

2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA 2023)

**Montreal, Quebec, Canada
25 February - 1 March 2023**

Pages 1-650



**IEEE Catalog Number: CFP23013-POD
ISBN: 978-1-6654-7653-9**

**Copyright © 2023 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:	CFP23013-POD
ISBN (Print-On-Demand):	978-1-6654-7653-9
ISBN (Online):	978-1-6654-7652-2
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

CURRAN ASSOCIATES INC.
proceedings
.com

TABLE OF CONTENTS

SESSION 1A: NEURAL NETWORKS AND ACCELERATORS 1

SGCN: Exploiting Compressed-Sparse Features in Deep Graph Convolutional Network Accelerators.....	1
<i>Mingi Yoo, Jaeyong Song, Jounghoo Lee, Namhyung Kim, Youngsok Kim, Jinho Lee</i>	
PhotoFourier: A Photonic Joint Transform Correlator-Based Neural Network Accelerator.....	15
<i>Shurui Li, Hangbo Yang, Chee Wei Wong, Volker J. Sorger, Puneet Gupta</i>	
INCA: Input-Stationary Dataflow at Outside-The-Box Thinking About Deep Learning Accelerators	29
<i>Bokyung Kim, Shiyu Li, Hai Li</i>	
GROW: A Row-Stationary Sparse-Dense GEMM Accelerator for Memory-Efficient Graph Convolutional Neural Networks.....	42
<i>Ranggi Hwang, Minhoo Kang, Jiwon Lee, Dongyun Kam, Youngjoo Lee, Minsoo Rhu</i>	
Logical/Physical Topology-Aware Collective Communication in Deep Learning Training	56
<i>Sanghun Cho, Hyojun Son, John Kim</i>	
Sibia: Signed Bit-Slice Architecture for Dense DNN Acceleration with Slice-Level Sparsity Exploitation	69
<i>Dongseok Im, Gwangtae Park, Zhiyong Li, Junha Ryu, Hoi-Jun Yoo</i>	

SESSION 1B: NVRAM AND HYBRID MEMORY

AstriFlash a Flash-Based System for Online Services	81
<i>Siddharth Gupta, Yunho Oh, Lei Yan, Mark Sutherland, Abhishek Bhattacharjee, Babak Falsafi, Peter Hsu</i>	
Thoth: Bridging the Gap Between Persistently Secure Memories and Memory Interfaces of Emerging NVMs	94
<i>Xijing Han, James Tuck, Amro Awad</i>	
Multi-Granularity Shadow Paging with NVM Write Optimization for Crash-Consistent Memory-Mapped I/O.....	108
<i>Hongchao Du, Qiao Li, Riwei Pan, Tei-Wei Kuo, Chun Jason Xue</i>	
MGC: Multiple-Gray-Code for 3D NAND Flash Based High-Density SSDs	122
<i>Yina Lv, Liang Shi, Qiao Li, Congming Gao, Yunpeng Song, Longfei Luo, Youtao Zhang</i>	
Baryon: Efficient Hybrid Memory Management with Compression and Sub-Blocking.....	137
<i>Yiwei Li, Mingyu Gao</i>	
Root Crash Consistency of SGX-Style Integrity Trees in Secure Non-Volatile Memory Systems	152
<i>Jianming Huang, Yu Hua</i>	

SESSION 1C: CACHING AND MEMORY MANAGEMENT

ACIC: Admission-Controlled Instruction Cache.....	165
<i>Yunjin Wang, Chia-Hao Chang, Anand Sivasubramaniam, Niranjan Soundararajan</i>	

Compression-Aware and Performance-Efficient Insertion Policies for Long-Lasting Hybrid LLCs	179
<i>Carlos Escuin, Asif Ali Khan, Pablo Ibáñez, Teresa Monreal, Jeronimo Castrillon, Víctor Viñals</i>	
NOMAD: Enabling Non-Blocking OS-Managed DRAM Cache Via Tag-Data Decoupling	193
<i>Youngin Kim, Hyeonjin Kim, William J. Song</i>	
Safety Hints for HTM Capacity Abort Mitigation.....	206
<i>Anirudh Jain, Divya Kiran Kadiyala, Alexandros Daglis</i>	
ICache: An Importance-Sampling-Informed Cache for Accelerating I/O-Bound DNN Model Training	220
<i>Weijian Chen, Shuibing He, Yaowen Xu, Xuechen Zhang, Siling Yang, Shuang Hu, Xian-He Sun, Gang Chen</i>	
Are Randomized Caches Truly Random? Formal Analysis of Randomized-Partitioned Caches.....	233
<i>Anirban Chakraborty, Sarani Bhattacharya, Sayandeep Saha, Debdeep Mukhopadhyay</i>	

SESSION 2A: ACCELERATORS

HIRAC: A Hierarchical Accelerator with Sorting-Based Packing for SpGEMMs in DNN Applications.....	247
<i>Hesam Shabani, Abhishek Singh, Bishoy Youhana, Xiaochen Guo</i>	
VEGETA: Vertically-Integrated Extensions for Sparse/Dense GEMM Tile Acceleration on CPUs.....	259
<i>Geonhwa Jeong, Sana Damani, Abhimanyu Rajeshkumar Bambhaniya, Eric Qin, Christopher J. Hughes, Sreenivas Subramoney, Hyesoon Kim, Tushar Krishna</i>	
ViTCoD: Vision Transformer Acceleration Via Dedicated Algorithm and Accelerator Co-Design	273
<i>Haoran You, Zhanyi Sun, Huihong Shi, Zhongzhi Yu, Yang Zhao, Yongan Zhang, Chaojian Li, Baopu Li, Yingyan Lin</i>	
Leveraging Domain Information for the Efficient Automated Design of Deep Learning Accelerators	287
<i>Chirag Sakhuja, Zhan Shi, Calvin Lin</i>	
DIMM-Link: Enabling Efficient Inter-DIMM Communication for Near-Memory Processing.....	302
<i>Zhe Zhou, Cong Li, Fan Yang, Guangyu Sun</i>	

SESSION 2B: SECURITY

AutoCAT: Reinforcement Learning for Automated Exploration of Cache-Timing Attacks.....	317
<i>Mulong Luo, Wenjie Xiong, Geunbae Lee, Yueying Li, Xiaomeng Yang, Amy Zhang, Yuandong Tian, Hsien-Hsin S. Lee, G. Edward Suh</i>	
SHADOW: Preventing Row Hammer in DRAM with Intra-Subarray Row Shuffling	333
<i>Minbok Wi, Jaehyun Park, Seoyoung Ko, Michael Jaemin Kim, Nam Sung Kim, Eojin Lee, Jung Ho Ahn</i>	
Efficient Distributed Secure Memory with Migratable Merkle Tree.....	347
<i>Erhu Feng, Dong Du, Yubin Xia, Haibo Chen</i>	
AB-ORAM: Constructing Adjustable Buckets for Space Reduction in Ring ORAM.....	361
<i>Mehrnoosh Raoufi, Jun Yang, Xulong Tang, Youtao Zhang</i>	

Scalable and Secure Row-Swap: Efficient and Safe Row Hammer Mitigation in Memory Systems	374
<i>Jeonghyun Woo, Gururaj Saileshwar, Prashant J. Nair</i>	

SESSION 2C: APPLICATIONS 1

Post0-VR: Enabling Universal Realistic Rendering for Modern VR Via Exploiting Architectural Similarity and Data Sharing	390
<i>Yu Wen, Chenhao Xie, Shuaiwen Leon Song, Xin Fu</i>	
ParallelNN: A Parallel Octree-Based Nearest Neighbor Search Accelerator for 3D Point Clouds.....	403
<i>Faquan Chen, Rendong Ying, Jianwei Xue, Fei Wen, Peilin Liu</i>	
ViTALiTy: Unifying Low-Rank and Sparse Approximation for Vision Transformer Acceleration with a Linear Taylor Attention.....	415
<i>Jyotikrishna Dass, Shang Wu, Huihong Shi, Chaojian Li, Zhifan Ye, Zhongfeng Wang, Yingyan Lin</i>	
CTA: Hardware-Software Co-Design for Compressed Token Attention Mechanism	429
<i>Haoran Wang, Haobo Xu, Ying Wang, Yinhe Han</i>	
HeatViT: Hardware-Efficient Adaptive Token Pruning for Vision Transformers.....	442
<i>Peiyan Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Kenneth Liu, Zhenglun Kong, Xin Meng, Zhengang Li, Xue Lin, Zhenman Fang, Yanzhi Wang</i>	

SESSION 3B: DATACENTERS AND HPC

Trans-FW: Short Circuiting Page Table Walk in Multi-GPU Systems Via Remote Forwarding.....	456
<i>Bingyao Li, Jieming Yin, Anup Holey, Youtao Zhang, Jun Yang, Xulong Tang</i>	
Ah-Q: Quantifying and Handling the Interference Within a Datacenter from a System Perspective.....	471
<i>Yuhang Liu, Xin Deng, Jiapeng Zhou, Mingyu Chen, Yungang Bao</i>	
Market Mechanism-Based User-In-The-Loop Scalable Power Oversubscription for HPC Systems	485
<i>Md Rajib Hossen, Kishwar Ahmed, Mohammad A. Islam</i>	
Rambda: RDMA-Driven Acceleration Framework for Memory-Intensive μ s-Scale Datacenter Applications.....	499
<i>Yifan Yuan, Jinghan Huang, Yan Sun, Tianchen Wang, Jacob Nelson, Dan R. K. Ports, Yipeng Wang, Ren Wang, Charlie Tai, Nam Sung Kim</i>	

SESSION 3C: GPUS

FinePack: Transparently Improving the Efficiency of Fine-Grained Transfers in Multi-GPU Systems.....	516
<i>Harini Muthukrishnan, Daniel Lustig, Oreste Villa, Thomas Wenisch, David Nellans</i>	
Mitigating GPU Core Partitioning Performance Effects	530
<i>Aaron Barnes, Fangjia Shen, Timothy G. Rogers</i>	
Plutus: Bandwidth-Efficient Memory Security for GPUs	543
<i>Rahaf Abdullah, Huiyang Zhou, Amro Awad</i>	

MPress: Democratizing Billion-Scale Model Training on Multi-GPU Servers Via Memory-Saving Inter-Operator Parallelism	556
<i>Quan Zhou, Haiquan Wang, Xiaoyan Yu, Cheng Li, Youhui Bai, Feng Yan, Yinlong Xu</i>	

SESSION 4A: NEURAL NETWORKS AND ACCELERATORS 2

DeFiNES: Enabling Fast Exploration of the Depth-First Scheduling Space for DNN Accelerators Through Analytical Modeling	570
<i>Linyan Mei, Koen Goetschalckx, Arne Symons, Marian Verhelst</i>	
CEGMA: Coordinated Elastic Graph Matching Acceleration for Graph Matching Networks.....	584
<i>Yue Dai, Youtao Zhang, Xulong Tang</i>	
ISOSceles: Accelerating Sparse CNNs Through Inter-Layer Pipelining.....	598
<i>Yifan Yang, Joel S. Emer, Daniel Sanchez</i>	
OptimStore: In-Storage Optimization of Large Scale DNNs with On-Die Processing	611
<i>Junkyum Kim, Myeonggu Kang, Yunki Han, Yang-Gon Kim, Lee-Sup Kim</i>	
KRISP: Enabling Kernel-Wise RIGht-Sizing for Spatial Partitioned GPU Inference Servers.....	624
<i>Marcus Chow, Ali Jahanshahi, Daniel Wong</i>	
MERCURY: Accelerating DNN Training by Exploiting Input Similarity.....	638
<i>Vahid Janfaza, Kevin Weston, Moein Razavi, Shantanu Mandal, Farabi Mahmud, Alex Hilty, Abdullah Muzahid</i>	

SESSION 4B: PIMS AND PERSISTENT MEMORY

Silo: Speculative Hardware Logging for Atomic Durability in Persistent Memory	651
<i>Ming Zhang, Yu Hua</i>	
Reconciling Selective Logging and Hardware Persistent Memory Transaction.....	664
<i>Chen Cheng Ye, Yuan Chao Xu, Xipeng Shen, Yan Sha, Xiaofei Liao, Hai Jin, Yan Solihin</i>	
SecPB: Architectures for Secure Non-Volatile Memory with Battery-Backed Persist Buffers	677
<i>Alexander Freij, Huiyang Zhou, Yan Solihin</i>	
EVE: Ephemeral Vector Engines.....	691
<i>Khalid Al-Hawaj, Tuan Ta, Nick Cebry, Shady Agwa, Olalekan Afuye, Eric Hall, Courtney Golden, Alyssa B. Apsel, Christopher Batten</i>	
On Consistency for Bulk-Bitwise Processing-In-Memory	705
<i>Ben Perach, Ronny Ronen, Shahar Kvatinsky</i>	
Dalorex: A Data-Local Program Execution and Architecture for Memory-Bound Applications	718
<i>Marcelo Orenes-Vera, Esin Tureci, David Wentzlaff, Margaret Martonosi</i>	

SESSION 4C: QUANTUM AND FPGAS

HyQSAT: A Hybrid Approach for 3-SAT Problems by Integrating Quantum Annealer with CDCL	731
<i>Siwei Tan, Mingqian Yu, Andre Python, Yongheng Shang, Tingting Li, Liqiang Lu, Jianwei Yin</i>	
Duet: Creating Harmony Between Processors and Embedded FPGAs	745
<i>Ang Li, August Ning, David Wentzlaff</i>	

Co-Designed Architectures for Modular Superconducting Quantum Computers	759
<i>Evan McKinney, Mingkang Xia, Chao Zhou, Pinlei Lu, Michael Hatridge, Alex K. Jones</i>	
A Pulse Generation Framework with Augmented Program-Aware Basis Gates and Criticality Analysis	773
<i>Yanhao Chen, Yuwei Jin, Fei Hua, Ari Hayes, Ang Li, Yunong Shi, Eddy Z. Zhang</i>	
The Imitation Game: Leveraging CopyCats for Robust Native Gate Selection in NISQ Programs.....	787
<i>Poulami Das, Eric Kessler, Yunong Shi</i>	

SESSION 5A: CLOUD AND EDGE COMPUTING

ENODE: Energy-Efficient and Low-Latency Edge Inference and Training of Neural ODEs.....	802
<i>Junkang Zhu, Yaoyu Tao, Zhengya Zhang</i>	
SpecFaaS: Accelerating Serverless Applications with Speculative Function Execution.....	814
<i>Jovan Stojkovic, Tianyin Xu, Hubertus Franke, Josep Torrellas</i>	
MoCA: Memory-Centric, Adaptive Execution for Multi-Tenant Deep Neural Networks.....	828
<i>Seah Kim, Hasan Genc, Vadim Vadimovich Nikiforov, Krste Asanovic, Borivoje Nikolic, Yakun Sophia Shao</i>	
Know Your Enemy to Save Cloud Energy: Energy-Performance Characterization of Machine Learning Serving	842
<i>Junyeol Yu, Jongseok Kim, Euseong Seo</i>	
Adrias: Interference-Aware Memory Orchestration for Disaggregated Cloud Infrastructures.....	855
<i>Dimosthenis Masouros, Christian Pinto, Michele Gazzetti, Sotirios Xydis, Dimitrios Soudris</i>	

SESSION 5B: ENCRYPTION AND SGX

Poseidon: Practical Homomorphic Encryption Accelerator	870
<i>Yinghao Yang, Huaizhi Zhang, Shengyu Fan, Hang Lu, Mingzhe Zhang, Xiaowei Li</i>	
FAB: An FPGA-Based Accelerator for Bootstrappable Fully Homomorphic Encryption	882
<i>Rashmi Agrawal, Leo De Castro, Guowei Yang, Chiraag Juvekar, Rabia Yazicigil, Anantha Chandrakasan, Vinod Vaikuntanathan, Ajay Joshi</i>	
FxHENN: FPGA-Based Acceleration Framework for Homomorphic Encrypted CNN Inference.....	896
<i>Yilan Zhu, Xinyao Wang, Lei Ju, Shanqing Guo</i>	
D-Shield: Enabling Processor-Side Encryption and Integrity Verification for Secure NVMe Drives.....	908
<i>Md Hafizul Islam Chowdhury, Myoungsoo Jung, Fan Yao, Amro Awad</i>	
TensorFHE: Achieving Practical Computation on Encrypted Data Using GPGPU	922
<i>Shengyu Fan, Zhiwei Wang, Weizhi Xu, Rui Hou, Dan Meng, Mingzhe Zhang</i>	

SESSION 5C: RELIABILITY

AVGI: Microarchitecture-Driven, Fast and Accurate Vulnerability Assessment.....	935
<i>George Papadimitriou, Dimitris Gizopoulos</i>	
Thales: Formulating and Estimating Architectural Vulnerability Factors for DNN Accelerators	949
<i>Abhishek Tyagi, Yiming Gan, Shaoshan Liu, Bo Yu, Paul Whatmough, Yuhao Zhu</i>	

Realizing Extreme Endurance Through Fault-Aware Wear Leveling and Improved Tolerance 964
Jiangwei Zhang, Chong Wang, Zhenhua Zhu, Donald Kline, Alex K. Jones, Huazhong Yang, Yu Wang

ESD: An ECC-Assisted and Selective Deduplication for Encrypted Non-Volatile Main Memory 977
Chunfeng Du, Suzhen Wu, Jiapeng Wu, Bo Mao, Shengzhe Wang

SESSION 6A: INDUSTRY TRACK SESSION

A Systematic Study of DDR4 DRAM Faults in the Field 991
Majed Valad Beigi, Yi Cao, Sudhanva Gurumurthi, Charles Recchia, Andrew Walton, Vilas Sridharan

High Performance and Power Efficient Accelerator for Cloud Inference 1003
Jianguo Yao, Hao Zhou, Yalin Zhang, Ying Li, Chuang Feng, Shi Chen, Jiaoyan Chen, Yongdong Wang, Qiaojuan Hu

LightTrader: A Standalone High-Frequency Trading System with Deep Learning Inference Accelerators and Proactive Scheduler 1017
Sungyeob Yoo, Hyunsung Kim, Jinseok Kim, Sunghyun Park, Joo-Young Kim, Jinwook Oh

BM-Store: A Transparent and High-Performance Local Storage Architecture for Bare-Metal Clouds Enabling Large-Scale Deployment..... 1031
Yiquan Chen, Jiexiong Xu, Chengkun Wei, Yijing Wang, Xin Yuan, Yangming Zhang, Xulin Yu, Yi Chen, Zeke Wang, Shuibing He, Wenzhi Chen

SESSION 6B: NICS AND NETWORKS

Turbo: SmartNIC-Enabled Dynamic Load Balancing of μ s-Scale RPCs..... 1045
Hamed Seyedroudbari, Srikar Vanavasam, Alexandros Daglis

A Scalable Methodology for Designing Efficient Interconnection Network of Chiplets 1059
Yinxiao Feng, Dong Xiang, Kaisheng Ma

VVQ: Virtualizing Virtual Channel for Cost-Efficient Protocol Deadlock Avoidance..... 1072
Hans Kasan, John Kim

SESSION 7A: NEURAL NETWORK AND ACCELERATORS 3

Mix-GEMM: An Efficient HW-SW Architecture for Mixed-Precision Quantized Deep Neural Networks Inference on Edge Devices..... 1085
Enrico Reggiani, Alessandro Pappalardo, Max Doblas, Miquel Moreto, Mauro Olivieri, Osman Sabri Unsal, Adrián Cristal

FlowGNN: A Dataflow Architecture for Real-Time Workload-Agnostic Graph Neural Network Inference 1099
Rishov Sarkar, Stefan Abi-Karam, Yuqi He, Lakshmi Sathidevi, Cong Hao

Chimera: An Analytical Optimizing Framework for Effective Compute-Intensive Operators Fusion..... 1113
Size Zheng, Siyuan Chen, Peidi Song, Renze Chen, Xiuhong Li, Shengen Yan, Dahua Lin, Jingwen Leng, Yun Liang

Securator: A Fast and Secure Neural Processing Unit 1127
Nivedita Shrivastava, Smruti Ranjan Sarangi

Tensor Movement Orchestration in Multi-GPU Training Systems	1140
<i>Shao-Fu Lin, Yi-Jung Chen, Hsiang-Yun Cheng, Chia-Lin Yang</i>	

SESSION 7B: MICROARCHITECTURE AND MEMORY SYSTEMS

A Storage-Effective BTB Organization for Servers	1153
<i>Truls Asheim, Boris Grot, Rakesh Kumar</i>	
HoPP: Hardware-Software Co-Designed Page Prefetching for Disaggregated Memory	1168
<i>Hai Feng Li, Ke Liu, Ting Liang, Zuojun Li, Tianyue Lu, Hui Yuan, Yinben Xia, Yungang Bao, Mingyu Chen, Yizhou Shan</i>	
Speculative Register Reclamation	1182
<i>Sanyam Mehta</i>	
SnakeByte: A TLB Design with Adaptive and Recursive Page Merging in GPUs.....	1195
<i>Jiwon Lee, Ju Min Lee, Yunho Oh, William J. Song, Won Woo Ro</i>	
CARE: A Concurrency-Aware Enhanced Lightweight Cache Management Framework.....	1208
<i>Xiaoyang Lu, Rujia Wang, Xian-He Sun</i>	
Memory-Efficient Hashed Page Tables	1221
<i>Jovan Stojkovic, Namrata Mantri, Dimitrios Skarlatos, Tianyin Xu, Josep Torrellas</i>	

SESSION 7C: APPLICATIONS 2 & POTPOURRI

NvWa: Enhancing Sequence Alignment Accelerator Throughput Via Hardware Scheduling	1236
<i>Yewen Li, Xueqi Li, Ruihao Gao, Wanqi Liu, Guangming Tan</i>	
Efficient Supernet Training Using Path Parallelism	1249
<i>Ying Xu, Long Cheng, Xuyi Cai, Xiaohan Ma, Weiwei Chen, Lei Zhang, Ying Wang</i>	
Phloem: Automatic Acceleration of Irregular Applications with Fine-Grain Pipeline Parallelism	1262
<i>Quan M. Nguyen, Daniel Sanchez</i>	
CHOPPER: A Compiler Infrastructure for Programmable Bit-Serial SIMD Processing Using Memory in DRAM	1275
<i>Xiangjun Peng, Yaohua Wang, Ming-Chang Yang</i>	
VAQUERO: A Scratchpad-Based Vector Accelerator for Query Processing	1289
<i>Julián Pavón, Ivan Vargas Valdivieso, Joan Marimon, Roger Figueras, Francesc Moll, Osman Unsal, Mateo Valero, Adrian Cristal</i>	

Author Index