

2025 IEEE International Symposium on Workload Characterization (IISWC 2025)

**Irvine, California, USA
12-14 October 2025**

IEEE Catalog Number:
ISBN:

CFP25236-POD
979-8-3315-4918-3



Copyright © 2025, 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:	CFP25236-POD
ISBN (Print-On-Demand):	979-8-3315-4918-3
ISBN (Online):	979-8-3315-4917-6

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

Belenos: Bottleneck Evaluation to Link Biomechanics to Novel Computing Optimizations	1
<i>Hana Chitsaz, Johnson Umeike, Amirmahdi Namjoo, Babak N. Safa, Bahar Asgari</i>	
Workload Characterization Using Cross-Layer Features and Multilevel PCA	16
<i>Lina Sawalha, Grant Deljevic</i>	
Athena: A Plug-and-Play Advisor for Retrieval-Augmented Generation using VectorDB	28
<i>Ning Liang, Fabian Wenz, Jana Giceva, Lisa Wu Wills</i>	
The Fake-Busy and True-Idle Problems of Running Graph Applications on Chiplet-Based Multi-Cores	42
<i>Rashid Aligholipour, Yuan Yao</i>	
WANify: Gauging and Balancing Runtime WAN Bandwidth for Geo-distributed Data Analytics	56
<i>Anshuman Das Mohapatra, Kwangsung Oh</i>	
Understanding Distributed Training of Large Language Models with Unified Virtual Memory	70
<i>Jane Rhee, Eunbi Jeong, Jiwon Lee, Myung Kuk Yoon</i>	
Confidential LLM Inference: Performance and Cost Across CPU and GPU TEEs	84
<i>Marcin Chrapek, Marcin Copik, Etienne Mettaz, Torsten Hoefler</i>	
EdgeReasoning: Characterizing Reasoning LLM Deployment on Edge GPUs	99
<i>Benjamin Kubwimana, Qijing Huang</i>	
Keeping up with Large Language Models: A Holistic Methodology of Compute, Memory, Communication, and Cost Modeling	116
<i>Wenzhe Guo, Joyjit Kundu, Uras Tos, Weijiang Kong, Giuliano Sisto, Timon Evenblij, Manu Perumkunnil</i>	
DABench-LLM: Standardized and In-Depth Benchmarking of Post-Moore Dataflow AI Accelerators for LLMs	127
<i>Ziyu Hu, Zhiqing Zhong, Weijian Zheng, Zhijing Ye, Xuwei Tan, Xueru Zhang, Zheng Xie Rajkumar Kettimuthu, Xiaodong Yu</i>	
ZKProphet: Understanding Performance of Zero-Knowledge Proofs on GPUs	142
<i>Tarunesh Verma, Yichao Yuan, Nishil Talati, Todd Austin</i>	
The Curious Case of Global Stable Loads	155
<i>Shagnik Pal, Jeeho Ryoo, Lizy K. John</i>	
vACE: Exploring the Design Space of Vector Processing Units for Soft Error Vulnerability	167
<i>George-Marios Fragkoulis, Dimitris Gizopoulos</i>	
CASM: A Generalizable and Accessible Security Metric to Evaluate Security of Cache Architectures	179
<i>Phaedra Sophia Curlin, Tamara Silbergleit Lehman</i>	
Learning Architectural Cache Simulator Behaviour	193
<i>Pranjali Jain, Meiru Han, Zhizhou Zhang, Brandon Lee, Jonathan Balkind</i>	
HALO: Hybrid Systolic Arrays via Logical Partitioning for Acceleration of Complex-Valued Neural Networks	206
<i>Ji Yeong Yi, Eunbi Jeong, SungHee Yum, Jane Rhee, Sangun Choi, Gunjae Koo, Yunho Oh Myung Kuk Yoon</i>	

Exploring Lossy Compression of Activation Data for Emerging AI Accelerators: A Case Study on the Graphcore IPU	219
<i>Milan Shah, Xiaodong Yu, Sheng Di, Michela Becchi, Franck Cappello</i>	
BetterTogether: An Interference-Aware Framework for Fine-grained Software Pipelining on Heterogeneous SoCs	233
<i>Yanwen Xu, Rithik Sharma, Zheyuan Chen, Shaan Mistry, Tyler Sorensen</i>	
PRISM: Processing-In-Memory Sparse MTTKRP for Tensor Decomposition Acceleration	246
<i>Daniel Pacheco, Leonel Sousa, Aleksandar Ilic</i>	
ALPHA-PIM: Analysis of Linear Algebraic Processing for High-Performance Graph Applications on a Real Processing-In-Memory System	257
<i>Marzieh Barkhordar, Alireza Tabatabaeian, Mohammad Sadrosadati, Christina Giannoula, Juan Gomez Luna, Izzat El Hajj, Onur Mutlu, Alaa R. Alameldeen</i>	
PangenomicsBench: A Benchmark Suite and Characterization of Pangenomics	272
<i>Noah Kaplan, Jan-Niklas Schmelzle, Yufeng Gu, Erik Garrison, Christopher Batten, Reetuparna Das</i>	
decoder-bench: Benchmarking Decoders for Quantum Error Correction	286
<i>Satvik Maurya, Joshua Vizlai, Nithin Raveendran, Poulami Das, Swamit Tannu</i>	
A Comprehensive Analysis of Graph Neural Networks Training at Different Scales	296
<i>Mostafa Eghbali Zarch, Michela Becchi</i>	
EntoBench: A Benchmark Suite and Evaluation Framework for Insect-Scale Robotics	309
<i>Derin Ozturk, Nick Cebry, Angela Cui, Hang Gao, Julie Villamil, E. Farrell Helbling, Christopher Batten</i>	
Improving the Performance of Out-of-Core LLM Inference Using Heterogeneous Host Memory	325
<i>Sudhanshu Gupta, Sandhya Dwarkadas</i>	
miniGiraffe: A Pangenomic Mapping Proxy App	339
<i>Jessica I. Dagostini, Joseph B. Manzano, Tyler Sorensen, Scott Beamer</i>	
Does Linux Provide Performance Isolation for NVMe SSDs? Configuring cgroups for I/O Control in the NVMe Era	353
<i>Krijn Doekemeijer, Zebin Ren, Tiziano De Matteis, Balakrishnan Chandrasekaran, Animesh Trivedi</i>	
Dissecting CPU-GPU Unified Physical Memory on AMD MI300A APUs	368
<i>Jacob Wahlgren, Gabin Schieffer, Ruimin Shi, Edgar A. León, Roger Pearce, Maya Gokhale, Ivy Peng</i>	
Design and accuracy trade-offs in Computational Statistics	381
<i>Tiancheng Xu, Alan L. Cox, Scott Rixner</i>	
An Analysis of Ethereum Workloads from a Key-Value Storage Perspective	394
<i>Yanjing Ren, Jia Zhao, Jingwei Li, Patrick P. C. Lee</i>	
Storage-Based Approximate Nearest Neighbor Search: What are the Performance, Cost, and I/O Characteristics?	407
<i>Zebin Ren, Krijn Doekemeijer, Padma Apparao, Animesh Trivedi</i>	
Sweet or Sour CHERI: Performance Characterization of the Arm Morello Platform	423
<i>Xiaoyang Sun, Jeremy Singer, Zheng Wang</i>	

Characterizing Adaptive Mesh Refinement on Heterogeneous Platforms with Parthenon-VIBE	439
<i>Akash Poptani, Alireza Khadem, Scott Mahlke, Jonah Miller, Joshua Dolence, Galen Shipman</i>	
<i>Reetuparna Das</i>	
XRSight: An End-to-End Hardware-Software Co-Design Platform for XR SoC Evaluation	452
<i>Prashanth Ganesh, Zekai Lin, Yakun Sophia Shao</i>	
Icicle: Open-Source Hardware Support for Top-Down Microarchitectural Analysis on RISC-V	464
<i>Matthew Edwin Weingarten, Michael Grieco, Stephen Edwards, Tanvir Ahmed Khan</i>	
AlphaFold3 Workload Characterization: A Comprehensive Analysis of Bottlenecks and Performance	
Scaling	478
<i>Jinpyo Kim, Mingi Kwon, Jishen Zhao</i>	
Characterizing and Optimizing Real-Time Optimal Control for Embedded SoCs	489
<i>Shengjun Kris Dong, Dima Nikiforov, Widyadewi Soedarmadji, Minh Nguyen, Vikram Jain</i>	
<i>Christopher W. Fletcher, Yakun Sophia Shao</i>	
ClusterSim: Modeling Thread Block Clusters in Hopper GPUs	504
<i>Tim Lühnen, Jyotirman Behera, Devashree Tripathy, Sohan Lal</i>	
Adaptive Graphical Settings Optimization for Energy-Efficient Mobile 3D Rendering	516
<i>Nolin Szafranski, Hong Bing Tang, Lina Sawalha</i>	
Benchmarking Support for RISC-V CPUs in Serverless Computing	520
<i>Georgios Pournaras, Vasileios Karakostas, George Papadimitriou, Dimitris Gizopoulos</i>	
CoBloom: An FPGA Accelerator System for Bloom Filter Insertion in Genomics Applications	524
<i>Patrick Hardison, Chris Kjellqvist, Ning Liang, Lisa Wu Wills</i>	
PartnerM ² LB: Personal Assistant Multi-device Machine Learning Benchmark	529
<i>Yu-Ching Hu, Owen Lam, Yulian Li, Hung-Wei Tseng</i>	