# 2026 IEEE International Symposium on High Performance Computer Architecture (HPCA 2026)

## Sydney, Australia
## 31 January - 4 February 2026

## Pages 1-628

**IEEE Catalog Number:** CFP26013-POD
**ISBN:** 979-8-3315-9303-2

**Copyright © 2026 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: | CFP26013-POD |
| ISBN (Print-On-Demand): | 979-8-3315-9303-2 |
| ISBN (Online): | 979-8-3315-9302-5 |
| ISSN: | 1530-0897 |

**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



# TABLE OF CONTENTS

Count2Multiply: Reliable In-Memory High-Radix Counting ..... 1
*Joao P. C. De Lima, Ben Morris, Asif A. Khan, Jeronimo Castrillon, Alex K. Jones*

Conduit: Programmer-Transparent Near-Data Processing using Multiple Compute-Capable Resources in Solid State Drives ..... 16
*Rakesh Nadig, Vamanan Arulchelvan, Mayank Kabra, Harshita Gupta, Rahul Bera, Nika M. Ghiasi, Nanditha Rao, Qingcai Jiang, Andreas K. Kakolyris, Yu Liang, Mohammad Sadrosadati, Onur Mutlu*

UniFHE: Faster Accelerator for FHE with Diverse Algebraic Structure and Balanced Memory System ..... 36
*Qingyun Niu, Lutan Zhao, Ming Cai, Kai Li, Dan Meng, Rui Hou*

QuCo: Efficient and Flexible Hardware-Driven Automatic Configuration of Tile Transfers in GPUs ..... 50
*Nicolas Meseguer, Daoxuan Xu, Yifan Sun, Michael Pellauer, Jose L. Abellan, Manuel E. Acacio*

Compression-Aware Gradient Splitting for Collective Communications in Distributed Training ..... 64
*Pranati Majhi, Sabuj Laskar, Abdullah Muzahid, Eun J. Kim*

GenPairX: A Hardware-Algorithm Co-Designed Accelerator for Paired-End Read Mapping ..... 80
*Julien Eudine, Chu Li, Zhuo Cheng, Renzo Andri, Can Firtina, Mohammad Sadrosadati, Nika M. Ghiasi, Konstantina Koliogeorgi, Anirban Nag, Arash Tavakkol, Haiyu Mao, Onur Mutlu, Shai Bergman, Ji Zhang*

ASPA: Reassigning DDR5 Parity Bandwidth ..... 96
*Fan Li, Qiufeng Li, Yanan Guo, Weidong Cao, Xin Xin*

PADE: A Predictor-Free Sparse Attention Accelerator via Unified Execution and Stage Fusion .....110
*Huizheng Wang, Hongbin Wang, Zichuan Wang, Zhiheng Yue, Yang Wang, Chao Li, Yang Hu, Shouyi Yin*

Athena: Synergizing Data Prefetching and Off-Chip Prediction via Online Reinforcement Learning ..... 129
*Rahul Bera, Zhenrong Lang, Caroline Hengartner, Konstantinos Kanellopoulos, Rakesh Kumar, Mohammad Sadrosadati, Onur Mutlu*

AQPIM: Breaking the PIM Capacity Wall for LLMs with in-Memory Activation Quantization ..... 148
*Kosuke Matsushima, Yasuyuki Okoshi, Masato Motomura, Daichi Fujiki*

GyRot: Leveraging Hidden Synergy Between Rotation and Fine-Grained Group Quantization for Low-Bit LLM Inference ..... 165
*Sangjin Kim, Yuseon Chou, Byeongcheol Kim, Jungjun Oh, Hoi-Jun Yoo*

I-POP: Ignite Positive Prefetchers ..... 180
*Yiquan Lin, Wenhai Lin, Yiquan Chen, Jiexiong Xu, Shishun Cai, Jiarong Ye, Zonghui Wang, Wenzhi Chen*

A Deadlock-Free Bridge Module for Inter-Chiplet Cache-Coherent Communication in an Open Chiplet Ecosystem ..... 196
*Zhiqiang Chen, Wenwen Fu, Yongwen Wang, Hongwei Zhou*

SFD: Towards Segment Fusion Dataflow for Spatial Accelerators ..... 209
*Fuyu Wang, Minghua Shen, Yufei Ding, Nong Xiao, Yutong Lu*

WATOS: Efficient LLM Training Strategies and Architecture Co-Exploration for Wafer-Scale Chip ..... 223
*Huizheng Wang, Zichuan Wang, Hongbin Wang, Jingxiang Hou, Taiquan Wei, Chao Li, Yang Hu, Shouyi Yin*

REASON: Accelerating Probabilistic Logical Reasoning for Scalable Neuro-Symbolic Intelligence ..... 242
*Zishen Wan, Che-Kai Liu, Jiayi Qian, Hanchen Yang, Arijit Raychowdhury, Tushar Krishna*

A PN-Free Digital 3-SAT Accelerator using Crossbar Architecture and Frequency-Controlled Counters ..... 258
*Zhezheng Ren, Chenao Yuan, Yuke Zhang, Shiyu Su*

Towards Compute-Aware In-Switch Computing for LLMs Tensor-Parallelism on Multi-GPU Systems ..... 272
*Chen Zhang, Qijun Zhang, Zhuoshan Zhou, Yijia Diao, Haibo Wang, Zhe Zhou, Zhipeng Tu, Zhiyao Li, Guangyu Sun, Zhuoran Song, Zhigang Ji, Jingwen Leng, Minyi Guo*

IVE: An Accelerator for Single-Server Private Information Retrieval using Versatile Processing Elements ..... 287
*Sangpyo Kim, Hyesung Ji, Jongmin Kim, Wonseok Choi, Jaiyoung Park, Jung H. Ahn*

SMTcheck: Accurate SMT Interference Prediction to Improve Scheduling Efficiency in Datacenters ..... 302
*Sanghyun Kim, Jinhyeok Oh, Taehun Kim, Gyutae Kim, Youngsok Kim, Jaehyun Hwang, Joonsung Kim*

ESTroM: Element-Flow Architecture for Processing Sparse Tractable Probabilistic Models ..... 317
*Anjunyi Fan, Xuejie Liu, Anji Liu, Qiuping Wu, Jiaqi Yang, Yuchao Qin, Guy Van Den Broeck, Yitao Liang, Bonan Yan*

Pinball: A Cryogenic Predecoder for Quantum Error Correction Decoding Under Circuit-Level Noise ..... 332
*Alexander Knapen, Guanchen Tao, Jacob Mack, Tomas Bruno, Mehdi Saligane, Dennis Sylvester, Qirui Zhang, Gokul S. Ravi*

SSBleed: Non-Speculative Side-Channel Attacks via Speculative Store Bypass on Armv9 CPUs ..... 349
*Chang Liu, Hongpei Zheng, Xin Zhang, Dapeng Ju, Dongsheng Wang, Yinqian Zhang, Trevor E. Carlson*

Advancing Full-Stack Acceleration for Schrödinger-Style Quantum Simulation ..... 364
*Shuang Liang, Yuncheng Lu, Ce Guo, Paul H. J. Kelly, Wayne Luk, Hongxiang Fan*

Scaling Graph Neural Network Training via Geometric Optimization ..... 378
*Fangzhou Ye, Lingxiang Yin, Hao Zheng*

C³: CXL Coherence Controllers for Heterogeneous Architectures ..... 393
*Anatole Lefort, David Schall, Nicolo Carpentieri, Julian Pritzi, Soham Chakraborty, Nicolai Oswald, Pramod Bhatotia*

Streamlined On-Chip Temporal Prefetching ..... 410
*Quang Duong, Calvin Lin*

LRM-GPU: Alleviating Synchronization Overhead for Multi-Chiplet GPU Architecture ..... 425
*Baiqing Zhong, Zhirong Ye, Xiaojie Li, Peilin Wang, Haiqiu Huang, Zhaolin Li, Zhiyi Yu, Mingyu Wang*

Predicting DRAM Failures at Scale: A Two-Stage Approach for Heterogeneous Systems ..... 439
*Chenglin Wang, Shouxin Wang, Zhirong Shen, Lu Tang, Shuyue Zhou, Ronglong Wu, Min Zhou, Jialiang Yu, Yiming Zhang*

ReThermal: Co-Design of Thermal-Aware Static and Dynamic Scheduling for LLM Training on Liquid-Cooled Wafer-Scale Chips ..... 453
*Chengran Li, Huizheng Wang, Jiaxin Liu, Jingyao Liu, Zhiheng Yue, Xia Li, Shenfei Jiang, Jinyi Deng, Yang Hu, Shouyi Yin*

LEGO: Supporting LLM-Enhanced Games with One Gaming GPU ..... 468
*Han Zhao, Weihao Cui, Zeshen Zhang, Wenhao Zhang, Jiangtong Li, Quan Chen, Pu Pang, Zijun Li, Zhenhua Han, Yuqing Yang, Minyi Guo*

Swift: High-Performance Sparse-Dense Matrix Multiplication on GPUs ..... 482
*Jinyu Hu, Huizhang Luo, Hong Jiang, Marc Casas, Kenli Li, Chubo Liu*

zkPHIRE: A Programmable Accelerator for ZKPs Over HIgh-DegRee, Expressive Gates ..... 498
*Alhad Daftardar, Jianqiao Mo, Joey Ah-Kiow, Benedikt Bunz, Siddharth Garg, Brandon Reagen*

BitDecoding: Unlocking Tensor Cores for Long-Context LLMs with Low-Bit KV Cache ..... 513
*Dayou Du, Shijie Cao, Jianyi Cheng, Luo Mai, Ting Cao, Mao Yang*

HR-DCIM: High-Reliability Floating-Point Digital CIM Architecture with Unified Low-Cost Iterative Error Correction ..... 526
*Zhen He, Yiqi Wang, Zhiheng Yue, Zihan Wu, Huiming Han, Shaojun Wei, Yang Hu, Fengbin Tu, Shouyi Yin*

HERO-Sign: Hierarchical Tuning and Efficient Compiler-Time GPU Optimizations for SPHINCS$^+$ Signature Generation ..... 541
*Yaoyun Zhou, Qian Wang*

CROPHE: Cross-Operator Dataflow Optimization for Fully Homomorphic Encryption Accelerators ..... 554
*Xinhua Chen, Jiangbin Dong, Hongren Zheng, Tian Tang, Mingyu Gao*

NPUWattch: ML-Based Power, Area, and Timing Modeling for Neural Accelerators ..... 568
*Sehyeon Kim, Minkwan Kim, Chanho Park, Hanmok Park, Seonghoon Kim, Taigon Song, William J. Song*

TurboFuzz: FPGA Accelerated Hardware Fuzzing for Processor Agile Verification ..... 582
*Yang Zhong, Haoran Wu, Xueqi Li, Sa Wang, David Boland, Yungang Bao, Kan Shi*

RPU – A Reasoning Processing Unit ..... 598
*Matthew J. Adiletta, Gu-Yeon Wei, David Brooks*

ELORA: Efficient LoRA and KV Cache Management for Multi-LoRA LLM Serving ..... 615
*Jiuchen Shi, Hang Zhang, Yixiao Wang, Quan Chen, Yizhou Shan, Kaihua Fu, Wei Wang, Minyi Guo*

PASCAL: A Phase-Aware Scheduling Algorithm for Serving Reasoning-Based Large Language Models ..... 629
*Eunyeong Cho, Jehyeon Bang, Ranggi Hwang, Minsoo Rhu*

FlashFuser: Expanding the Scale of Kernel Fusion for Compute-Intensive Operators via Inter-Core Connection ..... 645
*Ziyu Huang, Yangjie Zhou, Zihan Liu, Xinhao Luo, Yijia Diao, Minyi Guo, Jidong Zhai, Yu Feng, Chen Zhang, Anbang Wu, Jingwen Leng*

Cyclone: Designing Efficient and Highly Parallel QCCD Architectural Codesigns for Fault Tolerant Quantum Memory ..... 659
*Sahil Khan, Abhinav Anand, Kenneth R. Brown, Jonathan M. Baker*

Uni-STC: Unified Sparse Tensor Core ..... 673
*Haocheng Lian, Qiyue Zhang, Xinran Zhao, Meichen Dong, Yijie Nie, Zhengyi Zhao, Junzhong Shen, Wei Guo, Chun Huang, Bingcai Sui, Weifeng Liu*

Sassy: SmartNIC-Assisted Notification Delivery for µs-Scale RDMA Workloads..... 691
*Hamed Seyedroudbari, Alexandros Daglis*

SPLATONIC: Architectural Support for 3D Gaussian Splatting SLAM via Sparse Processing ..... 705
*Xiaotong Huang, He Zhu, Tianrui Ma, Yuxiang Xiong, Fangxin Liu, Zhezhi He, Yiming Gan, Zihan Liu, Jingwen Leng, Yu Feng, Minyi Guo*

Leveraging ASIC AI Chips for Homomorphic Encryption ..... 719
*Jianming Tong, Tianhao Huang, Jingtian Dang, Leo De Castro, Anirudh Itagi, Anupam Golder, Asra Ali, Jeremy Kun, Jevin Jiang, G. Edward Suh, Tushar Krishna*

PhasedStore: Supporting High-Performance Write-Through Cache-Coherence Protocols Under TSO ..... 737
*Burak Ocalan, Chloe Alverti, Shashwat Jaiswal, Antonis Psistakis, David Koufaty, Suyash Mahar, Steven Swanson, Josep Torrellas*

SAGe: A Lightweight Algorithm-Architecture Co-Design for Mitigating the Data Preparation Bottleneck in Large-Scale Genome Sequence Analysis ..... 751
*Nika M. Ghiasi, Talu Guloglu, Harun Mustafa, Can Firtina, Konstantina Koliogeorgi, Konstantinos Kanellopoulos, Haiyu Mao, Rakesh Nadig, Mohammad Sadrosadati, Jisung Park, Onur Mutlu*

AutoGNN: End-to-End Hardware-Driven Graph Preprocessing for Enhanced GNN Performance ..... 774
*Seungkwan Kang, Seungjun Lee, Donghyun Gouk, Miryeong Kwon, Hyunkyu Choi, Junhyeok Jang, Sangwon Lee, Huiwon Choi, Jie Zhang, Wonil Choi, Mahmut T. Kandemir, Myoungsoo Jung*

Peregrine: Accelerating TFHE Bootstrapping on GPUs via Multi-Level External Product Co-Design..... 791
*Haoqi He, Zhiwei Wang, Lutan Zhao, Dian Jiao, Dan Meng, Rui Hou*

FACE: Fully Overlapped PD Scheduling and Multi-Level Architecture Co-Exploration on Wafer ..... 805
*Zheng Xu, Dehao Kong, Jiaxin Liu, Dingcheng Jiang, Xu Dai, Jinyi Deng, Yang Hu, Shouyi Yin*

DSAssassin: Cross-VM Side-Channel Attacks by Exploiting Intel Data Streaming Accelerator ..... 821
*Ben Chen, Kunlin Li, Shuwen Deng, Dongsheng Wang, Yun Chen*

Cambricon-CIM: Enabling Energy-Efficient and Error-Resilient Analog CIM Acceleration via Reformation of Coding Bases..... 836
*Hongrui Guo, Tianrui Ma, Zidong Du, Mo Zou, Yifan Hao, Yongwei Zhao, Rui Zhang, Wei Li, Xing Hu, Zhiwei Xu, Qi Guo, Tianshi Chen*

Tempranillo: Non-Speculative Early Register Release..... 852
*Carlos Escuin, Paolo S. Galfano, Davide B. Bartolini, Leeor Peled, Mehdi Alipour*

LoCaLUT: Harnessing Capacity–Computation Tradeoffs for LUT-Based Inference in DRAM-PIM ..... 869
*Junguk Hong, Changmin Shin, Sukjin Kim, Si U. Noh, Taehee Kwon, Seongyeon Park, Hanjun Kim, Youngsok Kim, Jinho Lee*

VeloxGNN: Efficient Out-Of-Core GNN Training with Delayed Gradient Propagation ..... 885
*Yi Li, Tsun-Yu Yang, Zhaoyan Shen, Ming-Chang Yang, Bingzhe Li*

Focus: A Streaming Concentration Architecture for Efficient Vision-Language Models ..... 901
*Chiyue Wei, Cong Guo, Junyao Zhang, Haoxuan Shan, Yifan Xu, Ziyue Zhang, Yudong Liu, Qinsi Wang, Changchun Zhou, Hai H. Li, Yiran Chen*

Characterizing Cloud-Native LLM Inference at Bytedance and Exposing Optimization Challenges and Opportunities for Future AI Accelerators ..... 919
*Jingwei Cai, Dehao Kong, Hantao Huang, Zishan Jiang, Zixuan Ma, Qingyu Guo, Zhenxing Zhang, Guiming Shi, Mingyu Gao, Kaisheng Ma, Minghui Yu*

ReScue: Reliable and Secure CXL Memory ..... 938
*Chihun Song, Austin A. Cruz, Michael J. Kim, Minbok Wi, Gaohan Ye, Kyungsan Kim, Sangyeol Lee, Jung H. Ahn, Nam S. Kim*

DRACO: A Hardware-Efficient Robot Rigid Body Dynamics Accelerator with Precision-Aware Quantization Framework ..... 954
*Xingyu Liu, Jiawei Liang, Yipu Zhang, Linfeng Du, Chaofang Ma, Hui Yu, Jiang Xu, Wei Zhang*

GRTX: Efficient Ray Tracing for 3D Gaussian-Based Rendering ..... 967
*Junseo Lee, Sangyun Jeon, Jungi Lee, Junyong Park, Jaewoong Sim*

AutoHAAP: Automated Heterogeneity-Aware Asymmetric Partitioning for LLM Training ..... 981
*Yuanyuan Wang, Nana Tang, Yuyang Wang, Shu Pan, Dingding Yu, Zeyue Wang, Mou Sun, Kejie Fu, Fangyu Wang, Yunchuan Chen, Ning Sun, Fei Yang*

d'ArQ: A QOC Framework with Causality-Aware Grouping and Basis Selection ..... 998
*Changheon Lee, Hyungseok Kim, Seungwoo Choi, Youngmin Kim, Won W. Ro*

Intermittence-Aware Cache Compression ..... 1011
*Gan Fang, Jianping Zeng, Yuchen Zhou, Changhee Jung*

COMET: Communication and Memory Co-Design for Fine-Grained AI Inference in MCM Accelerators ..... 1028
*Taishu Sheng, Guangyu Sun, Dezun Dong*

HDPAT: Hierarchical Distributed Page Address Translation for Wafer-Scale GPUs ..... 1042
*Daoxuan Xu, Ying Li, Yuwei Sun, Jie Ren, Yifan Sun*

AUM: Unleashing the Efficiency Potential of Shared Processors with Accelerator Units for LLM Serving ..... 1054
*Xinkai Wang, Chao Li, Yiming Zhuansun, Jinyang Guo, Xiaofeng Hou, Jing Wang, Luping Wang, Weigao Chen, Cheng Huang, Guodong Yang, Liping Zhang, Minyi Guo*

Area Bloating and the Future of Specialization ..... 1069
*Qixuan Yu, David Wentzlaff*

RoMe: Row Granularity Access Memory System for Large Language Models ..... 1083
*Hwayong Nam, Seungmin Baek, Jumin Kim, Michael J. Kim, Jung H. Ahn*

DP-HLS: A High-Level Synthesis Framework for Accelerating Dynamic Programming Algorithms in Bioinformatics ..... 1098
*Anshu Gupta, Yingqi Cao, Jason Liang, Yatish Turakhia*

TraceQ: Trace-Based Reconstruction of Quantum Circuit Dataflow in Surface-Code Fault-Tolerant Quantum Computing ..... 1115
*Theodoros Trochatos, Christopher Kang, Andrew Wang, Frederic T. Chong, Jakub Szefer*

Towards Resource-Efficient Serverless LLM Inference with SLINFER....1129
*Chuhao Xu, Zijun Li, Quan Chen, Han Zhao, Xueyan Tang, Minyi Guo*

Near-Zero-Overhead Freshness for Recommendation Systems via Inference-Side Model Updates....1147
*Wenjun Yu, Sitian Chen, Cheng Chen, Amelie C. Zhou*

NP-CAM: Efficient and Scalable DNA Classification using a NoC-Partitioned CAM Architecture....1162
*Benjamin F. Morris, Tergel Molom-Ochir, Changchun Zhou, Yiran Chen, Alex K. Jones, Hai Li*

VectorLiteRAG: Latency-Aware and Fine-Grained Resource Partitioning for Efficient RAG....1176
*Junkyum Kim, Divya Mahajan*

eGPU: Production-Scale Elastic Sharing Over 10,000 GPUs....1191
*Xiaochuan Tang, Hao Qi, Jianbo Dong, Yinghao Yu, Zhennan Xue, Zhengyu Zhang, Daocheng Ying, Zheng Cao, Xiaoyi Lu*

SpotCC: Facilitating Coded Computation for Prediction Serving Systems on Spot Instances.... 1205
*Lin Wang, Yuchong Hu, Ziling Duan, Mingqi Li, Chenxuan Yao, Feifan Liu, Xiaolu Li, Leihua Qin, Dan Feng*

N-DIPPER: A Distributed Inter-Die Peak Power Management Network for Nand Systems .... 1219
*Jinwoo Park, John Kim*

MemSOS: OS-Guided Selective Memory Mirroring .... 1233
*Junghoon Kim, Jongheon Jeong, Seokwon Moon, Seong H. Seo, Yeonhong Park, Jinkyu Jeong, Nam S. Kim, Jae W. Lee*

µShare: Non-Intrusive Kernel Co-Locating on NVIDIA GPUs .... 1248
*Wenhao Huang, Zhaolin Duan, Laiping Zhao, Yuhao Zhang, Yanjie Wang, Yiming Li, Yihan Wang, Yichi Chen, Zhihang Tang, Kang Chen, Deze Zeng, Wenxin Li, Keqiu Li*

Cambricon-GS: An Accelerator for 3D Gaussian Splatting Training with Gaussian-Pixel Hybrid Parallelism .... 1262
*Rui Wen, Zhifei Yue, Tianbo Liu, Xinkai Song, Jin Li, Di Huang, Jiaming Guo, Xing Hu, Zidong Du, Qi Guo, Tianshi Chen*

ARIADNE: Adaptive UVM Management for Efficient GPU Memory Oversubscription .... 1276
*Hyunkyun Shin, Seongtae Bang, Hyungwon Park, Daehoon Kim*

BARD: Reducing Write Latency of DDR5 Memory by Exploiting Bank-Parallelism .... 1291
*Suhas Vittal, Moinuddin Qureshi*

The Last-Level Branch Predictor Revisited .... 1306
*David Schall, Maria Durackova, Boris Grot*

TEMP: A Memory Efficient Physical-Aware Tensor Partition-Mapping Framework on Wafer-Scale Chips.... 1322
*Huizheng Wang, Taiquan Wei, Zichuan Wang, Dingcheng Jiang, Qize Yang, Jiaxin Liu, Jingxiang Hou, Chao Li, Jinyi Deng, Yang Hu, Shouyi Yin*

The Cost of Dynamic Reasoning: Demystifying AI Agents and Test-Time Scaling from an AI Infrastructure Perspective .... 1340
*Jiin Kim, Byeongjun Shin, Jinha Chung, Minsoo Rhu*

Protean: A Programmable Spectre Defense .... 1356
*Nicholas Mosier, Hamed Nemati, John C. Mitchell, Caroline Trippel*

MIRZA: Efficiently Mitigating Rowhammer with Randomization and ALERT ..... 1376
*Hritvik Taneja, Ali Hajiabadi, Michele Marazzi, Kaveh Razavi, Moinuddin Qureshi*

TENET-V2: Applying Relation-Centric Notation to Model and Optimize Data Swizzle in the Cache of Modern NPU ..... 1389
*Hanyu Zhang, Fangxu Guo, Liqiang Lu, Long Wang, Yunfei Du, Zhe Wang, Jinghan Zhang, Jie Zhang, Chenli Xue, Chengpeng Wu, Ziyi Zhang, Yun Liang, Size Zheng, Jianwei Yin*

ORANGE: Exploring Ockham's Razor for Neural Rendering by Accelerating 3DGS on NPUs with GEMM-Friendly Blending and Balanced Workloads ..... 1404
*Haomin Li, Yue Liang, Fangxin Liu, Bowen Zhu, Zongwu Wang, Yu Feng, Liqiang Lu, Li Jiang, Haibing Guan*

LiLo: Harnessing the on-Chip Accelerators in Intel CPUs for Compressed LLM Inference Acceleration ..... 1419
*Hyungyo Kim, Qirong Xia, Jinghan Huang, Nachuan Wang, Younjoo Lee, Jung H. Ahn, Wajdi K. Feghali, Ren Wang, Nam S. Kim*

AccelFlow: Orchestrating an On-Package Ensemble of Fine-Grained Accelerators for Microservices ..... 1436
*Jovan Stojkovic, Abraham Farrell, Zhangxiaowen Gong, Christopher J. Hughes, Josep Torrellas*

SCALE: Tackling Communication Bottlenecks in Confidential Distributed Machine Learning ..... 1453
*Joongun Park, Yongqin Wang, Huan Xu, Hanjiang Wu, Mengyuan Li, Tushar Krishna*

LowCarb: Carbon-Aware Scheduling of Serverless Functions ..... 1467
*Rohan B. Roy, Devesh Tiwari*

GustavSNN: Unleashing the Power of Gustavson's Algorithm on SNN Acceleration with Column-Parallel Tick-Batch Dataflow ..... 1483
*Sangwoo Hwang, Donghun Lee, Jahyun Koo, Jaeha Kung*

PIM-Malloc: A Fast and Scalable Dynamic Memory Allocator for Processing-In-Memory (PIM) Architectures ..... 1497
*Dongjae Lee, Bongjoon Hyun, Youngjin Kwon, Minsoo Rhu*

FractalCloud: A Fractal-Inspired Architecture for Efficient Large-Scale Point Cloud Processing ..... 1514
*Yuzhe Fu, Changchun Zhou, Hancheng Ye, Bowen Duan, Qiyu Huang, Chiyue Wei, Cong Guo, Hai H. Li, Yiran Chen*

PIMphony: Overcoming Bandwidth and Capacity Inefficiency in PIM-Based Long-Context LLM Inference System ..... 1529
*Hyucksung Kwon, Kyungmo Koo, Janghyeon Kim, Woongkyu Lee, Minjae Lee, Gyeonggeun Jung, Hyungdeok Lee, Yousub Jung, Jaehan Park, Yosub Song, Byeongsu Yang, Haerang Choi, Guhyun Kim, Jongsoon Won, Woojae Shin, Changhyun Kim, Gyeongcheol Shin, Yongkee Kwon, Ilkon Kim, Euicheol Lim, John Kim, Jungwook Choi*

MoEntwine: Unleashing the Potential of Wafer-Scale Chips for Large-Scale Expert Parallel Inference ..... 1550
*Xinru Tang, Jingxiang Hou, Dingcheng Jiang, Taiquan Wei, Jiaxin Liu, Jinyi Deng, Huizheng Wang, Qize Yang, Haoran Shang, Chao Li, Yang Hu, Shouyi Yin*

CoCoTree: A Computation-Capable Architecture for Collective Communication in Scalable PIM ..... 1565
*Shunchen Shi, Qijia Yang, Fan Yang, Yu Huang, Youwei Zhuo, Zhichun Li, Ninghui Sun, Xueqi Li*

Exploration of LLM Workload Reliability Based on di/dt Effects and Voltage Droops ..... 1581
*Zhixing Jiang, Justin Garrigus, Allison Seigler, Ethan Syed, Yan-Lun Huang, Mehdi Sadi, Tawfik Rahal-Arabi, Lizy K. John*

Adaptive Draft Sequence Length: Enhancing Speculative Decoding Throughput on PIM-Enabled Systems ..... 1596
*Runze Wang, Qinggang Wang, Haifeng Liu, Long Zheng, Xiaofei Liao, Hai Jin, Jingling Xue*

The Memory Processing Unit: A Generalized Interface for End-To-End In-Memory Execution ..... 1611
*Minh S. Q. Truong, Yiqiu Sun, Dawei Xiong, Amol Shah, Alexander Glass, Abraham Farrell, James A. Bain, L. Richard Carley, Saugata Ghose*

Secret Caching Sauce for High-Performance Secure Memory ..... 1627
*Xu Jiang, Xueliang Wei, Yifei Qu, Dan Feng, Yulai Xie, Wei Tong*

RidgeWalker: Perfectly Pipelined Graph Random Walks on FPGAs ..... 1641
*Hongshi Tan, Yao Chen, Xinyu Chen, Qizhen Zhang, Cheng Chen, Weng-Fai Wong, Bingsheng He*

SALT: Track-and-Mitigate Subarrays, Not Rows, for Blast-Radius-Free Rowhammer Defense ..... 1656
*Moinuddin K. Qureshi*

V-Rex: Real-Time Streaming Video LLM Acceleration via Dynamic KV Cache Retrieval ..... 1672
*Donghyuk Kim, Sejeong Yang, Wonjin Shin, Joo-Young Kim*

Nugget: Portable Program Snippets ..... 1686
*Zhantong Qiu, Mahyar Samani, Jason Lowe-Power*

VAR-Turbo: Unlocking the Potential of Visual Autoregressive Models Through Dual Redundancy ..... 1703
*Xujiang Xiang, Fengbin Tu*

Toward Scalable Gate-Level Parallelism on Trapped-Ion Processors with Racetrack Electrodes ..... 1719
*Enhyeok Jang, Hyungseok Kim, Yongju Lee, Jaewon Kwon, Yipeng Huang, Won W. Ro*

CLINE: Improving Control Flow Compilation of Quantum Programs with Control Line Encoding ..... 1736
*Anbang Wu, Liqiang Lu, Jianwei Yin, Jingwen Leng, Minyi Guo*

Cohet: A CXL-Driven Coherent Heterogeneous Computing Framework with Hardware-Calibrated Full-System Simulation ..... 1749
*Yanjing Wang, Lizhou Wu, Sunfeng Gao, Yibo Tang, Junhui Luo, Zicong Wang, Yang Ou, Dezun Dong, Nong Xiao, Mingche Lai*

DC-MBQC: A Distributed Compilation Framework for Measurement-Based Quantum Computing ..... 1765
*Yecheng Xue, Rui Yang, Zhiding Liang, Tongyang Li*

Pulse: Fine-Grained Hierarchical Hashing Index for Disaggregated Memory ..... 1779
*Guangyang Deng, Zixiang Yu, Zhirong Shen, Qiangsheng Su, Zhinan Cheng, Jiwu Shu*

An Efficient and Scalable Hardware Architecture for Number Theoretic Transform on FPGA with Design Automation ..... 1793
*Yilan Zhu, Geng Yang, Xingyu Tian, Dilshan Kumarathunga, Liang Kong, Xianglong Deng, Shengyu Fan, Guang Fan, Guiming Shi, Lei Chen, Bo Zhang, Yisong Chang, Shoumeng Yan, Zhenman Fang, Mingzhe Zhang*

Conflux: A High-Performance Keyword Private Retrieval System for Dynamic Datasets ..... 1807
*Zehao Chen, Zhaoyan Shen, Qian Wei, Hang Lu, Lei Ju*

PinDrop: Breaking the Silence on SDCs in a Large-Scale Fleet ..... 1821
*Peter W. Deutsch, Harish D. Dixit, Gautham Vunnam, Carl Moran, Eleanor Ozer, Sriram Sankar*

Fully Parallelized BP Decoding for Quantum LDPC Codes Can Outperform BP-OSD ..... 1835
*Ming Wang, Ang Li, Frank Mueller*

Enterprise Class On-Chip Accelerator Integration ..... 1849
*Deanna Berger, Alper Buyuktosunoglu, Craig Walters, Robert Sonnelitter, Hailey Nicholson, Ashraf Elsharif, Yamil Rivera, Avery Francois, Cedric Lichtenau, Jason Kohl*

TraceRTL: Agile Performance Evaluation for Microarchitecture Exploration ..... 1864
*Zifei Zhang, Yinan Xu, Sa Wang, Dan Tang, Yungang Bao*

**Author Index**