2025 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2025)

Ghent, Belgium
11-13 May 2025

IEEE Catalog Number: CFP25PER-POD
ISBN: 979-8-3315-0295-9

****** 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:	CFP25PER-POD
ISBN (Print-On-Demand):	979-8-3315-0295-9
ISBN (Online):	979-8-3315-0294-2
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

2025 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)

ISPASS 2025

Table of Contents

Message from the General Chairs ..... **xii**

Message from the Program Chairs ..... **xiii**

Conference Organization ..... **xv**

Program Committee ..... **xvi**

Steering Committee ..... **xviii**

Sponsors ..... **xix**

Machine Learning

Understanding the Performance Horizon of the Latest ML Workloads with NonGEMM Workloads .. 1
Rachid Karami (University of California, USA), Sheng-Chun Kao (Mountain View, USA), and Hyoukjun Kwon (University of California, USA)

ADOR: A Design Exploration Framework for LLM Serving with Enhanced Latency and Throughput .. 15
Junsoo Kim (HyperAccel Inc., South Korea), Hunjong Lee (HyperAccel Inc., South Korea), Geonwoo Ko (KAIST, South Korea), Gyubin Choi (HyperAccel Inc., South Korea), Seri Ham (KAIST, South Korea), Seongmin Hong (HyperAccel Inc., South Korea), and Joo-Young Kim (HyperAccel Inc., South Korea)

Generative AI in Embodied Systems: System-Level Analysis of Performance, Efficiency and Scalability 26
Zishen Wan (Georgia Institute of Technology), Jiayi Qian (Georgia Institute of Technology), Yuhang Du (University of Minnesota, Twin Cities), Jason Jabbour (Harvard University), Yilun Du (Harvard University), Yang Katie Zhao (University of Minnesota, Twin Cities), Arijit Raychowdhury (Georgia Institute of Technology), Tushar Krishna (Georgia Institute of Technology), and Vijay Janapa Reddi (Harvard University)

Exploring Constrained Dataflow Accelerators for Real-Time Multi-Task Multi-Model ML Workloads 38
Jamin Seo (Georgia Institute of Technology, USA), Jianming Tong (Georgia Institute of Technology, USA), Tushar Krishna (Georgia Institute of Technology, USA), and Hyoukjun Kwon (University of California, USA)

Characterizing and Optimizing LLM Inference Workloads on CPU-GPU Coupled Architectures 49
Prabhu Vellaisamy (Carnegie Mellon University), Thomas Labonte (Samsung Semiconductor, Inc.), Sourav Chakraborty (Samsung Semiconductor, Inc.), Matt Turner (Hewlett Packard Enterprise), Samantika Sury (Hewlett Packard Enterprise), and John Paul Shen (Carnegie Mellon University)

GPU/Ray Tracing

Hierarchical Traversal Stack Design using Shared Memory for GPU Ray Tracing 62
Eunsoo Jung (Ewha Womans University, Korea), Eunbi Jeong (Ewha Womans University, Korea), Gunjae Koo (Korea University, Korea), Yunho Oh (Korea University, Korea), and Myung Kuk Yoon (Ewha Womans University, Korea)

RayFlex: An Open-Source RTL Implementation of the Hardware Ray Tracer Datapath 73
Fangjia Shen (Purdue University, U.S.A.), Aaron Barnes (Purdue University, U.S.A.), Anusuya Nallathambi (Purdue University, U.S.A.), and Timothy G. Rogers (Purdue University, U.S.A.)

ConCCL: Optimizing ML Concurrent Computation and Communication with GPU DMA Engines . 85
Anirudha Agrawal (Advanced Micro Devices, Inc.), Shaizeen Aga (Advanced Micro Devices, Inc.), Suchita Pati (Advanced Micro Devices, Inc.), and Mahzabeen Islam (Advanced Micro Devices, Inc.)

FinGraV: Methodology for Fine-Grain GPU Power Visibility and Insights 96
Varsha Singhania (Advanced Micro Devices, Inc.), Shaizeen Aga (Advanced Micro Devices, Inc.), and Mohamed Assem Ibrahim (Advanced Micro Devices, Inc.)

Dissecting Performance Overheads of Confidential Computing on GPU-Based Systems 108
Yang Yang (University of Virginia, USA), Mohammad Sonji (University of Virginia, USA), and Adwait Jog (University of Virginia, USA)

System Characterization and Framework

Intel® In-Memory Analytics Accelerator: Performance Characterization and Guidelines 124
Jaeyoung Kang (University of Illinois at Urbana-Champaign, U.S.A), Qirong Xia (University of Illinois at Urbana-Champaign, U.S.A), Ipoom Jeong (Department of System Semiconductor Engineering, South Korea; Department of Electrical and Electronic Engineering, South Korea), Yongjoo Park (University of Illinois at Urbana-Champaign, U.S.A), and Nam Sung Kim (University of Illinois at Urbana-Champaign, U.S.A)

Luthier: A Dynamic Binary Instrumentation Framework Targeting AMD GPUs 137
Matin Raayai-Ardakani (Northeastern University, USA), Andrew Nguyen (Northeastern University, USA), Ivan Rosales (Northeastern University, USA), Daoxuan Xu (William & Mary, USA), Yuwei Sun (New York University, USA), Yifan Sun (William & Mary, USA), David Kaeli (Northeastern University, USA), and Norman Rubin (Northeastern University, USA)

Performance Analysis of GEMM Workloads on the AMD Versal Platform 150
Kaustubh Manohar Mhatre (Arizona State University, USA), Venkata Guru Prashanth Mulleti (Arizona State University, USA), Curt John Bansil (Arizona State University, USA), Endri Taka (The University of Texas at Austin, USA), and Aman Arora (Arizona State University, USA)

Beyond the Numbers: Measuring Android Performance Through User Perception 162
Jaeheon Lee (POSTECH), Juhyung Park (DGIST), Seonggyun Oh (DGIST), Jinhyung Koo (POSTECH), and Sungjin Lee (POSTECH)

Benchmarking and Tools

COCOSSim: A Cycle-Accurate Simulator for Heterogeneous Systolic Array Architectures 174
Mansi Choudhary (Duke University, USA), Chris Kjellqvist (Duke University, USA), Jiaao Ma (Duke University, USA), and Lisa Wu Wills (Duke University, USA)

SCALE-Sim v3: A Modular Cycle-Accurate Systolic Accelerator Simulator for End-to-End System Analysis 186
Ritik Raj (Georgia Institute of Technology), Sarbartha Banerjee (University of Texas Austin), Nikhil Chandra (Georgia Institute of Technology), Zishen Wan (Georgia Institute of Technology), Jianming Tong (Georgia Institute of Technology), Ananda Samajdar (IBM Research), and Tushar Krishna (Georgia Institute of Technology)

Evaluation of MindPalace for Chip Design Tradeoffs on Function-as-a-Service 201
Kaifeng Xu (Princeton University, USA), Georgios Tziantzioulis (Princeton University, USA), and David Wentzlaff (Princeton University, USA)

PowerSensor3: A Fast and Accurate Open Source Power Measurement Tool 213
Steven van der Vlugt (ASTRON (Netherlands Institute for Radio Astronomy), the Netherlands), Leon Oostrum (Netherlands eScience Center, the Netherlands), Gijs Schoonderbeek (ASTRON (Netherlands Institute for Radio Astronomy), the Netherlands), Ben van Werkhoven (Leiden University, the Netherlands; Netherlands eScience Center, the Netherlands), Bram Veenboer (ASTRON (Netherlands Institute for Radio Astronomy), the Netherlands), Krijn Doekemeijer (Vrije Universiteit Amsterdam, the Netherlands), and John W. Romein (ASTRON (Netherlands Institute for Radio Astronomy), the Netherlands)

Benchmarking 3D Gaussian Splatting Rendering 227
Saichand Samudrala (Texas A&M University), Sushant Kondguli (Meta), and Paul Gratz (Texas A&M University)

Modeling and Evaluation

An Analytical Cost Model for Fast Evaluation of Multiple Compute-Engine CNN Accelerators 239
Fareed Qararyah (Chalmers University of Technology and University of Gothenburg, Sweden), Mohammad Ali Maleki (Chalmers University of Technology and University of Gothenburg, Sweden), and Pedro Trancoso (Chalmers University of Technology and University of Gothenburg, Sweden)

MeMo: Enhancing Representative Sampling via Mechanistic Micro-Model Signatures 251
Chenji Han (CAS, China; University of Chinese Academy of Sciences, China), Huai Xu (CAS, China; University of Chinese Academy of Sciences, China), Guangyao Guo (CAS, China; University of Chinese Academy of Sciences, China), Yuxuan Wu (Loongson Technology, China), and Fuxin Zhang (CAS, China)

COSMOS: An LLC Contention Slowdown Model for Heterogeneous Multi-Core Systems 264
Yongju Lee (Yonsei University, Korea), Jaewon Kwon (Yonsei University, Korea), Cheolhwan Kim (Yonsei University, Korea), Enhyeok Jang (Yonsei University, Korea), Jiwon Lee (Yonsei University, Korea), Hyunwuk Lee (Yonsei University, Korea), and Won Woo Ro (Yonsei University, Korea)

Use Equal-Work or Equal-Time Speedup, Not Geomean Speedup 276
Lieven Eeckhout (Ghent University, Belgium)

Identifying Important Data Transformations for Synthesizing Effective Lossless Compressors....... 286
Noushin Azami (Texas State University, USA) and Martin Burtscher (Texas State University, USA)

Accelerators and Power

Beethoven: A Heterogeneous Multi-Core Accelerator System Composer 297
Chris Kjellqvist (Duke University, USA), Brendan Peercy (Duke University, USA), Alvin R. Lebeck (Duke University, USA), and Lisa Wu Wills (Duke University, USA)

SAGA: A Surrogate Assisted Genetic Algorithm for Fast CPU Power Virus Generation 309
Panteleimonas Chatzimiltis (University of Cyprus), Georgia Antoniou (University of Cyprus), Haris Volos (University of Cyprus), and Yiannakis Sazeides (University of Cyprus)

Concurrent PIM and Load/Store Servicing in PIM-Enabled Memory 320
Sudhanshu Gupta (University of Rochester), Niti Madan (AMD Research and Advanced Development), Sooraj Puthoor (AMD Research and Advanced Development), Nuwan Jayasena (AMD Research and Advanced Development), and Sandhya Dwarkadas (University of Virginia)

La Superba: Leveraging a Self-Comparison Method to Understand the Performance Benefits of Sparse Acceleration Optimizations 335
Nebil Ozer (University of Maryland, College Park), Gregory Kollmer (Rain AI; University of Texas, Austin), Ramyad Hadidi (Rain AI), and Bahar Asgari (University of Maryland, College Park)

Posters

The Fake-Busy and True-Idle Problems of Running Graph Applications on Chiplet-Based Multi-Cores 347
Rashid Aligholipour (Uppsala University, Sweden) and Yuan Yao (Uppsala University, Sweden)

The Future of Instruction-Level Parallelism (ILP) 350
Alexandra W. Chadwick (University of Cambridge, UK), Márton Erdős (University of Cambridge, UK), Utpal Bora (University of Cambridge, UK), Akshay Bhosale (University of Cambridge, UK), Bob Lytton (Arm, UK), Yuxin Guo (University of Cambridge, UK), Richard Cooper (Arm, UK), Giacomo Gabrielli (Arm, UK), and Timothy M. Jones (University of Cambridge, UK)

Characterizing Compute-Communication Overlap in GPU-Accelerated Distributed Deep Learning: Performance and Power Implications 353
Seonho Lee (Georgia Institute of Technology, USA), Jihwan Oh (KAIST, South Korea), Seokjin Go (Georgia Institute of Technology, USA), and Divya Mahajan (Georgia Institute of Technology, USA)

PIM-BEACON: A Benchmarking and Emulation Framework Supporting Adaptive CONfigurations in DRAM-Based Processing-in-Memory Systems 356
Inseong Hwang (Seoul National University of Science and Technology, Republic of Korea), Jihoon Jang (Seoul National University of Science and Technology, Republic of Korea), Chaewon Park (Seoul National University of Science and Technology, Republic of Korea), and Hyun Kim (Seoul National University of Science and Technology, Republic of Korea)

Profiling Concurrent Vision Inference Workloads on NVIDIA Jetson 359
Abhinaba Chakraborty (Ghent University-IMEC, Belgium), Wouter Tavernier (Ghent University-IMEC, Belgium), Akis Kourtis (NCSR Demokritos, Greece), Mario Pickavet (Ghent University-IMEC, Belgium), Andreas Oikonomakis (NCSR Demokritos, Greece), and Didier Colle (Ghent University-IMEC, Belgium)

ASLink: Modeling Multi-GPU Execution in Accel-Sim 362
Christin Bose (Purdue University, USA), Cesar Avalos (Purdue University, USA), Junrui Pan (Purdue University, USA), Yechen Liu (Purdue University, USA), Mahmoud Khairy (Purdue University, USA), Clay Hughes (Sandia National Laboratories, USA), and Timothy Rogers (Purdue University, USA)

FIDESlib: A Fully-Fledged Open-Source FHE Library for Efficient CKKS on GPUs 365
Carlos Agulló-Domingo (Universidad de Murcia), Óscar Vera-López (Universidad de Murcia), Seyda Guzelhan (Boston University), Lohit Daksha (Boston University), Aymane El Jerari (Northeastern University), Kaustubh Shivdikar (AMD), Rashmi Agrawal (Boston University), David Kaeli (Northeastern University), Ajay Joshi (Boston University), and José L. Abellán (Universidad de Murcia)

Carbon-Aware Server Replacement 368
Iris Uwizeyimana (University of Toronto, Canada) and Natalie Enright Jerger (University of Toronto, Canada)

A Flexible and Accurate Circuit-Level Substrate for Future DRAM Design and Analysis 371
S M Mojahidul Ahsan (The University of Kansas), Mohammad Nouri (Cornell University), Ramesh Reddy Ganapam (The University of Kansas), Mohammad Alian (Cornell University), and Tamzidul Hoque (The University of Kansas)

TPNM: A CXL Based General Purpose Tiered Process Near Memory Framework 374
Pingyi Huo (The Pennsylvania State University, USA), Anusha Devulapally (The Pennsylvania State University, USA), Hasan Al Maruf (AMD, Inc, USA), Meena Arunachalam (AMD, Inc, USA), Mahmut Taylan Kandemir (The Pennsylvania State University, USA), and Vijaykrishnan Narayanan (The Pennsylvania State University, USA)

A Real-Time, Auto-Regression Method for In-Situ Feature Extraction in Hydrodynamics Simulations 377
Kewei Yan (University of North Carolina at Charlotte, USA) and Yonghong Yan (University of North Carolina at Charlotte, USA)

Library of Networks: An Online Tool for Design and Analysis of Network Topologies 379
Aniket Chatterjee (Purdue University), Conor James Green (Purdue University), and Mithuna Thottethodi (Purdue University)

Multi-Core Aware Evaluation of Prefetchers 382
Marti Torrents (Intel Corporation), Paul Caheny (Intel Corporation), Stijn Eyerman (Intel Corporation), and Wim Heriman (Intel Corporation)

Evaluation and Comparison of the Energy Efficiency of Several Intel Multicore Processors 385
Thomas Rauber (University Bayreuth, Germany) and Gudula Rünger (Technical University Chemnitz, Germany)

Analysis of the RISC-V Vector Extension for Vulkan Graphics Kernels 388
Martin Troiber (Technical University of Munich, Germany), Martin Schulz (Technical University of Munich, Germany), Blaise Tine (University of California, USA), and Hyesoon Kim (Georgia Institute of Technology, USA)

GPU Simulation Acceleration via Parallelization 390
Rodrigo Huerta (Universitat Politècnica de Catalunya) and Antonio González (Universitat Politècnica de Catalunya)

Energon: A Sustainability-Driven Modeling Framework for AI Data Centers 393
Wenzhe Guo (imec, Belgium), Joyjit Kundu (imec, Belgium), Uras Tos (imec, Belgium), Giuliano Sisto (imec, Belgium), Cedric Rolin (imec, Belgium), Lars-Ake Ragnarsson (imec, Belgium), and Timon Evenblij (imec, Belgium)

Evaluating Compute in Memory Architectures for Matrix Multiplication: A Dataflow-Centric Perspective 396
Tanvi Sharma (Purdue University), Indranil Chakraborty (Google TPU), Mustafa Ali (Microsoft), and Kaushik Roy (Purdue University)

Measuring Performance Overheads of Software Memory Management using Functional-First Simulators 399
Yves Vandriessche (Intel Corporation), Wim Heirman (Intel Corporation), Ed Nutting (VyperCore), Jeremy Birch (VyperCore), Judah Daniels (VyperCore), Mae Hood (VyperCore), and Pascal Costanza (Intel Corporation)

Interconnect Performance Estimation for ML Accelerators via Lightweight Analytical Model 401
Rahul Tripathy (Indian Institute of Science, Bangalore) and Sumit K. Mandal (Indian Institute of Science, Bangalore)

Author Index **405**