

2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2024)

**5-7 May 2024
Indianapolis, Indiana, USA**



**IEEE Catalog Number: CFP24PER-POD
ISBN: 979-8-3503-7639-5**

**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:	CFP24PER-POD
ISBN (Print-On-Demand):	979-8-3503-7639-5
ISBN (Online):	979-8-3503-7638-8
ISSN:	2994-9513

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 Performance Analysis of Systems and Software (ISPASS) **ISPASS 2024**

Table of Contents

Message from the General Chairs	xi
Message from the Program Chairs	xii
Organizing Committee	xiv
Program Committee	xv
Steering Committee	xvii
Sponsors	xviii

Best Papers

Aiding Microprocessor Performance Validation with Machine Learning	1
<i>Erick Carvajal Barboza (Universidad de Costa Rica, Costa Rica), Mahesh Ketkar (Intel Corporation, USA), Paul Gratz (Texas A&M University, USA), and Jiang Hu (Texas A&M University, USA)</i>	
CiMLoop: A Flexible, Accurate, and Fast Compute-In-Memory Modeling Tool	10
<i>Tanner Andrulis (Massachusetts Institute of Technology, USA), Joel S. Emer (Massachusetts Institute of Technology, Nvidia, USA), and Vivienne Sze (Massachusetts Institute of Technology, USA)</i>	
Characterizing In-Kernel Observability of Latency-Sensitive Request-Level Metrics with eBPF	24
<i>Mohammadreza Rezoani (University of California Riverside, USA), Ali Jahanshahi (University of California Riverside, USA), and Daniel Wong (University of California Riverside, USA)</i>	

BTBench: A Benchmark for Comprehensive Binary Translation Performance Evaluation 36
Xinyu Li (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, China), Yanzhi Lan (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, China), Gen Niu (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, China), Feng Xue (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, China), and Fuxin Zhang (State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, China)

Performance Modeling & Analysis

MuchiSim: A Simulation Framework for Design Exploration of Multi-Chip Manycore Systems 48
Marcelo Orenes-Vera (Princeton University), Esin Tureci (Princeton University), Margaret Martonosi (Princeton University), and David Wentzlauff (Princeton University)

CiFlow: Dataflow Analysis and Optimization of Key Switching for Homomorphic Encryption 61
Negar Neda (New York University, USA), Austin Ebel (New York University, USA), Benedict Reynwar (Information Sciences Institute, University of Southern California, USA), and Brandon Reagen (New York University, USA)

Workload Characterization of Commercial Mobile Benchmark Suites 73
Victor Kariofillis (University of Toronto) and Natalie Enright Jerger (University of Toronto)

RTune: Towards Automated and Coordinated Optimization of Computing and Computational Objectives of Parallel Iterative Applications 85
Yonghong Yan (University of North Carolina at Charlotte, USA), Kewei Yan (University of North Carolina at Charlotte, USA), and Anjia Wang (University of North Carolina at Charlotte, USA)

Analysis of HW Systems

Characterizing Soft-Error Resiliency in Arm's Ethos-U55 Embedded Machine Learning Accelerator 96
Abhishek Tyagi (University of Rochester, USA), Reiley Jeyapaul (AMD, USA), Chuteng Zhou (ARM Inc, USA), Paul Whatmough (AI Research Qualcomm, USA), and Yuhao Zhu (University of Rochester, USA)

SAP: Silicon Authentication Platform for System-on-Chip Supply Chain Vulnerabilities 109
Sami Ul Islam Sami (University of Florida, USA), Jingbo Zhou (University of Florida, USA), Sujan Kumar Saha (University of Florida, USA), Fahim Rahman (University of Florida, USA), Farimah Farahmandi (University of Florida, USA), and Mark Tehranipoor (University of Florida, USA)

SimPoint-Based Microarchitectural Hotspot & Energy-Efficiency Analysis of RISC-V OoO CPUs ... 120
Odysseas Chatzopoulos (University of Athens, Greece), Maria Trakosa (University of Athens, Greece), George Papadimitriou (University of Athens, Greece), Wing Shek Wong (Intel, Austin, Texas), and Dimitris Gizopoulos (University of Athens, Greece)

On the Rise of AMD Matrix Cores: Performance, Power Efficiency, and Programmability 132
Gabin Schieffer (KTH Royal Institute of Technology, Sweden), Daniel Araújo de Medeiros (KTH Royal Institute of Technology, Sweden), Jennifer Faj (KTH Royal Institute of Technology, Sweden), Aniruddha Marathe (Lawrence Livermore National Laboratory, USA), and Ioy Peng (KTH Royal Institute of Technology, Sweden)

Simulation

DNA Storage Toolkit: A Modular End-to-End DNA Data Storage Codec and Simulator 144
Puru Sharma (National University of Singapore, Singapore), Gary Goh Yipeng (National University of Singapore, Singapore), Bin Gao (National University of Singapore, Singapore), Longshen Ou (National University of Singapore, Singapore), Dehui Lin (National University of Singapore, Singapore), Deepak Sharma (National University of Singapore, Singapore), and Djordje Jevdjic (National University of Singapore, Singapore)

Zatel: Sample Complexity–Aware Scale–Model Simulation for Ray Tracing 156
Davit Grigoryan (University of British Columbia, Canada), Yuan Hsi Chou (University of British Columbia, Canada), and Tor M. Aamodt (University of British Columbia, Canada)

BZSim: Fast, Large-Scale Microarchitectural Simulation with Detailed Interconnect Modeling 167
Panagiotis Strikos (Chalmers University of Technology, Sweden), Ahsen Ejaz (Chalmers University of Technology, Sweden), and Ioannis Sourdis (Chalmers University of Technology, Sweden)

Userspace Networking in gem5 179
Johnson Umeike (University of Kansas), Siddharth Agarwal (University of Illinois Urbana-Champaign), Nikita Lazarev (MIT, CSAIL), and Mohammad Alian (University of Kansas)

System-Level Optimization

Vision Transformer Computation and Resilience for Dynamic Inference 192
Kavya Sreedhar (Stanford University, USA), Jason Clemons (NVIDIA, USA), Rangharajan Venkatesan (NVIDIA, USA), Stephen W. Keckler (NVIDIA, USA), and Mark Horowitz (Stanford University, USA)

LIBRA: Enabling Workload-Aware Multi-Dimensional Network Topology Optimization for Distributed Training of Large AI Models 205
William Won (Georgia Institute of Technology, USA), Saeed Rashidi (Georgia Institute of Technology, USA), Sudarshan Srinivasan (Intel, India), and Tushar Krishna (Georgia Institute of Technology, USA)

SwiftRL: Towards Efficient Reinforcement Learning on Real Processing-In-Memory Systems	217
<i>Kailash Gogineni (George Washington University, USA), Sai Santosh Dayapule (George Washington University, USA), Juan Gómez-Luna (ETH Zürich, Switzerland), Karthikeya Gogineni (Independent), Peng Wei (George Washington University, USA), Tian Lan (George Washington University, USA), Mohammad Sadrosadati (ETH Zürich, Switzerland), Onur Mutlu (ETH Zürich, Switzerland), and Guru Venkataramani (George Washington University, USA)</i>	
Forward to the Past: An Alternative to Hybrid CPU Design	230
<i>Sanyam Mehta (HPE, USA) and Anna Yue (University of Minnesota, USA)</i>	

AI & LLM Models & Analysis

Bandwidth Characterization of DeepSpeed on Distributed Large Language Model Training	241
<i>Bagus Hanindhito (The University of Texas at Austin, USA), Bhavesh Patel (Dell Technologies, USA), and Lizy K. John (The University of Texas at Austin, USA)</i>	
Generative AI Beyond LLMs: System Implications of Multi-Modal Generation	257
<i>Alicia Golden (FAIR at Meta; Harvard University), Samuel Hsia (FAIR at Meta; Harvard University), Fei Sun (Meta), Bilge Acun (FAIR at Meta), Basil Hosmer (FAIR at Meta), Yejin Lee (FAIR at Meta), Zachary DeVito (FAIR at Meta), Jeff Johnson (FAIR at Meta), Gu-Yeon Wei (Harvard University), David Brooks (Harvard University), and Carole-Jean Wu (FAIR at Meta)</i>	
Towards Cognitive AI Systems: Workload and Characterization of Neuro-Symbolic AI	268
<i>Zishen Wan (Georgia Institute of Technology, USA), Che-Kai Liu (Georgia Institute of Technology, USA), Hanchen Yang (Georgia Institute of Technology, USA), Ritik Raj (Georgia Institute of Technology, USA), Chaojian Li (Georgia Institute of Technology, USA), Haoran You (Georgia Institute of Technology, USA), Yonggan Fu (Georgia Institute of Technology, USA), Cheng Wan (Georgia Institute of Technology, USA), Ananda Samajdar (IBM Research, USA), Yingyan Lin (Georgia Institute of Technology, USA), Tushar Krishna (Georgia Institute of Technology, USA), and Arijit Raychowdhury (Georgia Institute of Technology, USA)</i>	
Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production	280
<i>Chandra Irugalbandara (Jaseci Labs), Ashish Mahendra (Jaseci Labs), Roland Daynauth (University of Michigan), Tharuka Kasthuri Arachchige (Jaseci Labs), Jayanaka Dantanarayana (Jaseci Labs), Krisztian Flautner (University of Michigan), Lingjia Tang (University of Michigan), Yiping Kang (University of Michigan), and Jason Mars (University of Michigan)</i>	

Posters

Leveraging Memory Expansion to Accelerate Large-Scale DL Training	292
<i>Divya Kadiyala (Georgia Institute of Technology), Saeed Rashidi (Georgia Institute of Technology), Taekyung Heo (Georgia Institute of Technology), Abhimanyu Bambhaniya (Georgia Institute of Technology), Tushar Krishna (Georgia Institute of Technology), and Alexandros Daglis (Georgia Institute of Technology)</i>	
APGPM: Automated PMC-Based Power Modeling Methodology for Modern Mobile GPUs	295
<i>Pranab Dash (Purdue University), Y. Charlie Hu (Purdue University), and Abhilash Jindal (IIT Delhi)</i>	
gem5-Based Evaluation of CVA6 SoC: Insights into the Architectural Design	298
<i>Umer Shahid (University of Engineering & Technology, Pakistan), Ayesha Ahmad (University of Engineering & Technology, Pakistan), and Shanzay Wasim (University of Engineering & Technology, Pakistan)</i>	
Accel-Bench: Exploring the Potential of Programming using Hardware-Accelerated Functions	301
<i>Abenezer Wudenhe (University of California, USA), Yu-Chia Liu (University of California, USA), Chris Chen (University of California, USA), and Hung-Wei Tseng (University of California, USA)</i>	
SEFsim: A Statistically-Guided Fast DRAM Simulator	304
<i>Debpratim Adak (North Carolina State University), Hyokeun Lee (North Carolina State University), Ben Feinberg (Sandia National Laboratories), Gwendolyn Voskuilen (Sandia National Laboratories), Clayton Hughes (Sandia National Laboratories), Huiyang Zhou (North Carolina State University), and Amro Awad (North Carolina State University)</i>	
Architecture-Level Modeling of Photonic Deep Neural Network Accelerators	307
<i>Tanner Andrulis (Massachusetts Institute of Technology, USA), Gohar Irfan Chaudhry (Massachusetts Institute of Technology, USA), Vinith M. Suriyakumar (Massachusetts Institute of Technology, USA), Joel S. Emer (Massachusetts Institute of Technology, Nvidia, USA), and Vivienne Sze (Massachusetts Institute of Technology, USA)</i>	
Automatic Extraction of Network Configurations for Realistic Simulation and Validation	310
<i>Joshua Suetterlein (Pacific Northwest National Laboratory, USA), Stephen J. Young (Pacific Northwest National Laboratory, USA), Jesun Firoz (Pacific Northwest National Laboratory, USA), Joseph Manzano (Pacific Northwest National Laboratory, USA), Ryan Friese (Pacific Northwest National Laboratory, USA), Nathan Tallent (Pacific Northwest National Laboratory, USA), Kevin Barker (Pacific Northwest National Laboratory, USA), and Timothy Stavenger (Pacific Northwest National Laboratory, USA)</i>	
MindPalace: A Framework for Studying Microarchitecture Design of Function-as-a-Service	313
<i>Kaifeng Xu (Princeton University, USA), Georgios Tziantzioulis (Princeton University, USA), and David Wentzloff (Princeton University, USA)</i>	
Infrastructure for Exploring SIMT Architecture in General-Purpose Processors	316
<i>Nikitha Kannan (BITS, Pilani), Kevin Wei (Stony Brook University), Dylan Scott (Stony Brook University), Natheesan Ratnasegar (Stony Brook University), Oğuzhan Canpolat (TOBB ETÜ), Hieu Mai (Stony Brook University), and Michael Ferdman (Stony Brook University)</i>	

Distributed Training of Neural Radiance Fields: A Performance Characterization	319
<i>Adrian Zhao (University of Toronto, Canada), Louis Zhang (University of Toronto, Canada), Sankeerth Durvasula (University of Toronto, Canada), Fan Chen (University of Toronto, Canada), Nilesch Jain (Intel Labs, United States), Selvakumar Panneer (Intel Labs, United States), and Nandita Vijaykumar (University of Toronto, Canada; Vector Institute, Canada)</i>	
Bottleneck Scenarios in use of the Conveyors Message Aggregation Library	322
<i>Shubhendra Pal Singhal (Georgia Institute of Technology, USA), Akihiro Hayashi (Georgia Institute of Technology, USA), and Vivek Sarkar (Georgia Institute of Technology, USA)</i>	
A Profiling-Based Benchmark Suite for Warehouse-Scale Computers	325
<i>Andreas Abel (Google), Yuying Li (Google), Richard O’Grady (Google), Chris Kennelly (Google), and Darryl Gove (Google)</i>	
Characterizing Dynamic Memory Behavior in WebAssembly Workloads	328
<i>Yuxin Qin (University of Glasgow, UK), Dejice Jacob (University of Glasgow, UK), and Jeremy Singer (University of Glasgow, UK)</i>	
Probing Weaknesses in GPU Reliability Assessment: A Cross-Layer Approach	331
<i>Lishan Yang (George Mason University, USA), George Papadimitriou (University of Athens, Greece), Dimitrios Sartzetakis (University of Athens, Greece), Adwait Jog (University of Virginia, USA), Evgenia Smirni (William & Mary, USA), and Dimitris Gizopoulos (University of Athens, Greece)</i>	
Author Index	335