2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2021)

Virtual Conference
28 – 30 March 2021

IEEE Catalog Number: CFP21PER-POD
ISBN: 978-1-7281-8644-3

Copyright © 2021 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:	CFP21PER-POD
ISBN (Print-On-Demand):	978-1-7281-8644-3
ISBN (Online):	978-1-7281-8643-6

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

2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)

ISPASS 2021

Table of Contents

Message from the General Chair xii

Message from the Program Chair xiii

Organizing Committee xiv

Program Committee xv

Steering Committee xvi

Keynotes xvii

Sponsors xix

Paper Session I: Benchmarking

GenomicsBench: A Benchmark Suite for Genomics 1
Arun Subramaniyan (University of Michigan, USA), Yufeng Gu (University of Michigan, USA), Timothy Dunn (University of Michigan, USA), Somnath Paul (Intel Corporation, USA), Md Vasimuddin (Intel Corporation, India), Sanchit Misra (Intel Corporation, India), David Blaauw (University of Michigan, USA), Satish Narayanasamy (University of Michigan, USA), and Reetuparna Das (University of Michigan, USA)

GNNMark: A Benchmark Suite to Characterize Graph Neural Network Training on GPUs 13
Trinayan Baruah (Northeastern University), Kaustubh Shivdikar (Northeastern University), Shi Dong (Cerebras Systems), Yifan Sun (William & Mary), Saiful A Mojumder (Boston University), Kihoon Jung (KAIST), José L. Abellán (Universidad Católica de Murcia), Yash Ukidave (AMD), Ajay Joshi (Boston University), John Kim (KAIST), and David Kaeli (Northeastern University)

AIBench Training: Balanced Industry-Standard AI Training Benchmarking 24
Fei Tang (Chinese Academy of Sciences; University of Chinese Academy of Sciences), Wanling Gao (Chinese Academy of Sciences; International Open Benchmark Council (BenchCouncil); University of Chinese Academy of Sciences), Jianfeng Zhan (Chinese Academy of Sciences; International Open Benchmark Council (BenchCouncil); University of Chinese Academy of Sciences), Chuanxin Lan (Chinese Academy of Sciences), Xu Wen (Chinese Academy of Sciences; University of Chinese Academy of Sciences), Lei Wang (Chinese Academy of Sciences; International Open Benchmark Council (BenchCouncil); University of Chinese Academy of Sciences), Chunjie Luo (Chinese Academy of Sciences; University of Chinese Academy of Sciences), Zheng Cao (Alibaba), Xingwang Xiong (Chinese Academy of Sciences), Zihan Jiang (Chinese Academy of Sciences), Tianshu Hao (Chinese Academy of Sciences), Fanda Fan (Chinese Academy of Sciences), Fan Zhang (Chinese Academy of Sciences), Yunyou Huang (International Open Benchmark Council (BenchCouncil)), Jianan Chen (Chinese Academy of Sciences), Mengjia Du (Chinese Academy of Sciences), Rui Ren (Chinese Academy of Sciences), Chen Zheng (Chinese Academy of Sciences), Daoyi Zheng (Baidu), Haoning Tang (Tencent), Kunlin Zhan (58.com), Biao Wang (NetEase), Defei Kong (ByteDance), Minghe Yu (Zhihu), Chongkang Tan (Lenovo), Huan Li (Paypal), Xinhui Tian (Moqi), Yatao Li (Microsoft Research Asia), Junchao Shao (JD.com), Zhenyu Wang (CloudTa), Xiaoyu Wang (Intellifusion), Jiahui Dai (International Open Benchmark Council (BenchCouncil)), and Hainan Ye (International Open Benchmark Council (BenchCouncil))

Paper Session II: GPUs

CoCoPeLia: Communication-Computation Overlap Prediction for Efficient Linear Algebra on GPUs 36
Petros Anastasiadis (National Technical University of Athens, Greece), Nikela Papadopoulou (National Technical University of Athens, Greece), Georgios Goumas (National Technical University of Athens, Greece), and Nectarios Koziris (National Technical University of Athens, Greece)

Learning Sparse Matrix Row Permutations for Efficient SpMM on GPU Architectures 48
Atefeh Mehrabi (Duke University), Donghyuk Lee (NVIDIA), Niladrish Chatterjee (NVIDIA), Daniel J. Sorin (Duke University), Benjamin C. Lee (University of Pennsylvania), and Mike O'Connor (NVIDIA; UT Austin)

Analyzing Secure Memory Architecture for GPUs 59
Shougang Yuan (North Carolina State University, USA), Ardhi Wiratama Baskara Yudha (University of Central Florida), Yan Solihin (University of Central Florida), and Huiyang Zhou (North Carolina State University, USA)

Poster Session A:

MicroGrad: A Centralized Framework for Workload Cloning and Stress Testing .70
Gokul Subramanian Ravi (University of Chicago), Ramon Bertran (IBM Research), Pradip Bose (IBM Research), and Mikko Lipasti (University of Wisconsin-Madison)

ViStA: Video Streaming and Analytics Benchmark .73
Navneet Raju (PES University, India), Rahul M Koushik (PES University, India), Hari Om (PES University, India), and Subramaniam Kalambur (PES University, India)

Analysis of Factors Affecting Power Consumption and Energy Efficiency of SGEMM on the Low-Power Myriad-2 VPU .76
Suyash Bakshi (University of Houston, USA) and Lennart Johnsson (University of Houston, USA)

A Defense-Inspired Benchmark Suite .79
Pete Ehrett (University of Michigan), Nathan Block (University of Michigan), Bing Schaefer (University of Michigan), Adrian Berding (University of Michigan), John Paul Koenig (University of Michigan), Pranav Srinivasan (University of Michigan), Valeria Bertacco (University of Michigan), and Todd Austin (University of Michigan)

An Automated Traffic Generation Framework for Performance Evaluation of Networks-on-Chip for Real World Use Cases .81
Sri Harsha Gade (Arm Ltd., Bangalore, KA, India), Anup Gangwar (Arm Ltd., Austin, TX, USA), Ambica Prasad (Arm Ltd., Bangalore, KA, India), Nitin Kumar Agarwal (Arm Ltd., Bangalore, KA, India), and Ravishankar Sreedharan (Arm Ltd., Bangalore, KA, India)

How Do Graph Relabeling Algorithms Improve Memory Locality? .84
Mohsen Koohi Esfahani (Queen's University Belfast, UK), Peter Kilpatrick (Queen's University Belfast, UK), and Hans Vandierendonck (Queen's University Belfast, UK)

Designing GPU Architecture for Memory Bandwidth Reservation .87
Emir C. Marangoz (The State University of New York at Binghamton), Kyoung-Don Kang (The State University of New York at Binghamton), and Seunghee Shin (The State University of New York at Binghamton)

Reducing BERT Computation by Padding Removal and Curriculum Learning .90
Wei Zhang (Alibaba Group), Wei Wei (Alibaba Group), Wen Wang (Alibaba Group), Lingling Jin (Alibaba Group), and Zheng Cao (Alibaba Group)

Efficient Split Counter Mode Encryption for NVM .93
Qi Pei (The State University of New York at Binghamton) and Seunghee Shin (The State University of New York at Binghamton)

Paper Session III: Characterization

AI Tax in Mobile SoCs: End-to-End Performance Analysis of Machine Learning in Smartphones .96
Michael Buch (Harvard University), Zahra Azad (Boston University), Ajay Joshi (Boston University), and Vijay Janapa Reddi (Harvard University)

Performance Characterization of .NET Benchmarks 107
Aniket Deshmukh (The University of Texas at Austin), Ruihao Li (The University of Texas at Austin), Rathijit Sen (Microsoft Gray Systems Lab), Robert R. Henry (Microsoft), Monica Beckwith (Microsoft), and Gagan Gupta (Microsoft)

Performance Analysis of Graph Neural Network Frameworks 118
Junwei Wu (University of Science and Technology of China, China), Jingwei Sun (University of Science and Technology of China, China), Hao Sun (University of Science and Technology of China, China), and Guangzhong Sun (University of Science and Technology of China, China)

Paper Session IV: Software Analysis

Loopapalooza: Investigating Limits of Loop-Level Parallelism with a Compiler-Driven Approach 128
Ali Mustafa Zaidi (Arm Research), Konstantinos Iordanou (University of Manchester, UK), Mikel Luján (University of Manchester, UK), and Giacomo Gabrielli (Arm Research)

Real-Time Characterization of Data Access Correlations 139
Bryan Harris (University of Louisville, USA), Michael Marzullo (University of Louisville, USA), and Nihat Altiparmak (University of Louisville, USA)

Comparative Code Structure Analysis using Deep Learning for Performance Prediction 151
Tarek Ramadan (Texas State University), Tanzima Z. Islam (Texas State University), Chase Phelps (Texas State University), Nathan Pinnow (Lawrence Livermore National Laboratory), and Jayaraman J. Thiagarajan (Lawrence Livermore National Laboratory)

Paper Session V: Best Paper Nominations

Understanding Capacity-Driven Scale-Out Neural Recommendation Inference 162
Michael Lui (Facebook; Drexel University), Yavuz Yetim (Facebook), Özgür Özkan (Facebook), Zhuoran Zhao (Facebook), Shin-Yeh Tsai (Facebook), Carole-Jean Wu (Facebook), and Mark Hempstead (Facebook; Tufts University)

Re-Establishing Fetch-Directed Instruction Prefetching: An Industry Perspective 172
Yasuo Ishii (Arm, Austin, TX, USA), Jaekyu Lee (Arm, Austin, TX, USA), Krishnendra Nathella (Arm, Austin, TX, USA), and Dam Sunwoo (Arm, Austin, TX, USA)

Enabling Reproducible and Agile Full-System Simulation 183
Bobby R. Bruce (UC Davis), Ayaz Akram (UC Davis), Hoa Nguyen (UC Davis), Kyle Roarty (University of Wisconsin), Mahyar Samani (UC Davis), Marjan Friborz (UC Davis), Trivikram Reddy (UC Davis), Matthew D. Sinclair (University of Wisconsin; Amd Research), and Jason Lowe-Power (UC Davis)

A Case against Hardware Managed DRAM Caches for NVRAM Based Systems 194
Mark Hildebrand (University of California), Julian T. Angeles (University of California), Jason Lowe-Power (University of California), and Venkatesh Akella (University of California)

Characterizing Massively Parallel Polymorphism 205
Mengchi Zhang (Purdue University, USA), Ahmad Alawneh (Purdue University, USA), and Timothy G. Rogers (Purdue University, USA)

Poster Session B:

Pinpointing the Memory Behaviors of DNN Training 217
Jiansong Li (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Xiao Dong (Youtu Lab, Tencent, China), Guangli Li (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Peng Zhao (Huawei Technology Co., Ltd, China), Xueying Wang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Xiaobing Chen (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Xianzhi Yu (Huawei Technology Co., Ltd, China), Yongxin Yang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Zihan Jiang (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China), Wei Cao (Chinese Academy of Sciences, China), Lei Liu (Chinese Academy of Sciences, China), and Xiaobing Feng (Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China)

Thermal-Aware Overclocking for Smartphones 220
Guru Prasad Srinivasa (University at Buffalo), David Werner (Tufts University), Mark Hempstead (Tufts University), and Geoffrey Challen (University of Illinois)

The Impact of SoC Integration and OS Deployment on the Reliability of Arm Processors 223
Pablo Bodmann (Federal University of Rio Grande do Sul, Brasil), George Papadimitriou (University of Athens, Greece), Dimitris Gizopoulos (University of Athens, Greece), and Paolo Rech (Politecnico di Torino, Italy)

Memory-Efficient Hardware Performance Counters with Approximate-Counting Algorithms 226
Jingyi Xu (University of California), Sehoon Kim (University of California), Borivoje Nikolic (University of California), and Yakun Sophia Shao (University of California)

Architecture-Level Energy Estimation for Heterogeneous Computing Systems 229
Francis Wang (Massachusetts Institute of Technology), Yannan Nellie Wu (Massachusetts Institute of Technology), Matthew Woicik (Massachusetts Institute of Technology), Joel S. Emer (Massachusetts Institute of Technology), and Vivienne Sze (Massachusetts Institute of Technology)

Sparseloop: An Analytical, Energy-Focused Design Space Exploration Methodology for Sparse Tensor Accelerators 232
Yannan Nellie Wu (MIT, US), Po-An Tsai (NVIDIA, US), Angshuman Parashar (NVIDIA, US), Vivienne Sze (MIT, US), and Joel S. Emer (MIT/NVIDIA, US)

Splash-4: Improving Scalability with Lock-Free Constructs 235
Eduardo José Gómez-Hernández (University of Murcia, Spain), Ruixiang Shao (University of Murcia, Spain), Christos Sakalis (Uppsala University, Sweden), Stefanos Kaxiras (Uppsala University, Sweden), and Alberto Ros (University of Murcia, Spain)

Accelerating Fully Homomorphic Encryption through Microarchitecture-Aware Analysis and Optimization 237
Wonkyung Jung (Seoul National University), Eojin Lee (Samsung Electronics), Sangpyo Kim (Seoul National University), Namhoon Kim (Seoul National University), Keewoo Lee (Seoul National University), Chohong Min (Ewha Woman's University), Jung Hee Cheon (Seoul National University), and Jung Ho Ahn (Seoul National University)

Efficient Management of Scratch-Pad Memories in Deep Learning Accelerators 240
Subhankar Pal (University of Michigan, Ann Arbor, MI), Swagath Venkataramani (IBM TJ Watson Research Center, Yorktown Heights, NY), Viji Srinivasan (IBM TJ Watson Research Center, Yorktown Heights, NY), and Kailash Gopalakrishnan (IBM TJ Watson Research Center, Yorktown Heights, NY)

Paper Session VI: Datacenters and HPC

Hardware Acceleration for DBMS Machine Learning Scoring: Is It Worth the Overheads? 243
Zahra Azad (Boston University), Rathijit Sen (Microsoft Gray Systems Lab), Kwanghyun Park (Microsoft Gray Systems Lab), and Ajay Joshi (Boston University)

TPUPoint: Automatic Characterization of Hardware-Accelerated Machine-Learning Behavior for Cloud Computing 254
Abenezer Wudenhe (University of California, Riverside) and Hung-Wei Tseng (University of California, Riverside)

Pitfalls of InfiniBand with On-Demand Paging 265
Takuya Fukuoka (The University of Tokyo), Shigeyuki Sato (The University of Tokyo), and Kenjiro Taura (The University of Tokyo)

Analyzing the Interplay between Random Shuffling and Storage Devices for Efficient Machine Learning 276
Zhi-Lin Ke (National Taiwan University), Hsiang-Yun Cheng (Academia Sinica), Chia-Lin Yang (National Taiwan University), and Han-Wei Huang (National Taiwan University)

Paper Session VII: HW and Co-Design

E3: A HW/SW Co-Design Neuroevolution Platform for Autonomous Learning in Edge Device 288
Sheng-Chun Kao (Georgia Institute of Technology) and Tushar Krishna (Georgia Institute of Technology)

FireMarshal: Making HW/SW Co-Design Reproducible and Reliable 299
Nathan Pemberton (University of California, Berkeley) and Alon Amid (University of California, Berkeley)

COBRA: A Framework for Evaluating Compositions of Hardware Branch Predictors 310
Jerry Zhao (University of California, Berkeley), Abraham Gonzalez (University of California, Berkeley), Alon Amid (University of California, Berkeley), Sagar Karandikar (University of California, Berkeley), and Krste Asanovic (University of California, Berkeley)

Author Index 321