

2026 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2026)

**Seoul, Republic of Korea
26-28 April 2026**

IEEE Catalog Number:
ISBN:

CFP26PER-POD
979-8-3315-5104-9



Copyright © 2026, 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:	CFP26PER-POD
ISBN (Print-On-Demand):	979-8-3315-5104-9
ISBN (Online):	979-8-3315-5103-2
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

TABLE OF CONTENTS

LLMServingSim 2.0: A Unified Simulator for Heterogeneous and Disaggregated LLM Serving Infrastructure	1
<i>Jaehong Cho, Hyunmin Choi, Guseul Heo, Jongse Park</i>	
Characterizing State Space Model and Hybrid Language Model Performance with Long Context	15
<i>Saptarshi Mitra, Rachid Karami, Haocheng Xu, Sitao Huang, Hyoukjun Kwon</i>	
TaxBreak: Unmasking the Hidden Costs of LLM Inference Through Overhead Decomposition	28
<i>Prabhu Vellaisamy, Shreesh Tripathi, Vignesh Natarajan, Surya Santhan Thenarasu, Shawn Blanton John P. Shen</i>	
The xPU-athalon: Quantifying the Competition of AI Acceleration	39
<i>Alicia Golden, Carole-Jean Wu, Gu-Yeon Wei, David Brooks</i>	
HARMONI: Hierarchical ARchitecture MOdeling for LLMs with Near/In Memory Computing	51
<i>Khyati Kiyawat, Yasas Seneviratne, Zhenxing Fan, Morteza Baradaran, Kevin Skadron</i>	
TensorDynamic: Bridging Application- and Instruction-Level Fault Injection for DNN Tensor Core Execution	64
<i>Yuxiao Jia, Euijun Chung, Huanzhi Pu, Ben Feinberg, Hyesoon Kim</i>	
GTPin: Enhancing Intel GPU Profiling With High-Level Binary Instrumentation	77
<i>Konstantin Levit-Gurevich, Alex Skaletsky, Ortal Geller, Moshe Goren, Rawan Ziadat, Rami Burstein Tal Oved</i>	
FlexNoC: Fast and Flexible Analysis for NoCs with Arbitrary Topologies and Hybrid Arbitration	90
<i>Anuparna Ganguly, Rahul Tripathy, Tyson Loveless, Mohammad Majharul Islam, Sumit K. Mandal</i>	
Characterization of Coherent Data Movement in Scaled-Out Shared Memory Systems	104
<i>Maryam Babaie, Michael Miller, Wendy Elsasser, Steven Woo, Jason Lowe-Power</i>	
Characterization and Performance Analysis of Aging Effect on Enterprise HDDs	115
<i>Ava Dezhban, Mohammadamin Ajdari, Hossein Asadi</i>	
It's Not the Storage: Exposing Linux Kernel Bottlenecks in Page Fault Handling on NVMe SSD-backed Systems	125
<i>Alireza Haddadi, David Black-Schaffer, Chang Hyun Park</i>	
Do Not Get Fooled: Use True Device Utilization on High-Performance Storage Devices	137
<i>Ali Sedaghatgoo, Mohammadamin Ajdari, Erfan Teymouri, Hossein Asadi</i>	
Compiler and System Optimizations for Gem5 Simulator	149
<i>Haneul Park, Siddharth Agarwal, Pradyun Narkadamilli, Kiung Jung, Yongjun Park, Ipoom Jeong Nam Sung Kim</i>	
Understanding Simulated Architecture via gem5 Call-Stack Profiling	161
<i>Johan Söderström, Rashid Aligholipour, Yuan Yao</i>	
Multi-Level TDA for Heterogeneous Arm Processors Using a Standardized Telemetry Framework	173
<i>Jumana Mundichipparakkal, Darin Greene, Joshua Randall, Vincent Coubard, Michael Williams Rene De Jong</i>	

Toward Reproducible and Standardized Computer Architecture Simulation with gem5	184
<i>Kunal Pai, Harshil Patel, Erin Le, Noah Krim, Mahyar Samani, Bobby R. Bruce, Jason Lowe-Power</i>	
Hibiscus: End-to-end Architectural Simulation Framework for Hybrid SFQ/CMOS-Memory Compute Systems	197
<i>Ryan Marsala, Yerzhan Mustafa, Prabhath Tangella, Mohammad Sonji, George Michelogiannakis, Selcuk Kose, Adwait Jog, Mehmet E. Belviranli</i>	
Cleaning up the Mess: Re-Evaluating the Real-System Modeling Accuracy of Ramulator 2.0	209
<i>F. Nisa Bostancı, Haocong Luo, Ataberk Olgun, Maria Makeenkova, Geraldo F. de Oliveira, A. Giray Yağlıkcı, Onur Mutlu</i>	
A Heterogeneous Mapping of a Speech Neuroprosthesis Application	220
<i>Iris Uwizeyimana, Oscar Sun, Natalie Enright Jerger</i>	
CCPA: A Concurrent Content Processing Architecture for Hardware Firewalls	232
<i>Shuangliang Chen, Saptadeep Pal, Rakesh Kumar</i>	
UPSORT: Design and Analysis of Processing-in-Memory Sorting Algorithms using UPMEM	244
<i>Ivan Vargas Vadivieso, Julian Pavon, Carlos Rojas, Javier Chavez Rubio, Marco Ramirez, Luis Villa, Mateo Valero, Osman Unsal, Adrian Cristal</i>	
A Comprehensive Study of Performance Degradation for Diverse Intensive Workloads in RocksDB	256
<i>Haoyu Gong, Zhaokang Ke, David Du</i>	
Scalable Assembler System: Optimizing Performance through Unified Data Structures and Procedure	267
<i>Xinyu Pan, Xi Li, Jian Zhang, Meiguang Zheng</i>	
DyPNet-MSc: Dynamic Bandwidth Allocation in Photonic Network-on-Wafer GPU Architectures	277
<i>Ziyue Zhang, Hossein SeyyedAghaei, Benyamin Eslami, Xin Wang, Gunther Roelkens, Didier Colle, Mario Pickavet, Lieven Eeckhout</i>	
Understanding the Performance Implications of Oblivious RAM (ORAM)	288
<i>Biniyam Tiruye, Sitota Mersha, Kalab Assefa, Samuel Owens, Lauren Biernacki, Todd Austin</i>	
Theodosian: A Deep Dive into Memory-Hierarchy-Centric FHE Acceleration	299
<i>Wonseok Choi, Hyunah Yu, Jongmin Kim, Hyesung Ji, Jaiyoung Park, Jung Ho Ahn</i>	
An Empirical Study of LLM Serving in Confidential GPUs	311
<i>Eunseong Park, Timo Thans, Vishnu Kumar Kalidasan, Qinghao Hu, Wenjie Xiong</i>	
Beyond Latency: A System-Level Characterization of MPC and FHE for PPML	323
<i>Pengzhi Huang, Kiwan Maeng, G. Edward Suh</i>	
CryptOracle: A Modular Framework to Characterize FHE	335
<i>Cory Brynds, Parker McLeod, Lauren Caccamise, Asmita Pal, Dewan Saiham, Sazadur Rahman, Joshua San Miguel, Di Wu</i>	
FDPEmu: How to Separate Workloads for Better WAF on FDP SSDs	348
<i>Nakyeong Kim, Kwanghee Lee, Bryan S. Kim, Seehwan Yoo, Jaedong Lee, Jongmoo Choi</i>	
FYI: A Foundational Counter Architecture for Building CPI Stacks on Out-of-Order Processors	358
<i>Silvio Campelo de Santana, Lieven Eeckhout, Magnus Jahre</i>	

Understanding BTB Tag Sizing	369
<i>Roman K. Brunner, Rakesh Kumar</i>	
HOTP - High-conf Only Taken Predictor for Efficient Superscalar Multi-level Branch Predictor	380
<i>Bhawandeep Singh Harsh, Jose Renau</i>	
Correct Wrong Path Simulation	391
<i>Chrysanthos Pepi, Krishnam Tibrewala, Bhargav Reddy Godala, Sankara Prasad Ramesh, Alberto Ros</i>	
<i>Daniel A. Jiménez, Gilles A. Pokam, Paul V. Gratz</i>	
Pesto: Diagnosing Performance Pathologies in Out-of-Order Processors	405
<i>Silvio Campelo de Santana, Joseph Rogers, Lieven Eeckhout, Magnus Jahre</i>	
Characterizing Synonyms and TLB Invalidations for Virtually-Indexed Virtually-Tagged Caches	416
<i>Sotiris Totomis, Kallia Chronaki, Vassilis Papaefstathiou</i>	
Hardware Generation and Exploration of Lookup Table-Based Accelerators for 1.58-bit LLM Inference	426
<i>Robin Geens, Joran Heldens, Joren Dumoulin, Marian Verhelst</i>	
EnergAIzer: Fast and Accurate GPU Power Estimation Framework for AI Workloads	435
<i>Kyungmi Lee, Zhiye Song, Eun Kyung Lee, Xin Zhang, Tamar Eilam, Anantha P. Chandrakasan</i>	
Evaluating Cross-Architecture Performance Modeling of Distributed ML Workloads Using StableHLO	448
<i>Jonas Svedas, Nathan Laubeuf, Ryan Harvey, Arjun Singh, Changhai Man, Abubakr Nada, Tushar Krishna</i>	
<i>James Myers, Debjyoti Bhattacharjee</i>	
RACER: A Caching Layer based Optimization for Accelerating Reinforcement Learning Workloads	461
<i>Kailash Gogineni, Yongsheng Mei, Karthikeya Gogineni, Peng Wei, Tian Lan, Guru Venkataramani</i>	
A Cluster-Based Sampling Scheme for High-Performance Graph Neural Networks	473
<i>Surendra Kumar Raut, Kishan Tamboli, Vishwesh Jatala</i>	
HeadSkip: Characterizing and Accelerating ‘Small’ Language Models on Edge CPUs	484
<i>Ian Sartor, Karthik Ganesan, Anh Nguyen, Andreas Moshovos</i>	
ASI: A Unifying Metric to Assess and Improve Processor Microarchitecture Sustainability	495
<i>Jaime Roelandts, Ajeya Naithani, Lieven Eeckhout</i>	
Linearizability Alone Is Not Enough for Strict FIFO Semantics	507
<i>Ryan Huard, Md Sabbir Hossain Polak, Joshua Jurgenson, David Troendle, Byunghyun Jang</i>	
eBeeMetrics: An eBPF-based Library Framework for Feedback-free Observability of QoS Metrics	518
<i>Muntaka Ibnath, Mohammadreza Rezvani, Daniel Wong</i>	
PLCG: Parametric Loop Code Generator for Loop Transformation Optimization Benchmarking	531
<i>Yijie Zhi, Qinran Wu, Xiaoyang Han, Ming Cai</i>	
Reclaiming Energy: Quantifying the Energy Overheads of Garbage Collection on Mobile Devices	543
<i>Kunal Sareen, Stephen M. Blackburn</i>	
Exploring the Energy Storage and Voltage Control Unit Design Space in Battery-Less IoT	556
<i>Lukas Liedtke, Espen Holsen, Per Gunnar Kjeldsberg, Frank Alexander Kraemer, Magnus Jahre</i>	
Situla: Studying the Interplay of Sparse Formats and CPU/GPU Libraries	567
<i>Amirmahdi Namjoo, Sanjali Yadav, Helya Hosseini, Bahar Asgari</i>	

Breaking the I/O Bottleneck: I/O Coordination Optimization for Efficient Large-Scale LLM Fine-Tuning	578
<i>Ziyang Shen, Hongchao Du, Kaihuan Lin, Dawei Wu, Yin Lin, Ting Luo, Qiao Li, Chun Jason Xue</i>	
Towards System-2 AI: Workloads and Characterizations of Energy-Based Models	589
<i>Hanchen Yang, Jiayi Qian, Zishen Wan, Jingtian Dang, Ziwei Li, Yilun Du, Tushar Krishna</i>	
Surrogates, Spikes, and Sparsity: Performance Analysis and Characterization of SNN Hyperparameters on Hardware	601
<i>Ilkin Aliyev, Jesus Lopez, Tosiron Adegbiya</i>	
MINISA: Minimal Instruction Set Architecture for Next-gen Reconfigurable Inference Accelerator	614
<i>Jianming Tong, Devansh Jain, Yujie Li, Charith Mendis, Tushar Krishna</i>	
Characterizing Machine Learning Force Fields as Emerging Molecular Dynamics Workloads on Graphics Processing Units	627
<i>Udari De Alwis, Benjamin E. Mayer, Thomas J. Ashby, Maria Barrera, Timon Evenblij, Joyjit Kundu</i>	
Macsim Mini: A Lightweight Cycle-Level GPU Simulator for Architecture Education	638
<i>Euijun Chung, Saurabh Singh, Huanzhi Pu, Yuxiao Jia, Anurag Kar, Sam Jijina, Scott Madeira Hyesoon Kim</i>	
RTG: Exploring the Dynamic Scheduling Space of Real-Time Generative AI workloads on Emerging Heterogeneous Systems	641
<i>Rachid Karami, Rajeev Patwari, Hyukjun Kwon, Ashish Sirasao</i>	
TTX: Towards Autotuning Triton Kernels via Latency Prediction with XGBoost	644
<i>Van-Sang Pham, Tien-Son Pham, Tuan-Duc Chu, Xuan Truong Nguyen, Thanh-Tuan Dao</i>	
CPU Bottlenecks in Rack-Stand Brain-Computer Interfaces: A Phase-Aware Characterisation	647
<i>Victor Kariofillis, Iris Uwizeyimana, Sara Ahmad, Tanvi Manku, Raghavendra Pradyumna Pothukuchi Abhishek Bhattacharjee, Natalie Enright Jerger</i>	
VCSR-MM: A Bandwidth-Optimized SpMM Design for Modern GPUs	650
<i>Mouad Tiahi, Elmira Karimi, David Kaeli</i>	
LLM4PART: From CallGraphs to Bottleneck Partitions with LLM-Enhanced Insights	653
<i>Vyuhita Bonthu, Venkatesh Pasumarti, Ashwin Krishnan, Manoj Nambiar</i>	
Characterizing and Optimizing Cache Placement for Secure Memory Metadata	655
<i>Samuel Thomas, Blake Cragen, R. Iris Bahar, Tamara Silbergleit Lehman</i>	
Revisiting the RAID Performance Bottleneck in the SSD Era: A Roofline Model Perspective	658
<i>Jun Zhao, Xiaoyang Wang, Shuo Wang, Congming Gao, Jiwu Shu</i>	
Power Ranger: A Comprehensive Framework for GPU Power Characterization	661
<i>Allison Seigler, Zhixing Jiang, Lizy Kurian John</i>	
COMET: A Framework for Modeling Compound Operation Dataflows with Explicit Collectives	664
<i>Shubham Negi, Manik Singhal, Aayush Ankit, Sudeep Bhoja, Kaushik Roy</i>	
Characterizing Cache-Asymmetric CPU Topologies on AMD 3D V-Cache Processors	667
<i>Sunwoo Kim, Eunbi Jeong, Myung Kuk Yoon</i>	