala (Lawrence Berkeley National Laboratory, USA), Neelay Fruitwala (Lawrence Berkeley National Laboratory, USA), and Gang Huang (Lawrence Berkeley National Laboratory, USA)

Poster Session 2

RapidPnR: Accelerating the Physical Design for FPGAs via Design-Level Parallelism 278
Wanzheng Weng (ShanghaiTech University, China) and Pingqiang Zhou (ShanghaiTech University, China)

A High-Throughput Implementation of the MUSIC Algorithm using AMD Versal AI Engine 279
Peifang Zhou (Fidus Systems Inc., Canada), Bachir Berkane (Fidus Systems Inc., Canada), Vlad Druz (Fidus Systems Inc., Canada), and Mark Rollins (AMD, Canada)

APR-OIS: A Near-Sensor Point Cloud Pre-Processing Accelerator on FPGA 280
Yiming Gao (University of Florida, USA) and Herman Lam (University of Florida, USA)

Optimized Coding and Parameter Selection for Efficient FPGA Design of Attention Mechanisms... 281
Ehsan Kabir (University of Arkansas, USA), Austin R.J. Downey (University of South Carolina, USA), Jason D. Bakos (University of South Carolina, USA), David Andrews (University of Arkansas, USA), and Miaoqing Huang (University of Arkansas, USA)

On Improving the HLS Compatibility of Large C/C++ Code Regions 282
Tiago Santos (Universidade do Porto, Portugal), João Bispo (Universidade do Porto, Portugal), João M. P. Cardoso (Universidade do Porto, Portugal), and James C. Hoe (Carnegie Mellon University, USA)

Accelerating Scientific Model Optimization with a Pipelined FPGA-Based Differential Evolution Engine 283
Manuel de Castro (Universidad de Valladolid, Spain), Roberto R. Osorio (Universidade da Coruña, Spain), Yuri Torres (Universidad de Valladolid, Spain), and Diego R. Llanos (Universidad de Valladolid, Spain)

ASTEF: FPGA-Based Enhancement of Event Camera Performance in Low-Light Conditions 284
Zhaoqi Wang (University of Florida, FL), Peter Mbua (University of Florida, FL), and Christophe Bobda (University of Florida, FL)

RV-ESMC: Efficient Sparse Matrix Convolution Processor Based on RISC-V Custom Instructions for Edge Platforms 285
Huachen Zhang (Jiangnan University, China), Jianyang Ding (Jiangnan University, China), Bowen Jiang (Jiangnan University, China), Tianshuo Lu (Jiangnan University, China), Wei Xu (Jiangnan University, China), and Zhilei Chai (Jiangnan University, China)

EVO-QNN: Efficient Mixed-Precision Quantization Inference on RISC-V-Based Edge Device 286
Tianshuo Lu (Jiangnan University, China), Jianyang Ding (Jiangnan University, China), Huachen Zhang (Jiangnan University, China), Bowen Jiang (Jiangnan University, China), Wei Xu (Jiangnan University, China), and Zhilei Chai (Jiangnan University, China)

DRSA: Accelerating Macro Placement on Commercial FPGAs 287
Menzo Bouaissi (EPFL, Switzerland), Paolo Ienne (EPFL, Switzerland), Lana Josipovic (ETH Zurich, Switzerland), and Andrea Guerrieri (HES-SO Valais-Wallis, Switzerland; EPFL, Switzerland)

Poster Session 3

Low-Latency FFT/iFFT RTL Implementation for the FALCON Post-Quantum Signature Algorithm 288
Alexandre Ortega (Université Jean Monnet Saint-Etienne, France), Lilian Bossuet (Université Jean Monnet Saint-Etienne, France), and Brice Colombier (Université Jean Monnet Saint-Etienne, France)

Breaking New Ground: Division Directly in Memory 289
Farzad Razi (University of Minnesota), Mehran Shoushtari Moghadam (Case Western Reserve University), M. Hassan Najafi (Case Western Reserve University), Sercan Aygun (University of Louisiana at Lafayette), and Marc Riedel (University of Minnesota)

Analog In-Memory Computing Enhanced FPGA for High-Throughput and Energy-Efficient Acceleration 290
Archit Gajjar (Hewlett Packard Labs, USA), Lei Zhao (Hewlett Packard Labs, USA), Omar Eldash (Hewlett Packard Labs, USA), Aishwarya Natarajan (Hewlett Packard Labs, USA), Xia Sheng (Hewlett Packard Labs, Portugal), Giacomo Pedretti (Hewlett Packard Labs, USA), Aman Arora (Arizona State University, USA), Paolo Faraboschi (Hewlett Packard Labs, USA), Jim Ignowski (Hewlett Packard Labs, USA), and Luca Buonanno (Hewlett Packard Labs, USA)

BCIM: A Bit-Serial Approach for Block-Cipher-In-Memory 291
Andrew Dervay (Binghamton University, USA), Omar Al Kailani (Binghamton University, USA), and Wenfeng Zhao (Binghamton University, USA)

LLM-IMC: Automating Analog In-Memory Computing Architecture Generation with Large Language Models 292
Deepak Vungarala (New Jersey Institute of Technology, USA), Md Hasibul Amin (University of South Carolina, USA), Pietro Mercati (Intel Labs, USA), Arman Roohi (University of Illinois Chicago, USA), Ramtin Zand (University of South Carolina, USA), and Shaahin Angizi (New Jersey Institute of Technology, USA)

A Multimodal AI Acceleration with Dynamic Pruning and Run-Time Configuration 293
Hyun Woo Oh (University of California, Irvine), Hanning Chen (University of California, Irvine), Sanggeon Yun (University of California, Irvine), Yang Ni (University of California, Irvine), Behnam Khaleghi (Qualcomm), Fei Wen (Samsung Semiconductor), and Mohsen Imani (University of California, Irvine)

Performance Modeling and Comparisons of an FPGA-Based Direct Simulation Monte Carlo Solver..... 294
Saleen Bhattarai (The University of New South Wales Canberra, Australia), Edwin Peters (The University of New South Wales Canberra, Australia), Sean O'Byrne (The Australian National University, Australia), and David Petty (The University of New South Wales Canberra, Australia)

SmartNIC-Based Distributed Shared Memory 295
Hemanth Ramesh (Virginia Tech), Naarayanan Rao VSathish (Virginia Tech), Edson Horta (Virginia Tech), Antonio Barbalace (The University of Edinburgh), and Binoy Ravindran (Virginia Tech)

iSEW: in-Sensor Embedded Watermarking for Secure Imaging 296
Sepehr Tabrizch (University of Illinois Chicago, USA), Shaahin Angizi (New Jersey Institute of Technology, USA), and Arman Roohi (University of Illinois Chicago, USA)

Efficient Adaptable Streaming Aggregation Engine 297
Philippos Papaphilippou (University of Southampton, UK), Wayne Luk (Imperial College London, UK), and David Gregg (Trinity College Dublin, Ireland)

Ph.D. Forum

Ph.D. Project ARIES: Efficient Mapping and Automated Compilation for AMD Versal Devices 298
Jinming Zhuang (Brown University) and Peipei Zhou (Brown University)

Ph.D. Project: An Efficient NTT Accelerator Supporting Various Lengths for HHE 300
Hang Gu (University of Science and Technology of China), Teng Wang (University of Science and Technology of China), and Chao Wang (University of Science and Technology of China)

Ph.D. Project: Floorplan Quality Prediction for HLS Design Exploration on Multi-Die FPGA 302
Haoran Xue (University of Science and Technology of China), Teng Wang (University of Science and Technology of China), and Chao Wang (University of Science and Technology of China)

Ph.D. Project: A Novel Compilation-Based Approach for Generating Sparse Tensor Accelerators 304
Xingyan Chen (University of Science and Technology of China), Lei Gong (University of Science and Technology of China), and Chao Wang (University of Science and Technology of China)

Ph.D. Project Hardware-Aware Neural Networks 306
Marta Andronic (Imperial College London, UK) and George A. Constantinides (Imperial College London, UK)

Ph.D. Project: Holistic Partitioning and Optimization of CPU-FPGA Applications Through Source-to-Source Compilation 308
Tiago Santos (Universidade do Porto, Portugal), João Bispo (Universidade do Porto, Portugal), and João M. P. Cardoso (Universidade do Porto, Portugal)

Ph.D. Project AIM: Accelerating Arbitrary-Precision Integer Multiplication on Heterogeneous Reconfigurable Computing Platform Versal ACAP 310
Zhuoping Yang (Brown University) and Peipei Zhou (Brown University)

Author Index **313**