2026 ACM/IEEE 53rd Annual International Symposium on Computer Architecture (ISCA 2026)

Raleigh, North Carolina, USA
27 June - 1 July 2026

Pages 1–676

IEEE Catalog Number: CFP26030-POD
ISBN: 979-8-3315-5066-0

Copyright © 2026, IEEE

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:	CFP26030-POD
ISBN (Print-On-Demand):	979-8-3315-5066-0
ISBN (Online):	979-8-3315-5065-3

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

CURRAN ASSOCIATES INC. proceedings .com

TABLE OF CONTENTS

MLX: Multi-Layer Execution for Structured LLM Workload Acceleration on Spatial Architectures 1
Haibin Wu, Wenming Li, Zhihua Fan, Zirui Ma, Yuqun Liu, Tengfei Xia, Yanhuan Liu, Kunming Zhang Xiaochun Ye, Dongrui Fan, Jian Weng

CODO: An Automated Compiler for Comprehensive Dataflow Optimization 17
Weichuang Zhang, Yiquan Wang, Xinzhou Zhang, Chi Zhang, Yu Feng, Xiaofeng Hou, Chao Li, Jieru Zhao Minyi Guo

CHIME: A Case for Efficient Long-Context Attention-FC Disaggregated Inference with DIMM-PIM 34
Qingyuan Liu, Liyan Chen, Haocheng Wang, Yanning Yang, Dong Du, Zhigang Mao, Naifeng Jing Yubin Xia, Haibo Chen

IroKnight: Ownership-Preserving Neural Acceleration for Inference Serving 51
Harsha Santhanam, Ashwin Rohit Alagiri Rajan, Hadi Esmaeilzadeh

Lippen: a Lightweight in-Place Pointer Encryption Architecture for Pointer Integrity 69
Erfan Iravani, Lalit Prasad Peri, Mohannad Ismail, Charitha Tumkur Siddalingaradhya, Changwoo Min Elif Bilge Kavun, Wenjie Xiong

SMoE: An Algorithm-System Co-Design for Pushing MoE to the Edge via Expert Substitution 86
Guoying Zhu, Meng Li, Haipeng Dai, Xuechen Liu, Weijun Wang, Keran Li, Jun Xiao, Ligeng Chen Wei Wang

Observability-Aided Gpu Memory Oversubscription 103
Pratheek B, Khushit Shah, Arkaprava Basu

sCROOGe: Circuit-level Design and Optimization Framework for RISC-V Out-of-Order GPUs 120
Maria Zerva, Panagiotis-Eleftherios Eleftherakis, Alexis Maras, Konstantinos Iliakis, Alexandros Moiras Sotirios Xydis

COSM: A Cooperative Scheduling Framework for Concurrent PIM and CPU Execution on Mobile Devices 137
Yilong Zhao, Fangxin Liu, Onur Mutlu, Mingyu Gao, Jian Liu, Haibing Guan, Li Jiang

Cerberus: Cross-Layer ECC Co-Design for Robust and Efficient Memory Protection 155
Junhwan Kim, Seunghyun Kim, Yesin Ryu, Saeid Gorgin, Jungrae Kim

Patterns Behind Chaos: Forecasting Data Movement for Efficient Large-Scale Moe LLM Inference 171
Zhongkai Yu, Yue Guan, Zihao Yu, Chenyang Zhou, Zhengding Hu, Shuyi Pei, Yangwook Kang, Yufei Ding Po-An Tsai

Cassandra: Enabling Reasoning LLMs at Edge via Self-Speculative Decoding 188
Soongyu Choi, Yuntae Kim, Muyoung Son, Joo-Young Kim

Combating the Memory Walls: Optimization Pathways for Long-Context Agentic Llm Inference 203
Haoran Wu, Can Xiao, Jiayi Nie, Xuan Guo, Binglei Lou, Jeffrey T.H. Wong, Zhiwen Mo, Cheng Zhang Przemyslaw Forys, Chengyang Ai, Timi Adeniran, Wayne Luk, Hongxiang Fan, Jianyi Cheng Timothy M. Jones, Rika Antonova, Robert Mullins, Aaron Zhao

Approaching Shannon Bound with Lossless LLM Weight Compression 219
Hongshi Tan, Yao Chen, Gustavo Alonso, Weng-Fai Wong, Bingsheng He

ECC Enabled Reliable and Performant Processing-in-Memory 234
Jeageun Jung, Margaret Lee, Mattan Erez

HBM-CASO: A Coordinated Approach to HBM System-Level and On-Die ECC 250
Ruizhi Zhu, Yanan Guo, Huize Li, Weidong Cao, Qian Lou, Xin Xin

ATX: Accelerator Task Extensions 266
Gerasimos Gerogiannis, Stijn Eyerman, Josep Torrellas, Wim Heirman

RHODES: Robust Optimization for Uncertainty-Aware Design of CO^2-Efficient Computing Systems 282
Mariam Elgamal, Abdulrahman Mahmoud, Gu-Yeon Wei, David Brooks, Gage Hills

Rearchitecting the Datacenter Lifecycle for AI 297
Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Ricardo Bianchini

CAPA: Manufacturing Carbon Estimation for Advanced-Packaged Architectures 314
Jingyang Liu, Gwenith Bowker-Bafna, Yuke Zhang, Natalie Enright Jerger

Five-Minute Rule 40 Years Later: A First-Principles Revisit for Modern Memory Hierarchy 331
Tong Zhang, Vikram Sharma Mailthody, Fei Sun, Linsen Ma, Chris J. Newburn, Teresa Zhang, Yang Liu Jiangpeng Li, Hao Zhong, Wen-Mei Hwu

LOONG: Utilizing Long-Stride Reprogramming to Enhance the Performance of SSDs 346
Congming Gao, Jiancong Zheng, Xufeng Yang, Qiao Li, Jian Chen, Tianyu Ren, Zheng Wan, Yina Lv Xin Xin, Min Ye, Chun Jason Xue, Jiwu Shu

Don't Surrender to Low QPS/$: Fast and Cost-Efficient ANNS with TridentANN 361
Yuchen Huang, Baiteng Ma, Erci Xu, Chuliang Weng

Shining Light on Silicon Photonic DNN Accelerators 377
Avilash Mukherjee, Mieszko Lis, Sudip Shekhar

TensorPrism: Rethinking Sparse High-Order Tensor Acceleration via Co-Occurrence Graph 394
Fangzhou Ye, Shilin Tian, Amir Ghazizadeh Ahsaei, Hao Zheng

OASIS: Outlier-Aware LUT-Based GEMM with Dual-Side Quantization for LLM Inference Acceleration 410
Xueying Wu, Baijun Zhou, Zhihui Gao, Yuzhe Fu, Qilin Zheng, Yintao He, Hai Li

Omni-LUT: Energy-Efficient LUT-Based Accelerator with Hardware-Aware KV Cache Quantization 425
Cheng-Han Tsai, Kuan-Chen Chou, Yu-Hsin Wang, Chieh-Dun Wen, Tsung Tai Yeh

QiMeng-Tensify: Scaling Up Tensor Computation Optimization via Architecture-Aware LLM-Guided MCTS 441
Shouyang Dong, Jun Bi, Yuanbo Wen, Xiyue Yu, Jianxing Xu, Guanglin Xu, Ling Li, Xuehai Zhou Tianshi Chen, Qi Guo

Taking Analytic Databases to the Bank 457
Alexandar Devic, Martin Prammer, Kevin Gaffney, Siddhartha Balakrishna Rai, Anand Sivasubramaniam Jignesh M. Patel, Ameen Akel

PuDghost: Experimental Analysis of Computation Result Corruption in Processing-Using-Dram Operations on Real Dram Chips and Implications for Future Systems 473
Daichi Tokuda, İsmail Emir Yüksel, Tatsuya Kubo, Ataberk Olgun, Haocong Luo, Nisa Bostanci Jikun Wang, A. Giray Yağlıkçı, Shinya Takamaeda-Yamazaki, Onur Mutlu

Meridian: In-Memory Acceleration for RAG with Document Attention Decomposition 490
Chaoqiang Liu, Yu Huang, Haifeng Liu, Yi Zhang, Qihang Qiu, Xueqi Li, Long Zheng, Xiaofei Liao, Hai Jin Jingling Xue

PipeIMC: A Pipelined In-SRAM Computing Architecture 505
Yikai Cui, Renhao Fan, Weike Li, Mingzhao Li, Mingyu Wang, Zhaolin Li

BAAP: Coupling Compute-in-SRAM with DRAM Banks for Near-Memory Processing 519
Cecilio C. Tamarit, Socrates S. Wong, Akshati Vaishnav, José F. Martínez

Towards Practical Interrupt Side-Channel Attacks on macOS for Apple Silicon 534
Xin Zhang, Chang Liu, Jiajun Zou, Yi Yang, Qingni Shen, Zhi Zhang, Trevor E. Carlson

Helium: Quantifying Microarchitectural Side-Channel Leakage with Probabilistic Guarantees 550
Samantha Archer, Mohammad Rahmani Fadiheh, Caroline Trippel

LÆGIS: Pinpointing and Addressing Performance Overheads of GPU-Based Confidential Computing 568
Yang Yang, Adwait Jog

MC-ORAM: A Mask-Assisted and Counter-Based Non-Deterministic ORAM Inside VM-Based TEEs 585
Yongqin Wang, Rachit Rajat, Jonghyun Lee, Mengyuan Li, Murali Annavaram

TimeGaps Channels: Exploiting CPU Halted Time for Fun and Profit 600
Yusi Feng, Xin Zhang, Sioli O'Connell, Liangwei Qiu, Chitchanok Chuengsatiansup, Daniel Genkin Yuval Yarom, Yinqian Zhang, Zhi Zhang

Dorado: Clustered Hardware Cache Coherence for 1,000+ Cores 617
Jovan Stojkovic, Abraham Farrell, Gerasimos Gerogiannis, Zhangxiaowen Gong, Christopher J. Hughes Josep Torrellas

Hierarchical Wakeup Logic of the Issue Queue for High Scalability 632
Hideki Ando, Hajime Shimada

RUNLTS: Branch Prediction with Register-Value Correlations and Hierarchical Table Orchestration 646
Toru Koizumi, Toshiki Maekawa, Masanari Mizuno, Maru Kuroki, Tomoaki Tsumura, Ryota Shioya

Augmenting the Branch Predictor with a Squashed-Branch Reuse Buffer 662
Rohit Singh, Jiayang Li, Eric Rotenberg

Revisiting Global Value Prediction: A Resurgent Complement to Local Predictors 677
Ling Yang, Libo Huang, Zhong Zheng, Bingcai Sui, Sheng Ma, Yongwen Wang, Li Shen, Junhui Wang Gang Chen, Qianming Yang, Songwen Pei, Weixia Xu

HybridSpec: Exploiting Hybrid-Bonding Memory to Accelerate LLM Serving Through Heterogeneous Architecture and Speculative Decoding 691
Zongle Huang, Wenbin Jia, Yaolei Li, Xinyuan Lin, Shupei Fan, Chen Tang, Shuwen Deng, Yongpan Liu

P^{3}-LLM: An Integrated NPU-PIM Accelerator for Edge LLM Inference Using Hybrid Numerical Formats 707
Yuzong Chen, Chao Fang, Xilai Dai, Yuheng Wu, Thierry Tambe, Marian Verhelst, Mohamed S. Abdelfattah

SMOOTH: Hardware-Assisted Fine-Grained On-Chip Memory Management for Efficient On-Device LLM Inference 723
Seulki Kim, Bokyeong Kim, Kyeonghyeon Ryu, Yeji Jung, Hwanjun Lee, Sungju Kim, Yunhyeong Jeon Daehoon Kim

SHyLA: 3D-Stacked NVM-DRAM Hybrid LLM-Inference Architecture Exploiting Data and Memory Heterogeneity 740
Liu He, Fuyao Zhou, Cheng Peng, Shunan Dong, Ziming Zhang, Huazhong Yang, Yongpan Liu Guowei Zhang, Hongyang Jia

SPEC CPU: The Next Generotion 757
Mahesh Madhav, Allen Lee, Andres Mejia, Branden Moore, Charan Soppadandi, Chris Cambly Christoph Müllner, Daniel Bowers, David Reiner, Denis Bakhvalov, Di Zhao, Duane Voth, Feng Xue Frédérique Silber-Chaussumier, James Bucek, James Southern, Jiangning Liu, Jim Himer, John Henning Kevin Smith, Kristen Yang, Kunal Kashyap, Mason Guy, Mat Colgrove, Michael Berg, Prasad Battini Prasad Joshi, Rohit Prasad, Shay Bhattacharya, Sriyash Caculo, Stefan Reimbold, Sundar Iyengar Van Smith, Zarko Todorovski

A Silicon-Proven Unified Low-Latency CXL Controller and Port-Based Routing Switch for Memory-Centric Fabrics 774
Miryeong Kwon, Seungjun Lee, Donghyun Gouk, Hongjoo Jung, Eojin Ryu, Seyeong Huh, Junseok Moon Hyein Woo, Junhee Kim, Kyungkuk Nam, Jinwoo Baek, Hyunkyu Choi, Woojin Choi, Yongjin Cho Myoungsoo Jung

Vistara: Making CXL Real-Full Path From ASIC Design and OS Support to Hyperscale Deployment 789
Neha Gholkar, Jovan Stojkovic, Hasan Al Maruf, Gregory Price, Prakash Chauhan, Hiral Patel Cedric Van Goethem Kiran Vemuri, Kiran Malwankar, Kishore Sriadibhatla, Kalyan Subramanian Shobhit Kanaujia, Chunqiang Tang, Abhishek Dhanotia

From Lab to Fleet: Building and Deploying a Practical Rowhammer Defense in Cloud SoCs 804
Stefan Saroiu, Sujay Yadalam, Alec Wolman, Will Remaklus, Daniel Berger, Isaac H. Luna, Ishwar Agarwal Jacob R. Lorch

KernelEvolve: Scaling Agentic Kernel Coding for Heterogeneous AI Accelerators at Meta 819
Gang Liao, Hongsen Qin, Ying Wang, Alicia Golden, Michael Kuchnik, Yavuz Yetim, Ruichao Xiao Jia Jiunn Ang, Chunli Fu, Yihan He, Samuel Hsia, Zewei Jiang, Roman Levenstein, Dianshi Li, Liyuan Li Ajit Mathews, Varna Puvvada, Feng Shi, Nathan Yan, Xiayu Yu, Uladzimir Pashkevich, Matt Steiner Carole-Jean Wu, Gaoxiang Liu

Graph.hls: A Compiler Framework for Composable Graph Accelerator Design 834
Feiyang Wu, Xuxiao Yang, Zhuohang Bian, Jing Wang, Ruifan Xu, Guangyu Sun, Yun Liang, Youwei Zhuo

Accelerator Polymorphism: Transcending Domain-Specific Architectures with Robotics 850
Hanyang Xu, Seongryong Oh, Yubin Lee, Ashwin Rohit Alagiri Rajan, Rohan Mahapatra, Om Patil Yuchuan Li, Jongse Park, Hadi Esmaeilzadeh

GRAINS: Enabling High-Performance and Low-Cost Graph-Based Genome Analysis via Storage-Aware Algorithm-Architecture Co-Design 871
Nika Mansouri Ghiasi, Harun Mustafa, Talu Güloglu, Rakesh Nadig, Konstantina Koliogeorgi Susana Rebolledo Ruiz, Marc Rautmann, Furkan Eris, Mohammad Sadrosadati, Jisung Park, Onur Mutlu

Lembas: Cost-Efficient Genome Alignment with External Memory and FPGA Acceleration 896
Seongyoung Kang, Se-Min Lim, Sang-Woo Jun

Lora: Towards Improved Applicability of Reconfigurable Architecture for Versatile Nonlinear Functions 911
Yuan Dai, Guibin Zou, Yuanda Yang, Huan Lin, Jiahang Lou, Yiwen Luo, Xinyu Cai, Wenbo Yin Wai-Shing Luk, Lingli Wang

Triage: An Adaptive Parallel Window Decoding Scheduler for Real-Time Fault-Tolerant Quantum Computation 928
Jiahan Chen, Chenghong Zhu, Ge Bai, Xin Wang

Coset Ensemble Decoder for Quantum Error Correction with Algorithm-Hardware Co-Design 943
Shuang Liang, Jubo Xu, Giulio Bassanino, Qianzhou Wang, Yidong Zhou, Yuncheng Lu, Zhiwen Mo Paul Kelly, Bo Yuan, Wayne Luk, Hongxiang Fan

A Streaming Architecture for Quantum Error Syndrome Compression at 4 Kelvin 958
Panagiotis Papanikolaou, Ryan Hou, Jennifer Volk, George Tzimpragos

Transpiler-Architecture Co-Design to Curb Clifford Costs in Fault-Tolerant Quantum Computing 975
Meng Wang, Chenxu Liu, Samuel Stein, Yufei Ding, Poulami Das, Prashant J. Nair, Ang Li

Kernpiler: Compiler Optimization for Quantum Hamiltonian Simulation with Partial Trotterization 991
Ethan Decker, Lucas Goetz, Evan McKinney, Erik Gustafson, Junyu Zhou, Yuhao Liu, Alex K. Jones, Ang Li Alexander Schuckert, Samuel Stein, Eleanor Crane, Gushu Li

Accelerating MoE with Dynamic In-Switch Computing on Multi-GPUs 1007
Qijun Zhang, Chen Zhang, Zhuoshan Zhou, Haibo Wang, Zhe Zhou, Zhipeng Tu, Guangyu Sun, Zhiyao Xie Yijia Diao, Zhigang Ji, Jingwen Leng, Guanghui He, Minyi Guo

ConServe: Contiguity-Preserving Memory Management for Multi-Turn LLM Serving 1022
Bingyao Li

Mapping and Communication Optimizations with Fault Tolerance for Wafer-Scale LLM Inference 1037
Junwei Cui, Le Qin, Weilin Cai, Jiayi Huang

DynoPipe: Heterogeneous Edge-Cloud LLM Serving with Dynamically Orchestrated Pipeline Boundaries 1055
Yanying Lin, Baicheng Chen, Xinyu Zhang, Chengzhong Xu, Kejiang Ye

DIAMoND: Dynamic Inference for Adaptive Edge MOE with Heterogeneous In-NAND and Near-DRAM Compute Architecture 1071
Ling Liang, Tianyang Luo, Shuzhang Zhong, Dongxue Zhao, Qichao Ma, Renjie Wei, Jingyu Wang, Meng Li Guangyu Sun, Zongwei Wang, Yimao Cai

SingularBit: Exploiting Synergy of Singular Value Decomposition and Low-Bit Quantization for Weight and KV Compression in LLM Inference 1087
Seongyon Hong, Hyundeok Kong, Jungwan Lee, Sangjin Kim, Hoi-Jun Yoo

Optimized Memory Tagging on Ampereone ® Processors 1102
Shivnandan Kaushik, Mahesh Madhav, Nagi Aboulenein, Jason Bessette, Sandeep Brahmadathan Benjamin Chaffin, Matthew Erler, Stephan Jourdan, Thomas Maciukenas, Ramya Jayaram Masti, Jon Perry Massimo Sutera, Scott Tetrick, Bret Toll, David Turley, Carl Worth, Atiq Bajwa

BoostX™-NTI: Fast, Scalable and Flexible Storage Architecture with NVMe/TCP Initiator Acceleration 1115
Hamin Jang, Jinha Jeong, Wonseok Lee, Jun Heo, Jongcheon Lee, Hyunjae Chu, Taehyun Kim Youngwoo Jeong, Dongu Kim, Heetaek Jeong, Changsu Kim, Dongup Kwon, Jangwoo Kim

M100: An Orchestrated Dataflow Architecture Powering General AI Computing 1130
Yan Xie, Changkui Mao, Changsong Wu, Chao Lu, Chao Suo, Cheng Qian, Chun Yang, Danyang Zhu Hengchang Xiong, Hongzhan Lu, Hongzhen Liu, Jiafu Liu, Jie Chen, Jie Dai, Junfeng Tang, Kai Liu, Kun Li Lipeng Ge, Meng Sun, Min Luo, Peng Chen, Peng Wang, Shaodong Yang, Shibin Tang, Shibo Chen Weikang Zhang, Xiao Ling, Xiaobo Du, Xin Wu, Yang Liu, Yi Jiang, Yihua Jin, Yin Huang, Yuli Zhang Zhen Yuan, Zhiyuan Man, Zhongxiao Yao, Li Auto

Prometheus: Toward Resilient Data Centers through Optimized Cooling Infrastructure 1142
Sourav Patel, Pratyush Kumar, Youzhi Liang, Thomas Kowalski, Nick Care, Wenjie Dong, Greg Imwalle Rita Lu, Jeremy Rice, Nick Saddock, Rick Vengalath, Urs Hölzle, Benjamin C. Lee Parthasarathy Ranganathan

Understanding Inference Scaling for LLMS: Bottlenecks, Trade-Offs, and Performance Principles 1156
Moiz Arif, Avinash Maurya, Sudharshan Vazhkudai, Bogdan Nicolae

MTIA 300: Meta's First Training Chip Featuring Built-in NICs and Collective Offloading Engines 1170
Chunqiang Tang

μRNG: A Framework for Assessing Randomness in Intermittent Computing Devices 1186
Prakhar Sah, Matthew Hicks

Intermittence-Aware Speculative Page Coloring for Secure NVM 1202
Jongouk Choi, Junyeong Park, Nicholas L'Heureux, Yan Solihin, Hyunwoo Joe, Changhee Jung

AutoFHE: An Automatic Hardware Generation Framework for Domain-Specific FHE Accelerators 1217
Yibo Du, Cangyuan Li, Bing Li, Mengdi Wang, Lian Liu, Shixin Zhao, Yinhe Han, Ying Wang

DarkStream: Exploiting Internal Throughput Contention in Data Streaming Accelerator for Timing Attacks 1232
Hyosang Kim, Ki-Dong Kang, Gyeongseo Park, Sungju Kim, Daehoon Kim

L-PCN: A Point Cloud Accelerator Exploiting Spatial Locality through Octree-Based Islandization 1248
Yiming Gao, Jieming Yin, Yuxiang Wang, Xiangru Chen, Zhilei Chai, Bowen Jiang, Jiliang Zhang Herman Lam

NS-FPS: Accelerating Farthest Point Sampling via Neighbor Search in Large-Scale Point Clouds 1262
Jiapei Zheng, Shuan Yang, Siqi He, Qi Liu, Chixiao Chen

RoCC: Harnessing Raster Operations Pipeline for Efficient Tensor Collective Communication 1277
Yuan Feng, Daniel Wong, Hyeran Jeon

STEP: Spatial Footprint Prefetcher with Multi-Point Temporal Triggers 1292
Yuanji Ye, Oliver Lenke, Thomas Wild, Andreas Herkersdorf

TDMSim: Enabling High-Density and Energy-Efficient GPU DRAM Caches with 2D-Materials for Data-Intensive Applications 1306
Chao Fu, Jingyang Zheng, Xinliu He, Xiangqi Dong, Zheng Cao, Yuning Zhan, WenZhong Bao, Peng Zhou Jun Han

RangeGuard: Efficient, Bounded Approximate Error Correction for Reliable DNNs 1322
Hanum Ko, Sangheum Yeon, Jong Hwan Ko, Jungrae Kim

Tetris: Efficient Long-context LLM Serving with Chunkwise Dynamic Sequence Parallelism 1338
Cong Li, Yuzhe Yang, Xuegui Zheng, Qifan Yang, Yijin Guan, Size Zheng, Li-Wen Chang, Shufan Liu Xin Liu, Guangyu Sun

ENEC: A Lossless AI Model Compression Method Enabling Fast Inference on Ascend NPUs 1353
Jinwu Yang, Jiaan Wu, Zedong Liu, Xinyang Ma, Hairui Zhao, Yida Gu, Yuanhong Huang, Xingchen Liu Wenjing Huang, Zheng Wei, Jing Xing, Yili Ma, Qingyi Zhang, Baoyi An, Zhongzhe Hu, Shaoteng Liu Xia Zhu, Jiaxun Lu, Guangming Tan, Dingwen Tao

STEP: Adaptive Spatio-Temporal Expert Prefetching for Low-Latency and Memory-Efficient MoE Inference 1370
Fangxin Liu, Ning Yang, Zongwu Wang, Chenyang Guan, Haomin Li, Yu Feng, Liqiang Lu, Xiang Li Siran Yang, Jiamang Wang, Lin Qu, Li Jiang, Haibing Guan

EVA: Accelerating LLM Decoding via an Efficient Vector Quantization Architecture 1385
Bowen Duan, Cong Guo, Chiyue Wei, Haoxuan Shan, Yuzhe Fu, Xinhua Chen, Yifan Xu, Ziyue Zhang Changchun Zhou, Hai Li, Yiran Chen

LoKA: Low-Precision Kernel Applications for Recommendation Models at Scale 1402
Liang Luo, Yinbin Ma, Quanyu Zhu, Vasiliy Kuznetsov, Yuxin Chen, Jian Jiao, Jiecao Yu, Buyun Zhang Tongyi Tang, Xiaohan Wei, Yanli Zhao, Zeliang Chen, Yuchen Hao, Venkatesh Ranganathan Sandeep Parab, Yantao Yao, Maxim Naumov, Chunzhi Yang, Shen Li, Ellie Wen, Wenlin Chen Chunqiang Tang

Bringing Near Data Processing Into the Low-Bit Floating-Point Era 1417
Tongxin Xie, Mingyu Gao, Zehao Wang, Zhihao Jia, Yuechen Xi, Bing Li, Mo Guang, Jiale Yan Kaiwen Long, Xingcheng Zhang, Huazhong Yang, Yuan Xie, Zhenhua Zhu, Yu Wang

NasZip: Software and Hardware Co-Design to Accelerate Approximate Nearest Neighbor Search with DIMM-Based Near-Data Processing 1434
Cheng Zou, Shuo Yang, Chen Nie, Yu Zou, Yu He, Chao Jiang, Limin Xiao, Weifeng Zhang, Zhezhi He

AQuant: Repurposing CODEC for VLM Acceleration via Adaptive Quantization 1451
Zhuoran Song, Chunyu Qi, Jian Weng, Xiaoyao Liang, Haibing Guan

Random-Access Hardware Sequence Compression 1466
Nolan Chu, Yoon Lee, Gagandeep Panwar, Xun Jian

Coarse-Grained Duplication First, Fine-Grained Deduplication Later: Duplication-Centric Multi-GPU Memory Management 1482
Xiangyue Huang, Yanan Guo, Yuanchao Xu

Reducing Page Faults via Invalidation-Based Mapping Propagation in Multi-GPU Systems 1497
Junsung Kim, Dongho Ha, Sungwoo Kim, Wonho Cho, Sungbin Kim, Yufei Ding, Won Woo Ro

æSIP: µArch-Aware ASIP-ISA Co-Design via Program Synthesis, Equality Saturation, and External Don't Cares 1512
Haoran Allen Jin, Jirong Yang, Barry Lyu, Ruijie Jerry Gao, Nathan Bleier

ColumnKeeper: Efficient Solutions to the Columndisturb Vulnerability in Dram-Based Systems 1529
Andreas Kosmas Kakolyris, F. Nisa Bostancı, Ataberk Olgun, İsmail Emir Yüksel, Harsh Songara Konstantinos Marios Sgouras, Umut Başer, Konstantinos Kanellopoulos, A. Giray Yağlıkçı, Onur Mutlu

PVAC: A Rowhammer Mitigation Architecture Exploiting Per-Victim-Row Counting 1546
Jumin Kim, Seungmin Baek, Hwayong Nam, Minbok Wi, Nam Sung Kim, Jung Ho Ahn

Loaded Dice: Solving the Non-Selection Problem for Scalable Probabilistic RowHammer Defense 1562
Jeonghyun Woo, Junsu Kim, Aamer Jaleel, Prashant J. Nair

PRowhammer: Propagating Bit-Flips from CPU to GPU 1579
Mrityunjay Shukla, Shubham Roy, Sayandeep Saha, Biswabandan Panda

DejaVu: Why You Should Write to Your DRAM Rows Twice, Carefully 1595
Haocong Luo, İsmail Emir Yüksel, Ataberk Olgun, Nisa Bostanci, Orhun Ecemiş, Abdullah Giray Yaglikci Onur Mutlu

DICE: Detailed Inter-Chiplet End-to-End Phy Modeling for Accurate Chiplet Simulation 1611
Rashid Aligholipour, Stefanos Kaxiras, Yuan Yao

Omelet: A Packaging-Aware Hierarchical Interconnect Simulator for 2.5D/3D Chiplet Architectures 1628
Jiho Kim, Danish Baig, Faaiq Waqar, Ashita Victor, Shimeng Yu, Muhannad Bakir, Cong Hao

PhaseWeave: Phase-Aware Execution on Heterogeneous Chiplet Architectures for Datacenters 1643
Joshua Kim, Chaojie Zhang, Íñigo Goiri, Christopher J. Rossbach, Jovan Stojkovic

ConBin: a Performance-Convergence Framework for Wafer-Scale Chip Binning 1659
Huiqing Xu, Mengdi Wang, Yinhe Han, Ying Wang

WaferBRAIN: Whole-Brain Scale Neuromorphic Architecture Based on Wafer-Scale Integration 1674
Yukun Feng, Hao Jia, Liangyu Gan, Haoming Chu, Yufan He, Jiaxin Yin, Lirong Zheng, Ning Ma Yuxiang Huan

DS-ISA: Instruction Set Architecture for Dynamical System Units 1689
Chunshu Wu, Ruibing Song, Chuan Liu, Tong Geng, Ang Li

NeRArch-Sim: A Unified Simulator for Benchmarking and DSE of Neural Rendering Accelerators 1704
Cheng-Jhih Shih, Chaojian Li, Chihao Yu, Hsuan-Chen Fang, Sixu Li, Wei-Po Hsin, Lexington Whalen Hyewon Suh, Greg Eisenhauer, Ling Liu, Yingyan Celine Lin

BulletTime: Time Dilation for High-Fidelity Tracing 1719
Michael Wu, Sibren Isaacman, Abhishek Bhattacharjee, Anurag Khandelwal

PipeWeave: Synergizing Analytical and Learning Models for Unified GPU Performance Prediction 1734
Kaixuan Zhang, Yunfan Cui, Shuhao Zhang, Chutong Ding, Shiyou Qian, Luping Wang, Jian Cao Guangtao Xue, Cheng Huang, Guodong Yang, Liping Zhang

SSBench: Automated Characterization of Memory Dependence Predictors on Modern CPUs 1750
Chang Liu, Yu Jin, Yuchen Fan, Tianrui Xiao, Lingfeng Yin, Trevor E. Carlson, Shuwen Deng Dongsheng Wang

R2d2: Robotized Reconfigurable Network for Disaggregated Datacenters 1766
Linus Y. Wong, Zhiyao Tang, Justin R. Yu, Zhilei Zheng, Jonathan M. Smith, André DeHon, Jing Li

Lotus: A Multi-FPGA Task Dataflow Architecture to Accelerate Cycle-Level Simulation 1783
Fares Elsabbagh, Joel S. Emer, Daniel Sanchez

SegFold: Accelerating Sparse Gemm with a Fine-Grained Dynamic Dataflow 1797
Xinrui Wu, Hanyu Wang, Jason Cong, Tony Nowatzki

ParetoES: Hardware-Accelerated Sparse Embedding Similarity via Pareto-Optimal Pruning 1814
Jiaqi Zhai, Xuanhua Shi, Wenju Zhao, Kaiyi Huang, Chencheng Ye, Shunsen Lv, Zhongtian Long Bingsheng He, Hai Jin

ECHO: Efficient Head-Orientation-Guided Real-Time Sound Spatialization for Virtual Reality 1829
Haiyu Wang, Tianhua Xia, Sai Qian Zhang

DESSCAM: An Event-Driven Architecture with In-Sensor Epitopological Sparse Sampling to Break the Latency-Power Tradeoff in Eye Tracking 1845
Zhijie Jian, Shangyu Yang, Yuan Hua, Jilin Zhang, Ziyi Cheng, Hong Chen

Slice: A Selective Local Inference Framework with Codec Exploitation for Accelerating Video Super-Resolution 1863
Mingu Jung, Sungbin Kim, Seunghyun Lee, Seung Hyun Jin, Hyunwuk Lee, Won Woo Ro

Enabling Continuous, in-Field Introspection: The Programmable IPU Architecture 1878
Ian McDougall, Shayne Wadle, Harish Babu Batchu, Karthikeyan Sankaralingam

Fenc²: Unifying Data Packing for Efficient Private Inference via Convolution and Architecture-Aware Fragment Encoding 1894
Ran Ran, Zhaoting Gong, Nuo Xu, Yuanchao Xu, Fan Yao, Wujie Wen

FlashTFHE: A Scalable Architecture for Efficient Multi-Bit Fully Homomorphic Encryption 1909
Jiaao Ma, Ceyu Xu, Ning Liang, Lisa Wu Wills

Unlocking Pipeline Parallelism for Bootstrapping: A Pipelined Multi-Chiplet TFHE Accelerator 1924
Yibo Du, Mengdi Wang, Cangyuan Li, Yinhe Han, Ying Wang

He²: A Communication-Light Heterogeneous Architecture for Efficient Fully Homomorphic Encryption 1940
Shangyi Shi, Husheng Han, Zhaoxuan Kan, Yinghao Yang, Jianan Mu, Tenghui Hua, Ge Yu, Xinyao Zheng Ling Liang, Zidong Du, Xing Hu

HyperDrive: Hierarchical Exploitation of Memory Efficiency for GPU-Based FHE Acceleration 1955
Guang Fan, Yi Chen, Lei Chen, Liang Kong, Chao Niu, Dian Jiao, Yilan Zhu, Geng Yang, Shengyu Fan Xianglong Deng, Fangyu Zheng, Jian Weng, Meng Li, Yisong Chang, Shoumeng Yan, Mingzhe Zhang

MNEMOS: A GPU-Based TFHE Acceleration Framework with Memory Access Optimization 1970
Junyi Zhang, Xianglong Deng, Yi Chen, Guang Fan, Lei Chen, Dian Jiao, Shengyu Fan, Zhiwei Wang Mingzhe Zhang

GauTracer: Extending Ray Tracing Accelerator for Gaussian-Based Scene Representation 1985
Lizhou Wu, Kunchen Zou, Yuzheng Lin, Chixiao Chen, Xiaoyang Zeng, Haozhe Zhu

TTP: A Hardware-Efficient Design for Precise Prefetching in Ray Tracing 2000
Yavuz Selim Tozlu, Anshul Naithani, Huiyang Zhou

Efficient 3D Gaussian Splatting with Axis-Shared Rasterization and Order-independent Transmittance 2014
Zhican Wang, Guanghui He, Lingjun Gao, Dantong Liu, Shell Xu Hu, Chen Zhang, Zhuoran Song Nicholas Lane, Hongxiang Fan

Bumper: Hinting Instruction Usefulness for Robust Unified Caches 2029
Georgios Vavouliotis, Tom Rollet, Davide Basilio Bartolini, Boris Grot, Leeor Peled, Lixia Yang

ICP: Exploiting Instruction Correlation for Prefetching Irregular Memory Accesses 2046
Mengming Li, Chenlu Miao, Buqing Xu, Qijun Zhang, Xiangfeng Sun, Ceyu Xu, Yuan Xie, Wenkai Li Shang Liu, Zhiyao Xie

Revelator: Rapid Data Fetching Via System-Software-Guided Hash-Based Speculative Address Translation 2061
Konstantinos Kanellopoulos, Konstantinos Sgouras, Harsh Songara, Andreas Kosmas Kakolyris Vlad-Petru Nitu, Spiros Galanopoulos, Rahul Bera, Konstantina Koliogeorgi, Rakesh Kumar, Onur Mutlu

Distilling Magic States in the Bicycle Architecture 2081
Shifan Xu, Kun Liu, Patrick Rall, Zhiyang He, Yongshan Ding

O3LS: Optimizing Lattice Surgery via Automatic Layout Searching and Loose Scheduling 2096
Chenghong Zhu, Xian Wu, Jiahan Chen, Keming He, Junjie Wu, Xin Wang, Lingling Lao

Leveraging Phase Polynomials for Quantum Circuit Optimization 2111
Zihan Chen, Henry Chen, Yuwei Jin, Enhyeok Jang, Mingkuan Xu, Vannessa Chan, Won Woo Ro Eddy Z. Zhang

Harmonia: A Unified Hierarchical Scheduling Framework for Sparse Matrix Multiplication 2127
Jingkui Yang, Fangxin Liu, Xin Ju, Ning Yang, Chenyang Guan, Junjie Wang, Zongwu Wang, Mei Wen Jian Liu, Li Jiang, Haibing Guan

PipeComm: Maximizing Link Utilization Through Pipeline-Aware Collective Communication Synthesis 2142
Ruifan Xu, Yuze Luo, Yuhao Meng, Size Zheng, Meng Li, Yun Liang

Breaking Barriers in Atomic Scaling: A Hardware-Software-Collaborated Framework to Deconstruct RDMA Atomic 2158
Guangyang Deng, Qiangsheng Su, Zhirong Shen, Qing Wang, Yina Lv, Ronglong Wu, Jiwu Shu

Democratizing and Accelerating Hardware Verification with Software-Native Optimization 2173
Yunlong Xie, Zhicheng Yao, Fangyuan Song, Jincheng Liu, Junyue Wang, Haojin Tang, Lu Chen, Yinan Xu Ziqing Zhang, Ziyuan Gao, Duan Yu, Hongtao Zhou, Jiayi Rao, Junyu Yue, Xiaolong Li, Yunqi Lu Zechen Yang, Hang Zhu, Shan Liu, Xu An, Qi Ge, Jiuyue Ma, Jianyi Meng, Kan Shi, Dan Tang, Tianyi Liu Sa Wang, Yungang Bao

HARTBREAKER: Deterministic Fuzzing of Multi-Hart RISC-V CPUs with Non-Deterministic Programs 2189
Quentin Bordier, Tobias Kovats, Flavien Solt, Kaveh Razavi

QED: Scalable Consistency Verification of Memory Instruction Reordering in Hardware 2205
Gokulan Ravi, Xiaokang Qiu, Mithuna Thottethodi, T. N. Vijaykumar

TäKōFormal: Enabling Robust Software for Programmable Memory Hierarchies 2221
Pranav Srinivasan, Manos Kapritsos, Yatin A. Manerkar

LIBRA: A High-Accuracy, Cost-Aware, and Coordinated Multi-GPU Page Prefetcher 2238
Xiangyue Huang, Yanan Guo, Yuanchao Xu

Enhancing Instruction Prefetching via Cache and TLB Management 2254
Alexandre Valentin Jamet, Georgios Vavouliotis, Martí Torrents, Dimitrios Chasapis, Marc Casas

R-Max: Extending BéLáDy's MIN with Prefetching to Bound Realistic Cache Performance 2270
Lei Wang, Chia-Hang Lee, Maccoy Merrell, Gino Chacon, Daniel A. Jiménez, Paul V. Gratz

From Memorization to Generalization: A Practical Neural Network Prefetching Framework 2287
Xuan Tang, Zicong Wang, Shuiyi He, Hao Tang, Dezun Dong, Xiangke Liao

Unifying Qubit Routing Across Diverse Quantum ISAs via Canonical Representation 2302
Zhaohui Yang, Kai Zhang, Xinyang Tian, Xiangyu Ren, Yingjian Liu, Yunfeng Li, Dawei Ding, Jianxin Chen Yuan Xie

TUSQ: Tracking, Uncomputation, and Sampling for Noisy Quantum Simulation 2318
Siddharth Dangwal, Tina Oberoi, Ajay Sailopal, Dhirpal Shah, Frederic T. Chong

Photonic Quantum Computing on Spin Memory Architecture with Tree-Encoded Fusion 2333
Xiangyu Ren, Yuexun Huang, Zhemin Zhang, Yuchen Zhu, Tsung-Yi Ho, Antonio Barbalace, Zhiding Liang

SATIC: An Optimizing Ising Compiler for SAT(isfiability) 2349
Ahmet Efe, Hüsrev Cılasun, Abhimanyu Kumar, Nafisa S. Prova, Ziqing Zeng, Tahmida Islam, Ruihong Yin Chaohui Li, Peter Kreye, Chris H. Kim, Sachin S. Sapatnekar, Ulya R. Karpuzcu

PowerGrad: Hierarchical Power Management for Power-Limited ML Inference Clusters 2364
Hyoungwook Nam, Raghavendra Pradyumna Pothukuchi, Alper Buyuktosunoglu, Aporva Amarnath Pradip Bose, Josep Torrellas

Power Sloshing in Compound Servers for Large-Scale AI Inference Workloads 2379
Albert Cho, Jovan Stojkovic, Leonardo Piga, Abhishek Dhanotia, Sultan Mahmud Sajal, Gefei Zuo Krishna Malladi, Devon Akers, Kalyan Subramanian, Shobhit Kanaujia, Alexandros Daglis

PowerWeave: Unlocking Energy-Efficient ML on GPUs with OS-Level Spatial Power Management 2394
Vasilis Kypriotis, Eric Dubberstein, Patrick H. Coppock, Eliot H. Solomon, Rayyan Zamir Tathagata Srimani, Dimitrios Skarlatos

Lit Silicon: A Case Where Thermal Imbalance Couples Concurrent Execution in Multiple GPUs 2409
Marco Kurzynski, Shaizeen Aga, Di Wu

Scalable Synthesis of Distributed Llm Workloads Through Symbolic Tensor Graphs 2425
Changhai Man, Joongun Park, Hanjiang Wu, Huan Xu, Srinivas Sridharan, Tushar Krishna

Disdp: Disaggregating Compute, Network, and Storage for Model-Sharded Data-Parallel Training 2442
Mo Sun, Zihan Yang, Changyue Liao, Yingtao Li, Jie Zhang, Kaiqi Chen, Fei Wu, Zeke Wang

MoE-Hub: Taming Software Complexity for Seamless MoE Overlap with Hardware-Accelerated Communication on Multi-GPU Systems 2459
Zhuoshan Zhou, Chen Zhang, Shuyi Zhang, Qijun Zhang, Haibo Wang, Zhe Zhou, Zhipeng Tu Guangyu Sun, Yijia Diao, Zhigang Ji, Jingwen Leng, Guanghui He, Minyi Guo

Symbiotic MLLM Serving: Dynamically Balancing Parallelism Across GPUs and Resources Within GPUs 2475
Zhicheng Li, Jiacheng Zhao, Yangyu Zhang, Zhaolin Duan, Xinyu Liu, Siqi Li, Shuoming Zhang Shuaijiang Li, Donglin Yu, Yuan Wen, Chunwei Xia, Xiyu Shi, Huimin Cui

Dynamic Scheduling for AI Accelerators via TISA 2490
Guanghui Song, Xiaoqiang Dan, Chengke Wang, Fei Liu, Wenyuan Lv, Zhongzhou Jiang, Jianjian Guan Teng Lu, Lin Tao, Cheng Li, Weixing Pan, Wei Huang, Zirong Shen, Yi Yang, Hui Liu, Jie Zhao

MXFFP: Microscaling Flexible Floating Point Format for Large-Scale AI Model Acceleration 2505
Sungwoo Kim, Sungbin Kim, Dongho Ha, Hyunwuk Lee, Junsung Kim, Seunghyun Lee, Mingu Jung Murali Annavaram, Won Woo Ro

UniCore: A Bit-Width Scalable GEMM Unit for Unified LLM Inference 2520
Yonghao Chen, Jiaxiang Zou, Xingyu Chen, Chenxi Xu, Jingyu Guo, Xinyu Chen

XtraMAC: An Efficient MAC Architecture for Mixed-Precision LLM Inference on FPGA 2536
Feng Yu, Hongshi Tan, Yao Chen, Weng-Fai Wong, Bingsheng He

ELSA: An Elastic Snn Inference Architecture for Efficient Neuromorphic Computing 2550
Kang You, Chen Nie, Lee Jun Yan, Ziling Wei, Cheng Zou, Zekai Xu, Yu Feng, Honglan Jiang, Zhezhi He

AXLE: Coordinated Offloading with Asynchronous Back-Streaming in Computational Memory Systems 2567
Suyeon Lee, Kangkyu Park, Kwangsik Shin, Ada Gavrilovska

DCC: Data-Centric Compilation of Machine Learning Kernels for Processing-In-Memory Architectures 2582
Peiming Yang, Sankeerth Durvasula, Ivan Fernandez, Mohammad Sadrosadati, Onur Mutlu Gennady Pekhimenko, Christina Giannoula

Optimizing Spatial Data Structure with Near-Cache Acceleration by Exploiting Physical Locality 2600
Hongyi Li, Yijia Liu, Haoran Pei, Qingyuan Yang, Zijian Pan, Songchen Ma, Leshan Li, Rong Zhao Xinglong Ji

Bridging Efficiency and Scalability in Llm System Via 3D Hybrid Pim With 2D in-Transit Computation 2615
Hongyi Li, Songchen Ma, Huanyu Qu, Weihao Zhang, Jia Chen, Junfeng Lin, Fengbin Tu, Rong Zhao

Early Silicon of Raptor: The First 3D-DRAM Accelerator for Generative Inference 2632
Prashant J. Nair, Ramyad Hadidi, Subramani Ganesh, Sangamesh Kodge, Shubhankit Rathore Neil Thanawala, Nikitha Reddy, Gyanesh Saharia, Vinayak Patankar, Arun Tiruvur, Nithesh Kurella Sudeep Bhoja

DICE: Enabling Efficient General-Purpose SIMT Execution with Statically Scheduled Coarse-Grained Reconfigurable Arrays 2648
Jiayi Wang, Ang Da Lu, Zhichen Zeng, Ang Li

TAGT: An Efficient Graph Transformer Accelerator with Topology-aware Sparsification and Merging 2664
Hui Yu, Wei Zhang, Ligang He, Jin Zhao, Yu Zhang, Zixiao Wang

GenZA: A General and Efficient Accelerator for Diverse Zero-Knowledge Proof Protocols 2679
Cheng Wang, Jiangbin Dong, Mingyu Gao

HiT: A Unified Sparsity-Adaptive Architecture for High-Throughput Matrix Multiplication 2694
Tingting Xiang, Xiaochen Wang, Miao Yu, Trevor E. Carlson

DiTPA: A DiT-Based Action Planner Accelerator Exploiting Action-Denoising-Multimodality Redundancy for Embodied Artificial Intelligence 2709
Xin Zhao, Longke Yan, Jiancong Li, Yongkun Wu, Fengbin Tu