2025 IEEE International Symposium on High Performance Computer Architecture (HPCA 2025)

Las Vegas, Nevada, USA 1-5 March 2025

Pages 1-593



IEEE Catalog Number: CFP25013-POD **ISBN:**

979-8-3315-0648-3

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:	CFP25013-POD
ISBN (Print-On-Demand):	979-8-3315-0648-3
ISBN (Online):	979-8-3315-0647-6
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



2025 IEEE International Symposium on High Performance Computer Architecture (HPCA) HPCA 2025

Table of Contents

Message from General Chairs	xxv
Message from Program Chairs	xxvi
Organizing Committee	xxix
IEEE TCCA Executive Committee	xxxi
Program Committee	xxxii
External Review Committee	xxxvii
Sponsors	xxxix
•	

Session 1A: Reliable Bets

 Veritas – Demystifying Silent Data Corruptions: µArch-Level Modeling and Fleet Data of Modern x86 CPUs
ChameleonEC: Exploiting Tunability of Erasure Coding for Low-Interference Repair
 DPUaudit: DPU-Assisted Pull-Based Architecture for Near-Zero Cost System Auditing

Session 1B: Speculating in Vegas - 1

Delinquent Loop Pre-Execution using Predicated Helper Threads
MASCOT: Predicting Memory Dependencies and Opportunities for Speculative Memory Bypassing 59
Karl H. Mose (University of Cambridge), Sebastian S. Kim (University of Murcia), Alberto Ros (University of Murcia), Timothy M. Jones (University of Cambridge), and Robert D. Mullins (University of Cambridge)
Architecting Value Prediction around In-Order Execution

Session 1C: Next Big Bets - 1

Efficient Optimization with Encoded Ising Models Devrath Iyer (Stanford University) and Sara Achour (Stanford University)	. 85
SPARK: Sparsity Aware, Low Area, Energy-Efficient, Near-Memory Architecture for Accelerating Linear Programming Problems Siddhartha Raman Sundara Raman (The University of Texas at Austin), Lizy John (The University of Texas at Austin), and Jaydeep P. Kulkarni (The University of Texas at Austin)	99
LegoZK: A Dynamically Reconfigurable Accelerator for Zero-Knowledge Proof Zhengbang Yang (University of Chinese Academy of Sciences), Lutan Zhao (University of Chinese Academy of Sciences), Peinan Li (University of Chinese Academy of Sciences), Han Liu (University of Chinese Academy of Sciences), Kai Li (University of Chinese Academy of Sciences), Boyan Zhao (University of Chinese Academy of Sciences), Dan Meng (University of Chinese Academy of Sciences), and Rui Hou (University of Chinese Academy of Sciences)	113

Session 2A: Quantum Slots-1

Reuse-Aware Compilation for Zoned Quantum Architectures Based on Neutral Atoms Wan-Hsuan Lin (University of California, Los Angeles), Daniel Bochen Tan (University of California, Los Angeles; Harvard University), and Jason Cong (University of California, Los Angeles)	. 127
HATT: Hamiltonian Adaptive Ternary Tree for Optimizing Fermion-to-Qubit Mapping Yuhao Liu (University of Pennsylvania), Kevin Yao (University of Pennsylvania), Jonathan Hong (University of Pennsylvania), Julien Froustey (University of California, Berkeley), Ermal Rrapaj (Lawrance Berkeley National Lab), Costin Iancu (Lawrance Berkeley National Lab), Gushu Li (University of Pennsylvania), and Yunong Shi (AWS Quantum	. 143
Technologies)	

Session 2B: Speculating in Vegas - 2

Gaze into the Pattern: Characterizing Spatial Patterns with Internal Temporal Correlations for Hardware Prefetching	73
To Cross, or Not to Cross Pages for Prefetching?	88
Integrating Prefetcher Selection with Dynamic Request Allocation Improves Prefetching 20 Efficiency 20 Mengming Li (Hong Kong University of Science and Technology), Qijun 20 Zhang (Hong Kong University of Science and Technology), Yongqing Ren 20 (Intel), and Zhiyao Xie (Hong Kong University of Science and Technology), Yongqing Ren 21 Technology) 21	04

Session 2C: High Stakes Vision - 1

VR-Pipe: Streamlining Hardware Graphics Pipeline for Volume Rendering Junseo Lee (Seoul National University), Jaisung Kim (Seoul National University), Junyong Park (Seoul National University), and Jaewoong Sim (Seoul National University)	217
IRIS: Unleashing ISP-Software Cooperation to Optimize the Machine Vision Pipeline Raúl Taranco (Universitat Politècnica de Catalunya (UPC), Spain), José-Maria Arnau (Universitat Politècnica de Catalunya (UPC), Spain), and Antonio González (Universitat Politècnica de Catalunya (UPC), Spain)	231
 Uni-Render: A Unified Accelerator for Real-Time Rendering Across Diverse Neural Renderers Chaojian Li (Georgia Institute of Technology), Sixu Li (Georgia Institute of Technology), Linrui Jiang (Georgia Institute of Technology), Jingqun Zhang (Georgia Institute of Technology), and Yingyan Celine Lin (Georgia Institute of Technology) 	246

Session 3A: Quantum Slots-2

 Interleaved Logical Qubits in Atom Arrays	61
Choco-Q: Commute Hamiltonian-Based QAOA for Constrained Binary Optimization	75
BOSS: Blocking Algorithm for Optimizing Shuttling Scheduling in Ion Trap	90
LSQCA: Resource-Efficient Load/Store Architecture for Limited-Scale Fault-Tolerant Quantum Computing	04

Session 3B: Best of CAL

The Importance of Generalizability in Machine Learning for Systems	321
R.I.P. Geomean Speedup Use Equal-Work (Or Equal-Time) Harmonic Mean Speedup Instead S Lieven Eeckhout (Ghent University, Belgium)	322
eDKM: An Efficient and Accurate Train-Time Weight Clustering for Large Language Models Minsik Cho (Apple, USA), Keivan A. Vahid (Apple, USA), Qichen Fu	323

(Apple, USA), Saurabh Adya (Apple, USA), Carlo C. Del Mundo (Apple, USA), Mohammad Rastegari (Apple, USA), Devang Naik (Apple, USA), and Peter Zatloukal (Apple, USA)

Session 3C: High Stakes Vision - 2

Sungbin Kim (Yonsei University, Republic of Korea), Hyunwuk Lee (Yonsei University, Republic of Korea), Wonho Cho (Yonsei University, Republic of Korea), Mincheol Park (Yonsei University, Republic of Korea), and Won Woo Ro (Yonsei University, Republic of Korea)
Gaussian Blending Unit: An Edge GPU Plug-in for Real-Time Gaussian-Based Rendering in
AR/VR
Zhifan Ye (Georgia Institute of Technology), Yonggan Fu (Georgia
Institute of Technology), Jinggun Zhang (Georgia Institute of
Technology), Leshu Li (Georgia Institute of Technology), Yongan Zhang
(Georgia Institute of Technology), Sixu Li (Georgia Institute of
Technology), Cheng Wan (Georgia Institute of Technology), Chenxi Wan
(Georgia Institute of Technology), Chaojian Li (Georgia Institute of
Technology), Sreemanth Prathipati (Georgia Institute of Technology),
and Yingyan Celine Lin (Georgia Institute of Technology)
GSArch: Breaking Memory Barriers in 3D Gaussian Splatting Training via Architectural
Support
Houshu He (Shanghai Jiao Tong University, China), Gang Li (Chinese
Academy of Sciences, China), Fangxin Liu (Shanghai Jiao Tong
University, China), Li Jiang (Shanghai Jiao Tong University, China),
Xiaoyao Liang (Shanghai Jiao Tong University, China), and Zhuoran Song
(Shanghai Jiao Tong University, China)

Session 3D: Securing your Chips

 Palermo: Improving the Performance of Oblivious Memory using Protocol-Hardware Co-Design 3 Haojie Ye (University of Michigan, USA), Yuchen Xia (University of Michigan, USA), Yuhan Chen (University of Michigan, USA), Kuan-Yu Chen (University of Michigan, USA), Yichao Yuan (University of Michigan, USA), Shuwen Deng (Tsinghua University, China), Baris Kasikci (University of Washington, USA), Trevor Mudge (University of Michigan, USA), and Nishil Talati (University of Michigan, USA) 	80
SpecMPK: Efficient In-Process Isolation with Speculative and Secure Permission Update Instruction	94
BrokenSleep: Remote Power Timing Attack Exploiting Processor Idle States	:09
Efficient Memory Side-Channel Protection for Embedding Generation in Machine Learning 4 Muhammad Umar (Cornell University), Akhilesh Parag Marathe (Virginia Tech), Monami Dutta Gupta (Virginia Tech), Shubham Jogprakash Ghosh (Virginia Tech), G. Edward Suh (Cornell University; NVIDIA), and Wenjie Xiong (Virginia Tech)	23

Session 4A: The Winning System - 1

Criticality-Aware Instruction-Centric Bandwidth Partitioning for Data Center Applications
Concord: Rethinking Distributed Coherence for Software Caches in Serverless Environments 458 Jovan Stojkovic (University of Illinois at Urbana-Champaign), Chloe Alverti (University of Illinois at Urbana-Champaign), Alan Andrade (University of Illinois at Urbana-Champaign), Nikoleta Iliakopoulou (University of Illinois at Urbana-Champaign), Hubertus Franke (IBM Research), Tianyin Xu (University of Illinois at Urbana-Champaign), and Josep Torrellas (University of Illinois at Urbana-Champaign)
Grad: Intelligent Microservice Scaling by Harnessing Resource Fungibility

Chengzhong Xu (University of Macau)

Session 4B: Doubling-down with PIM - 1

Zhen He (Tsinghua University, China), Yiqi Wang (Tsinghua University, China), Zihan Wu (Tsinghua University, China), Shaojun Wei (Tsinghua University, China), Yang Hu (Tsinghua University, China), Fengbin Tu (The Hong Kong University of Science and Technology, China), and Shouyi Yin (Tsinghua University, China)

 AsyncDIMM: Achieving Asynchronous Execution in DIMM-Based Near-Memory Processing 518 Liyan Chen (Shanghai Jiao Tong University), Dongxu Lyu (Shanghai Jiao Tong University), Jianfei Jiang (Shanghai Jiao Tong University), Qin Wang (Shanghai Jiao Tong University), Zhigang Mao (Shanghai Jiao Tong University), and Naifeng Jing (Shanghai Jiao Tong University)

Session 4C: Stacking the Layers

SoMa: Identifying, Exploring, and Understanding the DRAM Communication Scheduling Space for DNN Accelerators Jingwei Cai (Tsinghua University), Xuan Wang (Xi'an Jiaotong University; IIISCT), Mingyu Gao (Tsinghua University; Shanghai AI Laboratory; Shanghai Qi Zhi Institute), Sen Peng (Xi'an Jiaotong University; IIISCT), Zijian Zhu (Tsinghua University), Yuchen Wei (Tsinghua University), Zuotong Wu (Xi'an Jiaotong University; IIISCT), and Kaisheng Ma (Tsinghua University)	533
Adyna: Accelerating Dynamic Neural Networks with Adaptive Scheduling Zhiyao Li (Tsinghua University), Bohan Yang (University of Science and Technology of China), Jiaxiang Li (University of Toronto), Taijie Chen (Tsinghua University), Xintong Li (Tsinghua University), and Mingyu Gao (Tsinghua University; Shanghai AI Laboratory; Shanghai Qi Zhi Institute)	549
EDA: Energy-Efficient Inter-Layer Model Compilation for Edge DNN Inference Acceleration Bo Ren Pao (National Yang Ming Chiao Tung University), I-Chia Chen (National Yang Ming Chiao Tung University), En-Hao Chang (Google), and Tsung Tai Yeh (National Yang Ming Chiao Tung University)	563

Session 5A: Stashing your Winnings

 SkyByte: Architecting an Efficient Memory-Semantic CXL-Based SSD with OS and Hardware Co-Design
Zebra: Efficient Redundant Array of Zoned Namespace SSDs Enabled by Zone Random Write Area (ZRWA)
Reviving In-Storage Hardware Compression on ZNS SSDs through Host-SSD Collaboration 608 Yingjia Wang (The Chinese University of Hong Kong), Tao Lu (DapuStor Corporation), Yuhong Liang (The Chinese University of Hong Kong), Xiang Chen (DapuStor Corporation), and Ming-Chang Yang (The Chinese University of Hong Kong)

Session 5B: Doubling-down with PIM - 2

UniNDP: A Unified Compilation and Simulation Tool for Near DRAM Processing Architectures ... 624 Tongxin Xie (Tsinghua University), Zhenhua Zhu (Tsinghua University; HKUST), Bing Li (Chinese Academy of Sciences), Yukai He (Capital Normal University), Cong Li (Peking University), Guangyu Sun (Peking University), Huazhong Yang (Tsinghua University), Yuan Xie (HKUST), and Yu Wang (Tsinghua University)

 Piccolo: Large-Scale Graph Processing with Fine-Grained In-Memory Scatter-Gather
 GOPIM: GCN-Oriented Pipeline Optimization for PIM Accelerators

Session 5C: Stacking the Layers (Quantization)

LUT-DLA: Lookup Table as Efficient Extreme Low-Bit Deep Learning Accelerator Guoyu Li (University of Chinese Academy of Sciences; Microsoft Research), Shengyu Ye (Microsoft Research), Chunyun Chen (NTU Singapore), Yang Wang (Microsoft Research), Fan Yang (Microsoft Research), Ting Cao (Microsoft Research), Cheng Liu (University of Chinese Academy of Sciences), Mohamed M. Sabry Aly (NTU Singapore), and Mao Yang (Microsoft Research)	671
Exploring the Performance Improvement of Tensor Processing Engines through Transformation	
in the Bit-Weight Dimension of MACs	685
Qizhe Wu (University of Science and Technology of China), Huawen Liang	
(University of Science and Technology of China), Yuchen Gui	
(University of Science and Technology of China), Zhichen Zeng	
(University of Science and Technology of China; University of	
Washington), Zerong He (University of Science and Technology of	
China), Linfeng Tao (University of Science and Technology of China),	
Xiaotian Wang (University of Science and Technology of China; Raytron	
Technology), Letian Zhao (University of Science and Technology of	
China), Zhaoxi Zeng (University of Science and Technology of China),	
Wei Yuan (University of Science and Technology of China), Wei Wu	and Technology of China), Wei Wu
(University of Science and Technology of China), and Xi Jin	
(University of Science and Technology of China)	

Dongyun Rum (Ponung University of Science und Technology (POSTECH)),
Myeongji Yun (Pohang University of Science and Technology (POSTECH)),
Sunwoo Yoo (Pohang University of Science and Technology (POSTECH)),
Seungwoo Hong (Pohang University of Science and Technology (POSTECH)),
Zhengya Zhang (University of Michigan), and Youngjoo Lee (Pohang
University of Science and Technology (POSTECH))

Session 6A: Cacheno Royale

From Optimal to Practical: Efficient Micro-op Cache Replacement Policies for Data Center Applications Kan Zhu (University of Washington), Yilong Zhao (UC Berkeley), Yufei Gao (University of Washington; Tsinghua University), Peter Braun (UC Santa Cruz), Tanvir Ahmed Khan (Columbia University), Heiner Litz (UC Santa Cruz), Baris Kasikci (University of Washington), and Shuwen Deng (Tsinghua University)	. 716
Rethinking Dead Block Prediction for Intermittent Computing Gan Fang (Purdue University) and Changhee Jung (Purdue University)	. 732
Efficient Caching with a Tag-Enhanced DRAM Maryam Babaie (University of California, Davis), Ayaz Akram (Samsung Electronics), Wendy Elsasser (Rambus Inc.), Brent Haukness (Rambus Inc.), Michael R. Miller (Rambus Inc.), Taeksang Song (Samsung Electronics), Thomas Vogelsang (Rambus Inc.), Steven C. Woo (Rambus Inc.), and Jason Lowe-Power (University of California, Davis)	. 745

Session 6B: Next Big Bets - 2

PROCA: Programmable Probabilistic Processing Unit Architecture with Accept/Reject	
Prediction & Multicore Pipelining for Causal Inference	.761
Yihan Fu (Peking University, China), Anjunyi Fan (Peking University,	
China), Wenshuo Yue (Peking University, China), Hongxiao Zhao (Peking	
University, China), Daijing Šhi (Peking University, China), Qiuping Wu	
(Peking University, China), Jiayi Li (Peking University, China),	
Xiangyu Zhang (Independent Researcher), Yaoyu Tao (Peking University,	
China), Yuchao Yang (Peking University, China; Chinese Institute for	
Brain Research, China), and Bonan Yan (Peking University, China)	
CogSys: Efficient and Scalable Neurosymbolic Cognition System via Algorithm-Hardware	
Co-Design	775
Zishen Wan (Georgia Institute of Technology, GA), Hanchen Yang	
(Georgia Institute of Technology, GA), Ritik Raj (Georgia Institute of	
Technology, GA), Che-Kai Liu (Georgia Institute of Technology, GA),	
Ananda Samajdar (IBM Research, NY), Arijit Raychowdhury (Georgia	
Institute of Technology, GA), and Tushar Krishna (Georgia Institute of	
Technology, GA)	

 NeuVSA: A Unified and Efficient Accelerator for Neural Vector Search	.790
Chen (Huawei Technologies Co., Ltd., China), and Gong Zhang (Huawei Technologies Co., Ltd., China) Session 6C: Stacking the Layers (Sparsity) - 1	
Prosperity: Accelerating Spiking Neural Networks via Product Sparsity Chiyue Wei (Duke University), Cong Guo (Duke University), Feng Cheng (Duke University), Shiyu Li (Duke University), Hao Frank Yang (Johns Hopkins University), Hai Helen Li (Duke University), and Yiran Chen (Duke University)	. 806
Bit-Slice Architecture for DNN Acceleration with Slice-Level Sparsity Enhancement and Exploitation Insu Choi (Yonsei University, South Korea), Young-Seo Yoon (Yonsei University, South Korea), and Joon-Sung Yang (Yonsei University, South Korea)	. 821
A Hardware-Software Design Framework for SpMV Acceleration with Flexible Access Pattern Portfolio	. 836

Session 7A: Hammering the Odds - 1

Variable Read Disturbance: An Experimental Analysis of Temporal Variation in DRAM Read	
Disturbance	9
Ataberk Olgun (ETH Zurich), F. Nisa Bostanci (ETH Zurich), Ismail Emir	
Yüksel (ETH Zurich), Oğuzhan Canpolat (ETH Zurich), Haocong Luo (ETH	
Zurich), Geraldo F. Oliveira (ETH Żurich), A. Giray Yağlikçi (EŤH	
Zurich), Minesh Patel (Rutgers University), and Onur Mutlu (ETH	
Zurich)	
Understanding RowHammer Under Reduced Refresh Latency: Experimental Analysis of Real DRAM	1
Chips and Implications on Future Solutions	57
Yahya Can Tuğrul (ETH Zürich; TOBB University of Economics and	
Technology), A. Giray Yağlıkçı (ETH Zürich), İsmail Emir Yüksel (ETH	
Zürich), Ataberk Olgun (ETH Zürich), Oğuzhan Canpolat (ETH Zürich;	
TOBB University of Economics and Technology), Nisa Bostancı (ETH	
Zürich), Mohammad Sadrosadati (ETH Zürich), Oğuz Ergin (University of	
Sharjah; TOBB University of Economics and Technology), and Onur Mutlu	
(ETH Zürich)	

Chronus: Understanding and Securing the Cutting-Edge Industry Solutions to DRAM Read	
Disturbance	887
Oğuzhan Canpolat (ETH Zürich; TOBB University of Economics and	
Technology), A. Giray Yağlıkçı (ETH Zürich), Geraldo F. Oliveira (ETH	
Zürich), Ataberk Olgun (ETH Zürich), Nisa Bostancı (ETH Zürich),	
Ismail Emir Yuksel (ETH Zürich), Haocong Luo (ETH Zürich), Oğuz Ergin	
(University of Sharjah; TOBB University of Economics and Technology),	
and Onur Mutlu (ETH Zürich)	

Session 7B: Traversing Winning Paths - 1

NOVA: A Novel Vertex Management Architecture for Scalable Graph Processing
 MeHyper: Accelerating Hypergraph Neural Networks by Exploring Implicit Dataflows
Cambricon-DG: An Accelerator for Redundant-Free Dynamic Graph Neural Networks Based on Nonlinear Isolation

Session 7C: Stacking the Layers (Sparsity) - 2

 TB-STC: Transposable Block-Wise N:M Structured Sparse Tensor Core
CROSS: Compiler-Driven Optimization of Sparse DNNs using Sparse/Dense Computation Kernels 963
Fangxin Liu (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Shiyuan Huang (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Ning Yang (Shanghai Jiao Tong University), Zongwu Wang (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Haomin Li (Shanghai Jiao Tong University), and Li Jiang (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute)
AccelES: Accelerating Top-K SpMV for Embedding Similarity via Low-Bit Pruning

Session 8A: Hammering the Odds - 2

AutoRFM: Scaling Low-Cost In-DRAM Trackers to Ultra-Low Rowhammer Thresholds
DAPPER: A Performance-Attack-Resilient Tracker for RowHammer Defense
OPRAC: Towards Secure and Practical PRAC-Based Rowhammer Mitigation using Priority Oueues

QPRAC: Towards Secure and Practical PRAC-Based Rowhammer Mitigation using Priority Queues ... 1021

Jeonghyun Woo (University of British Columbia), Shaopeng Chris Lin (University of Toronto), Prashant J. Nair (University of British Columbia), Aamer Jaleel (NVIDIA), and Gururaj Saileshwar (University of Toronto)

Session 8B: Traversing Winning Paths - 2

Mithril: A Scalable System for Deep GNN Training	1052
Jingji Chen (Purdue University), Zhuoming Chen (Carnegie Mellon	
University), and Xuehai Qian (Tsinghua University)	
Buffalo: Enabling Large-Scale GNN Training via Memory-Efficient Bucketization	1066
Shuangyan Yang (University of California, Merced), Minjia Zhang	

(University of Illinois Urbana-Champaign), and Dong Li (University of California, Merced)

Session 8C: Viva Las Learning Models - 1

BitMoD: Bit-Serial Mixture-of-Datatype LLM Acceleration
FIGLUT: An Energy-Efficient Accelerator Design for FP-INT GEMM using Look-Up Tables 1098 Gunho Park (Pohang University of Science and Technology (POSTECH); NAVER Cloud), Hyeokjun Kwon (Pohang University of Science and Technology (POSTECH)), Jiwoo Kim (Pohang University of Science and Technology (POSTECH)), Jeongin Bae (NAVER Cloud), Baeseong Park (NAVER Cloud), Dongsoo Lee (NAVER Cloud), and Youngjoo Lee (Pohang University of Science and Technology (POSTECH))
 M-ANT: Efficient Low-Bit Group Quantization for LLMs via Mathematically Adaptive Numerical Type

Session 9A: The Poker Face of FHE

EFFACT: A Highly Efficient Full-Stack FHE Acceleration Platform
 Anaheim: Architecture and Algorithms for Processing Fully Homomorphic Encryption in Memory 1158 Jongmin Kim (Seoul National University), Sungmin Yun (Seoul National University), Hyesung Ji (Seoul National University), Wonseok Choi (Seoul National University), Sangpyo Kim (Seoul National University), and Jung Ho Ahn (Seoul National University)
 Hydra: Scale-Out FHE Accelerator Architecture for Secure Deep Learning on FPGA
 WarpDrive: GPU-Based Fully Homomorphic Encryption Acceleration Leveraging Tensor and CUDA Cores

Session 9B: Industry Session

 MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from μWatts to MWatts for Sustainable AI
Enterprise Class Modular Cache Hierarchy

Predicting DRAM-Caused Risky VMs in Large-Scale Clouds
Enhancing Large-Scale AI Training Efficiency: The C4 Solution for Real-Time Anomaly
Detection and Communication Optimization
Jianbo Dong (Alibaba Group), Bin Luo (Alibaba Group), Jun Zhang
(Alibaba Group), Pengcheng Zhang (Alibaba Group), Fei Feng (Alibaba
Group), Yikai Zhu (Alibaba Group), Ang Liu (Alibaba Group), Zian Chen
(Alibaba Group), Yi Shi (Alibaba Group), Hairong Jiao (Alibaba Group),
Gang Lu (Alibaba Group), Yu Guan (Alibaba Group), Ennan Zhai (Alibaba
Group), Wencong Xiao (Alibaba Group), Hanyu Zhao (Alibaba Group), Man
Yuan (Alibaba Group), Siran Yang (Alibaba Group), Xiang Li (Alibaba
Group), Jiamang Wang (Alibaba Group), Rui Men (Alibaba Group), Jianwei
Zhang (Alibaba Group), Chang Zhou (Alibaba Group), Dennis Cai (Alibaba
Group), Yuan Xie (Alibaba Group; Hong Kong University of Science and
Technology), and Binzhang Fu (Alibaba Group)
Revisiting Reliability in Large-Scale Machine Learning Research Clusters
Carole-Jean Wu (FAIR at Meta)

Session 9C: Finding the Best Table

HILP: Accounting for Workload-Level Parallelism in System-on-Chip Design Space Exploration 1275 Joseph Rogers (Norwegian University of Science and Technology (NTNU)), Lieven Eeckhout (Ghent University), and Magnus Jahre (Norwegian University of Science and Technology (NTNU))
CORDOBA: Carbon-Efficient Optimization Framework for Computing Systems
Architecting Space Microdatacenters: A System-level Approach
 ARTEMIS: Agile Discovery of Efficient Real-Time Systems-on-Chips in the Heterogeneous Era 1320 Subhankar Pal (IBM T. J. Watson Research Center, USA), Aporva Amarnath (IBM T. J. Watson Research Center, USA), Behzad Boroujerdian (University of Texas, USA), Augusto Vega (IBM T. J. Watson Research Center, USA), Alper Buyuktosunoglu (IBM T. J. Watson Research Center, USA), John-David Wellman (IBM T. J. Watson Research Center, USA), Vijay Janapa Reddi (Harvard University, USA), and Pradip Bose (IBM T. J. Watson Research Center, USA)

Session 10A: The Winning System - 2

Jovan Stojkovic (University of Illinois at Urbana-Champaign), Chaojie Zhang (Microsoft Azure Řesearch - Systems), Íñigo Goiri (Microsoft Azure Research - Systems), Josep Torrellas (University of Illinois at Urbana-Champaign), and Esha Choukse (Microsoft Azure Research -Systems) Andreas Kosmas Kakolyris (ETH Zürich; National Technical University of Athens), Dimosthenis Masouros (National Technical University of Athens), Petros Vavaroutsos (National Technical University of Athens), Sotirios Xydis (National Technical University of Athens), and Dimitrios Soudris (National Technical University of Athens) RpcNIC: Enabling Efficient Datacenter RPC Offloading on PCIe-Attached SmartNICs 1379 Jie Zhang (Zhejiang University), Hongjing Huang (Zhejiang University), Xuzheng Chen (Zhejiang University), Xiang Li (Zhejiang University), Jieru Zhao (Shanghai Jiao Tong University), Ming Liu (University of Wisconsin-Madison), and Zeke Wang (Zhejiang University) NVMePass: A Lightweight, High-Performance and Scalable NVMe Virtualization Architecture Yiquan Chen (Zhejiang University; Alibaba Group), Zhen Jin (Zhejiang University; Alibaba Group), Yijing Wang (Alibaba Group), Yi Chen (University of Michigan), Jiexiong Xu (Zhejiang University), Hao Yu (Alibaba Group), Jinlong Chen (Alibaba Group), Wenhai Lin (Zhejiang University), Kanghua Fang (Alibaba Group), Keyao Zhang (Zhejiang University), Chengkun Wei (Zhejiang University), Qiang Liu (Alibaba Group), Yuan Xie (HKUST), and Wenzhi Chen (Zhejiang University)

Session 10B: All in on GPUs - 1

Warped-Compaction: Maximizing GPU Register File Bandwidth Utilization via Operand Compaction	1408
Eunbi Jeong (Ewha Womans University, Korea), Ipoom Jeong (Yonsei University, Korea), Myung Kuk Yoon (Ewha Womans University, Korea), and Nam Sung Kim (University of Illinois Urbana-Champaign, U.S.A.)	
Cooperative Warp Execution in Tensor Core for RISC-V GPGPU Abubakr Nada (imec), Giuseppe Maria Sarda (imec; KU Leuven), and Erwan Lenormand (imec)	1422

SparseWeaver: Converting Sparse Operations as Dense Operations on GPUs for Graph Workloads..... 1437

Shinnung Jeong (Yonsei University, South Korea), Liam Paul Cooper	
(Georgia Institute of Technology, USA), Ju Min Lee (Yonsei University,	
South Korea), Heelim Choi (Yonsei University, South Korea), Nicholas	
Parnenzini (Georgia Institute of Technology, USA), Chihyo Ahn (Georgia	
Institute of Technology, USA), Yongwoo Lee (Yonsei University, South	
Korea), Hanjun Kim (Yonsei University, South Korea), and Hyesoon Kim	
(Georgia Institute of Technology, USA)	
HSMU-SpGEMM: Achieving High Shared Memory Utilization for Parallel Sparse General	
Matrix-Matrix Multiplication on Modern GPUs	2
Min Wu (Hunan University, China), Huizhang Luo (Hunan University,	
China), Fenfang Li (Hunan University, China), Yiran Zhang (Arizona	
State University, United States), Zhuo Tang (Hunan University, China),	
Kenli Li (Hunan University, China), Jeff Zhang (Arizona State	
· · · ·	
University, United States), and Chubo Liu (Hunan University, China)	

Session 10C: Viva Las Learning Models - 2

Anda: Unlocking Efficient LLM Inference with a Variable-Length Grouped Activation Data	167
Chao Fang (Nanjing University, China; KU Leuven, Belgium), Man Shi (KU	407
Leuven, Belgium), Robin Geens (KU Leuven, Belgium), Arne Symons (KU	
Leuven, Belgium), Zhongfeng Wang (Nanjing University, China), and Marian Varholst (KUL auzum, Balgium)	
Muriun Verneisi (RCI Leuden, Belgium)	
LAD: Efficient Accelerator for Generative Inference of LLM with Locality Aware Decoding 14	482
Haoran Wang (Chinese Academy of Sciences; University of Chinese	
Academy of Sciences), Yuming Li (Chinese Academy of Sciences;	
University of Chinese Academy of Sciences), Haobo Xu (Chinese Academy	
of Sciences), Ying Wang (Chinese Academy of Sciences), Liqi Liu	
(Chinese Academy of Sciences; University of Chinese Academy of	
Sciences). Jun Yang (Chinese Academy of Sciences), and Yinhe Han	

(Chinese Academy of Sciences)

VQ-LLM: High-Performance Code Generation for Vector Quantization Augmented LLM Inference 1496

Zihan Liu (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Xinhao Luo (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Junxian Guo (Shanghai Jiao Tong University), Wentao Ni (Shanghai Jiao Tong University), Yangjie Zhou (Shanghai Jiao Tong University), Yue Guan (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Cong Guo (Duke University), Weihao Cui (Shanghai Jiao Tong University; National University of Singapore), Yu Feng (Shanghai Jiao Tong University), Minyi Guo (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute), Yuhao Zhu (University of Rochester), Minjia Zhang (University of Illinois Urbana-Champaign), Chen Jin (Magik Compute), and Jingwen Leng (Shanghai Jiao Tong University; Shanghai Qi Zhi Institute)

InstAttention: In-Storage Attention Offloading for Cost-Effective Long-Context LLM
Inference
Xiurui Pan (Peking University), Endian Li (Peking University), Qiao Li
(University of Electronic Science and Technology of China), Shengwen
Liang (Chinese Academy of Sciences), Yizhou Shan (Huawei Cloud), Ke
Zhou (Wuhan National Laboratory for Optoelectronics of Huazhong
University of Science and Technology), Yingwei Luo (Peking
University), Xiaolin Wang (Peking University), and Jie Zhang (Peking
University)

Session 10D: Colluding to Even the Odds

TidalMesh: Topology-Driven AllReduce Collective Communication for Mesh Topology 1526 Dongkyun Lim (KAIST) and John Kim (KAIST)
Push Multicast: A Speculative and Coherent Interconnect for Mitigating Manycore CPU Communication Bottleneck
PIMnet: A Domain-Specific Network for Efficient Collective Communication in Scalable PIM 1557 Hyojun Son (KAIST), Gilbert Jonatan (KAIST), Xiangyu Wu (KAIST), Haeyoon Cho (KAIST), Kaustubh Shivdikar (Northeastern University), José L. Abellán (Universidad de Murcia), Ajay Joshi (Boston University), David Kaeli (Northeastern University), and John Kim (KAIST)
 EIGEN: Enabling Efficient 3DIC Interconnect with Heterogeneous Dual-Layer Network-on-Active-Interposer

Session 11A: The Winning System - 3

Ariadne: A Hotness-Aware and Size-Adaptive Compressed Swap Technique for Fast Application
Relaunch and Reduced CPU Usage on Mobile Devices
Yu Liang (ETH Zurich, Switzerland), Aofeng Shen (ETH Zurich,
Switzerland), Chun Jason Xue (MBZUAİ, Ünited Arab Emirates), Riwei Pan
(City University of Hong Kong, Hong Kong), Haiyu Mao (King's College
London, UK), Nika Mansouri Ghiasi (ETH Zurich, Switzerland), Qingcai
Jiang (ETH Zurich, Switzerland and University of Science and
Technology of China, China), Rakesh Nadig (ETH Zurich, Switzerland),
Lei Li (City University of Hong Kong, Hong Kong), Rachata
Ausavarungnirun (MangoBoost, UŠA), Mohammad Sadrosadati (ETH Zurich,
Switzerland), and Onur Mutlu (ETH Zurich, Switzerland)
Gemina: A Coordinated and High-Performance Memory Deduplication Engine
Zhehua Zhang (Xiamen University, China), Suzhen Wu (Xiamen University,
China), Wenyan You (Xiamen University, China), Chunfeng Du (Xiamen

University, China), and Bo Mao (Xiamen University, China)

No Rush in Executing Atomic Instructions	1618
Ashkan Asgharzadeh (University of Murcia), Josué Feliu (Universitat	
Politècnica de València), Manuel E. Acacio (University of Murcia),	
Stefanos Kaxiras (Uppsala University), and Alberto Ros (University of	
Murcia)	

Machine Learning-Guided Memory Optimization for DLRM Inference on Tiered Memory 1631 Jie Ren (William & Mary), Bin Ma (University of California, Merced), Shuangyan Yang (University of California, Merced), Benjamin Francis (Meta), Ehsan K. Ardestani (Meta), Min Si (Meta), and Dong Li (University of California, Merced)

Session 11B: All in on GPUs - 2

Let-Me-In: (Still) Employing In-Pointer Bounds Metadata for Fine-Grained GPU Memory Safety.. 1648 Jaewon Lee (Georgia Institute of Technology), Euijun Chung (Georgia Institute of Technology), Saurabh Singh (Georgia Institute of Technology), Seonjin Na (Georgia Institute of Technology), Yonghae Kim (Arm), Jaekyu Lee (Intel), and Hyesoon Kim (Georgia Institute of Technology)

Marching Page Walks: Batching and Concurrent Page Table Walks for Enhancing GPU Throughput.... 1662

Jiwon Lee (Yonsei University, Korea), Gun Ko (Yonsei University, Korea), Myung Kuk Yoon (Ewha Womans University, Korea), Ipoom Jeong (Yonsei University, Korea), Yunho Oh (Korea University, Korea), and Won Woo Ro (Yonsei University, Korea)

OASIS: Object-Aware Page Management for Multi-GPU Systems	678
Yueqi Wang (University of Pittsburgh), Bingyao Li (University of	
Pittsburgh), Mohamed Tarek Ibn Ziad (NVIDIA), Lieven Eeckhout (Ghent	
University), Jun Yang (University of Pittsburgh), Aamer Jaleel	
(NVIDIA), and Xulong Tang (University of Pittsburgh)	
NearFetch: Saving Inter-Module Bandwidth in Many-Chip-Module GPUs	693
Xia Zhao (Academy of Military Science), Guangda Zhang (Academy of	

Military Science), Lu Wang (Academy of Military Science), Shiqing Zhang (Academy of Military Science), and Huadong Dai (Academy of Military Science)

Session 11C: Viva Las Learning Models (with PIM)

Yeonhong Park (Seoul National University), and Jae W. Lee (Seoul National University)

 Lincoln: Real-Time 50~100B LLM Inference on Consumer Devices with LPDDR-Interfaced, Compute-Enabled Flash Memory	34
Make LLM Inference Affordable to Everyone: Augmenting GPU Memory with NDP-DIMM 175 Lian Liu (Chinese Academic of Sciences; University of Chinese Academy of Sciences; Zhongguancun Laboratory), Shixin Zhao (Chinese Academic of Sciences; University of Chinese Academy of Sciences), Bing Li (Chinese Academy of Sciences), Haimeng Ren (Chinese Academic of Sciences; ShanghaiTech University), Zhaohui Xu (Chinese Academic of Sciences; ShanghaiTech University), Mengdi Wang (Chinese Academic of Sciences; University of Chinese Academy of Sciences), Xiaowei Li (Chinese Academic of Sciences; University of Chinese Academy of Sciences; Zhongguancun Laboratory), Yinhe Han (Chinese Academic of Sciences; University of Chinese Academy of Sciences), and Ying Wang (Chinese Academic of Sciences; University of Chinese Academy of Sciences; University of Chinese Academy of Sciences), and Ying Wang (Chinese Academic of Sciences; University of Chinese Academy of Sciences)	51

Author Index