

2024 IEEE International Symposium on Workload Characterization (IISWC 2024)

**Vancouver, British Columbia, Canada
15-17 September 2024**



**IEEE Catalog Number: CFP24236-POD
ISBN: 979-8-3503-5604-5**

**Copyright © 2024 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:	CFP24236-POD
ISBN (Print-On-Demand):	979-8-3503-5604-5
ISBN (Online):	979-8-3503-5603-8
ISSN:	2835-222X

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

2024 IEEE International Symposium on Workload Characterization (IISWC) IISWC 2024

Table of Contents

Message from General Chairs	ix
Message from Program Chairs	x
Organizing Committee	xi
Program Committee	xii
Steering Committee	xiii
Artifact Evaluation Committee	xiv

2024 IEEE International Symposium on Workload Characterization

CRISP: Concurrent Rendering and Compute Simulation Platform for GPUs	1
<i>Junrui Pan (Purdue University, USA) and Timothy G. Rogers (Purdue University, USA)</i>	
LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving at Scale	15
<i>Jaehong Cho (KAIST, South Korea), Minsu Kim (KAIST, South Korea), Hyunmin Choi (KAIST, South Korea), Guseul Heo (KAIST, South Korea), and Jongse Park (KAIST, South Korea)</i>	
Lotus: Characterization of Machine Learning Preprocessing Pipelines via Framework and Hardware Profiling	30
<i>Rajveer Bachkaniwala (Georgia Tech, USA), Harshith Lanka (Georgia Tech, USA), Kexin Rong (Georgia Tech, USA), and Ada Gavrilovska (Georgia Tech, USA)</i>	
Mediator: Characterizing and Optimizing Multi-DNN Inference for Energy Efficient Edge Intelligence	44
<i>Seung Hun Choi (Korea University, South Korea), Myung Jae Chung (Korea University, South Korea), Young Geun Kim (Korea University, South Korea), and Sung Woo Chung (Korea University, South Korea)</i>	
Performance Modeling and Workload Analysis of Distributed Large Language Model Training and Inference	57
<i>Joyjit Kundu (IMEC, Belgium), Wenzhe Guo (IMEC, Belgium), Ali BanaGozar (IMEC, Belgium), Udari De Alwis (IMEC, Belgium), Sourav Sengupta (IMEC, Belgium), Puneet Gupta (UCLA, US), and Arindam Mallik (IMEC, Belgium)</i>	

CARM Tool: Cache-Aware Roofline Model Automatic Benchmarking and Application Analysis	68
<i>José Morgado (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal), Leonel Sousa (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal), and Aleksandar Ilic (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal)</i>	
SHARP: A Distribution-Based Framework for Reproducible Performance Evaluation	82
<i>Viyom Mittal (Hewlett Packard Labs, University of California, USA), Pedro Bruel (Hewlett Packard Labs, USA), Michalis Faloutsos (University of California, USA), Dejan Milojicic (Hewlett Packard Labs, USA), and Eitan Frachtenberg (Hewlett Packard Labs, USA)</i>	
Taming Performance Variability Caused by Client-Side Hardware Configuration	94
<i>Georgia Antoniou (University of Cyprus, The Republic of Cyprus), Haris Volos (University of Cyprus, The Republic of Cyprus), and Yiannakis Sazeides (University of Cyprus, The Republic of Cyprus)</i>	
HEX-SIM:Evaluating Multi-Modal Large Language Models on Multi-Chiplet NPUs	108
<i>Xinquan Lin (Fuzhou University, China), Haobo Xu (ICT, Chinese Academy of Sciences, China), Yinhe Han (ICT, Chinese Academy of Sciences, China), and Yiming Gan (ICT, Chinese Academy of Sciences, China)</i>	
Empowering the Quantum Cloud User with QRIO	121
<i>Shmeelok Chakraborty (University of Michigan, USA), Yuewen Hou (University of Michigan, USA), Ang Chen (University of Michigan, USA), and Gokul Subramanian Ravi (University of Michigan, USA)</i>	
Evergreen: Comprehensive Carbon Model for Performance-Emission Tradeoffs	132
<i>Tersiteab Adem (University of Michigan, USA), Andrew McCrabb (University of Michigan, USA), Vidushi Goyal (University of Michigan, USA), and Valeria Bertacco (University of Michigan, USA)</i>	
Performance Analysis of Zero-Knowledge Proofs	144
<i>Saichand Samudrala (Texas A&M University, USA), Jiawen Wu (Texas A&M University, USA), Chen Chen (Texas A&M University, USA), Haoxuan Shan (Duke University, USA), Jonathan Ku (Duke University, USA), Yiran Chen (Duke University, USA), and Jeyavijayan Rajendran (Texas A&M University, USA)</i>	
VelociTI: An Architecture-Level Performance Modeling Framework for Trapped Ion Quantum Computers	156
<i>Alexander Hankin (Harvard University, USA), Abdulrahman Mahmoud (Harvard University, USA), Mark Hempstead (Tufts University, USA), David Brooks (Harvard University, USA), and Gu-Yeon Wei (Harvard University, USA)</i>	
Understanding Performance Implications of LLM Inference on CPUs	169
<i>Seonjin Na (Georgia Institute of Technology, USA), Geonhwa Jeong (Georgia Institute of Technology, USA), Byung Hoon Ahn (University of California, USA), Jeffrey Young (Georgia Institute of Technology, USA), Tushar Krishna (Georgia Institute of Technology, USA), and Hyesoon Kim (Georgia Institute of Technology, USA)</i>	
Low-Bitwidth Floating Point Quantization for Efficient High-Quality Diffusion Models	181
<i>Cheng Chen (University of Toronto, Canada), Christina Giannoula (University of Toronto & Vector Institute, Canada), and Andreas Moshovos (University of Toronto & Vector Institute, Canada)</i>	

Characterizing the Accuracy-Efficiency Trade-off of Low-Rank Decomposition in Language Models	194
<i>Chakshu Moar (University of California, USA), Faraz Tahmasebi (University of California, USA), Michael Pellauer (NVIDIA, USA), and Hyoukjun Kwon (University of California, USA)</i>	
Understanding the Performance and Estimating the Cost of LLM Fine-Tuning	210
<i>Yuchen Xia (University of Michigan, USA), Jiho Kim (Georgia Institute of Technology, USA), Yuhan Chen (University of Michigan, USA), Haojie Ye (University of Michigan, USA), Souvik Kundu (Intel Labs, USA), Cong Hao (Georgia Institute of Technology, USA), and Nishil Talati (University of Michigan, USA)</i>	
Characterizing and Optimizing the End-to-End Performance of Multi-Agent Reinforcement Learning Systems	224
<i>Kailash Gogineni (George Washington University, USA), Yongsheng Mei (George Washington University, USA), Karthikeya Gogineni (Independent), Peng Wei (George Washington University, USA), Tian Lan (George Washington University, USA), and Guru Venkataramani (George Washington University, USA)</i>	
Understanding Address Translation Scaling Behaviours Using Hardware Performance Counters ..	236
<i>Nick Lindsay (Yale University, USA) and Abhishek Bhattacharjee (Yale University, USA)</i>	
Architectural Modeling and Benchmarking for Digital DRAM PIM	247
<i>Farzana Ahmed Siddique (University of Virginia, USA), Deyuan Guo (University of Virginia, USA), Zhenxing Fan (University of Virginia, USA), Mohammadhosein Gholamrezaei (University of Virginia, USA), Morteza Baradaran (University of Virginia, USA), Alif Ahmed (University of Virginia, USA), Hugo Abbot (University of Virginia, USA), Kyle Durrer (University of Virginia, USA), Kumaresh Nandagopal (University of Virginia, USA), Ethan Ermovick (University of Virginia, USA), Khyati Kiyawat (University of Virginia, USA), Beenish Gul (University of Virginia, USA), Abdullah Mughrabi (University of Virginia, USA), Ashish Venkat (University of Virginia, USA), and Kevin Skadron (University of Virginia, USA)</i>	
Kindle: A Comprehensive Framework for Exploring OS-Architecture Interplay in Hybrid Memory Systems	262
<i>Arun Kp (Indian Institute of Technology Kanpur, India) and Debadatta Mishra (Indian Institute of Technology Kanpur, India)</i>	
Enhanced System-Level Coherence for Heterogeneous Unified Memory Architectures	273
<i>Anoop Mysore Nataraja (University of Washington, USA), Ricardo Fernández-Pascual (University of Murcia, Spain), and Alberto Ros (University of Murcia, Spain)</i>	
Characterizing Emerging Page Replacement Policies for Memory-Intensive Applications	284
<i>Michael Wu (Yale University, USA), Sibren Isaacman (Loyola University Maryland, USA), and Abhishek Bhattacharjee (Yale University, USA)</i>	
Characterizing CUDA and OpenMP Synchronization Primitives	295
<i>Brandon Alexander Burtchell (Texas State University, USA) and Martin Burtcher (Texas State University, USA)</i>	

Evaluating Performance and Energy Efficiency of Parallel Programming Models in Heterogeneous Computing Systems	309
<i>Demirhan Sevim (Ozyegin University, Turkey), Baturalp Bilgin (Ozyegin University, Turkey), and Ismail Aktürk (Ozyegin University, Turkey)</i>	
Performance Impact of Removing Data Races from GPU Graph Analytics Programs	320
<i>Yiqian Liu (Texas State University, USA), Avery VanAusdal (Texas State University, USA), and Martin Burtscher (Texas State University, USA)</i>	
Author Index	333