

2022 IEEE International Symposium on Workload Characterization (IISWC 2022)

**Austin, Texas, USA
6 – 8 November 2022**



**IEEE Catalog Number: CFP22236-POD
ISBN: 978-1-6654-8799-3**

**Copyright © 2022 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:	CFP22236-POD
ISBN (Print-On-Demand):	978-1-6654-8799-3
ISBN (Online):	978-1-6654-8798-6

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

CURRAN ASSOCIATES INC.
proceedings
.com

2022 IEEE International Symposium on Workload Characterization (IISWC) IISWC 2022

Table of Contents

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

Microarchitecture/HW Performance Analysis

PInTE: Probabilistic Induction of Theft Evictions	1
<i>Cesar Gomes (Tufts University), Xuesi Chen (Carnegie Mellon University), and Mark Hempstead (Tufts University)</i>	
GRANITE: A Graph Neural Network Model for Basic Block Throughput Estimation	14
<i>Ondrej Sykora (Google Research), Mangpo Phothilimthana (Google Research, Brain Team), Charith Mendis (University of Illinois Urbana-Champaign), and Amir Yazdanbakhsh (Google Research, Brain Team)</i>	
UVM Discard: Eliminating Redundant Memory Transfers for Accelerators	27
<i>Weixi Zhu (Rice University), Guilherme Cox (NVIDIA), Jan Vesely (NVIDIA), Mark Hairgrove (NVIDIA), Alan L. Cox (Rice University), and Scott Rixner (Rice University)</i>	

HPC

FPChecker: Floating-Point Exception Detection Tool and Benchmark for Parallel and Distributed HPC	39
<i>Ignacio Laguna (Lawrence Livermore National Laboratory), Tanmay Tirpankar (University of Utah), Xinyi Li (University of Utah), and Ganesh Gopalakrishnan (University of Utah)</i>	
Splash-4: A Modern Benchmark Suite with Lock-Free Constructs	51
<i>Eduardo José Gómez-Hernández (University of Murcia), Juan M. Cebrian (University of Murcia), Stefanos Kaxiras (Uppsala University), and Alberto Ros (University of Murcia)</i>	

Characterizing Molecular Dynamics Simulation on Commodity Platforms	65
<i>Francesco Peverelli (DEIB, Politecnico di Milano), Davide Conficconi (DEIB, Politecnico di Milano), Davide Basilio Bartolini (Systems Laboratory, Zurich Research Center, Huawei Technologies), Alberto Scolari (Systems Laboratory, Zurich Research Center, Huawei Technologies), and Marco Domenico Santambrogio (DEIB, Politecnico di Milano)</i>	

AI Systems

An Evaluation of Edge TPU Accelerators for Convolutional Neural Networks	79
<i>Kiran Seshadri (Enfabrica), Berkin Akin (Google), James Laudon (Google Research, Brain Team), Ravi Narayanaswami (Cruise), and Amir Yazdanbakhsh (Google Research, Brain Team)</i>	
Accelerating Transformer Networks through Recomposing Softmax Layers	92
<i>Jaewan Choi (Seoul National University), Hailong Li (Seoul National University), Byeongho Kim (Samsung Electronics), Seunghwan Hwang (Seoul National University), and Jung Ho Ahn (Seoul National University)</i>	
A Slice and Dice Approach to Accelerate Compound Sparse Attention on GPU	104
<i>Hailong Li (Seoul National University), Jaewan Choi (Seoul National University), and Jung Ho Ahn (Seoul National University)</i>	
FedGPO: Heterogeneity-Aware Global Parameter Optimization for Efficient Federated Learning...	117
<i>Young Geun Kim (Korea University) and Carole-Jean Wu (Arizona State University / Meta)</i>	

Graph Neural Networks

Bottleneck Analysis of Dynamic Graph Neural Network Inference on CPU and GPU	130
<i>Hanqiu Chen (Georgia Institute of Technology), Yihan Jiang (Georgia Institute of Technology), Yahya Alhinai (Georgia Institute of Technology), Eunjee Na (Korea Advanced Institute of Science & Technology), and Cong Hao (Georgia Institute of Technology)</i>	
gSuite: A Flexible and Framework Independent Benchmark Suite for Graph Neural Network Inference on GPUs	146
<i>Taha Tekdoğan (Istanbul Technical University; ASELSAN Inc.), Serkan Göktaş (Istanbul Technical University), and Ayşe Yilmazer-Metin (Istanbul Technical University)</i>	
Characterizing the Efficiency of Graph Neural Network Frameworks with a Magnifying Glass	160
<i>Xin Huang (Texas State University), Jongryool Kim (SK hynix America), Bradley Rees (NVIDIA), and Chul-Ho Lee (Texas State University)</i>	

Graph Analytics and GPUs

Performance Characterization of AutoNUMA Memory Tiering on Graph Analytics	171
<i>Diego Moura (Federal University of Bahia), Daniel Mossé (University of Pittsburgh), and Vinicius Petrucci (Micron Technology)</i>	

Understanding the Power of Evolutionary Computation for GPU Code Optimization	185
<i>Jhe-Yu Liou (Arizona State University), Muaaz Awan (Lawrence Berkeley National Laboratory), Steven Hofmeyr (Lawrence Berkeley National Laboratory), Stephanie Forrest (Arizona State University), and Carole-Jean Wu (Arizona State University)</i>	
The Implications of Page Size Management on Graph Analytics	199
<i>Aninda Manocha (Princeton University), Zi Yan (NVIDIA), Esin Tureci (Princeton University), Juan Luis Aragon (University of Murcia), David Nellans (NVIDIA), and Margaret Martonosi (Princeton University)</i>	

Mobile, Web, and Cloud

Revisiting Temporal Storage I/O Behaviors of Smartphone Applications: Analysis and Synthesis	215
<i>Qiang Zou (Southwest University) and Bo Mao (Xiamen University)</i>	
How Far We've Come – A Characterization Study of Standalone WebAssembly Runtimes	228
<i>Wenwen Wang (University of Georgia)</i>	
SpotLake: Diverse Spot Instance Dataset Archive Service	242
<i>Sungjae Lee (Kookmin University), Jaeil Hwang (Kookmin University), and Kyungyong Lee (Kookmin University)</i>	
Leaps and Bounds: Analyzing WebAssembly's Performance with a Focus on Bounds Checking	256
<i>Raven Szezewczyk (University of Edinburgh), Kimberley Stonehouse (University of Edinburgh), Antonio Barbalace (University of Edinburgh), and Tom Spink (University of St Andrews)</i>	

AI Benchmarks & Characterization

Demystifying Map Space Exploration for NPUs	269
<i>Sheng-Chun Kao (Georgia Institute of Technology), Angshuman Parashar (NVIDIA), Po-An Tsai (NVIDIA), and Tushar Krishna (Georgia Institute of Technology)</i>	
LongTail-Bench: A Benchmark Suite for Domain-Specific Operators in Deep Learning	282
<i>Xiuhong Li (SenseTime Research; Shanghai AI Lab), Shengen Yan (Peking University), Lijuan Jiang (Shanghai AI Lab), Ping Xu (SenseTime Research), Jinming Ma (Shanghai AI Lab), Xingcheng Zhang (SenseTime Research; Shanghai AI Lab), and Dahua Lin (Shanghai AI Lab; The Chinese University of Hong Kong)</i>	
Demystifying BERT: System Design Implications	296
<i>Suchita Pati (University of Wisconsin-Madison; Advanced Micro Devices Inc.), Shaizeen Aga (Advanced Micro Devices Inc.), Nuwan Jayasena (Advanced Micro Devices Inc.), and Matthew D. Sinclair (University of Wisconsin-Madison; Advanced Micro Devices Inc.)</i>	

Author Index	311
---------------------------	------------