2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT 2021)

Atlanta, Georgia, USA
26-29 September 2021

IEEE

IEEE Catalog Number:	CFP21073-POD
ISBN:	978-1-6654-4279-4

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:	CFP21073-POD
ISBN (Print-On-Demand):	978-1-6654-4279-4
ISBN (Online):	978-1-6654-4278-7

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

2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT)

PACT 2021

Table of Contents

Message from the Chairs ix

PACT 2021 Organization xi

Conference Supporters xiv

Session 1: Tuning and Lifting

A Flexible Approach to Autotuning Multi-pass Machine Learning Compilers 1
Phitchaya Mangpo Phothilimthana (Google, USA), Amit Sabne (Google, USA), Nikhil Sarda (Google, USA), Karthik Srinivasa Murthy (Google, USA), Yanqi Zhou (Google, USA), Christof Angermueller (Google, USA), Mike Burrows (Google, USA), Sudip Roy (Google, USA), Ketan Mandke (Google, USA), Rezsa Farahani (Google, USA), Yu Emma Wang (Google, USA), Berkin Ilbeyi (Google, USA), Blake Hechtman (Google, USA), Bjarke Roune (Google, USA), Shen Wang (Google, USA), Yuanzhong Xu (Google, USA), and Samuel J. Kaufman (University of Washington, USA)

PolyGym: Polyhedral Optimizations as an Environment for Reinforcement Learning 17
Alexander Brauckmann (TU Dresden, Germany), Andrés Goens (Barkhausen Institut, Dresden, Germany), and Jeronimo Castrillon (TU Dresden, Germany)

Union: A Unified HW-SW Co-Design Ecosystem in MLIR for Evaluating Tensor Operations on Spatial Accelerators 30
Geonhwa Jeong (Georgia Institute of Technology), Gokcen Kestor (Pacific Northwest National Laboratory), Prasanth Chatarasi (IBM Research), Angshuman Parashar (NVIDIA), Po-An Tsai (NVIDIA), Sivasankaran Rajamanickam (Sandia National Laboratories), Roberto Gioiosa (Pacific Northwest National Laboratory), and Tushar Krishna (Georgia Institute of Technology)

Polygeist: Raising C to Polyhedral MLIR 45
William Moses (MIT CSAIL, USA), Lorenzo Chelini (TU Eindhoven, The Netherlands), Ruizhe Zhao (Imperial College London, UK), and Oleksandr Zinenko (Google Inc., France)

Program Lifting using Gray-Box Behavior 60
Bruce Collie (University of Edinburgh, United Kingdom) and Michael O'Boyle (University of Edinburgh, United Kingdom)

Session 2: Heterogeneous Systems

NLP-Fast: A Fast, Scalable, and Flexible System to Accelerate Large-Scale Heterogeneous NLP Models 75
Joonsung Kim (Seoul National University, Republic of Korea), Suyeon Hur (Seoul National University, Republic of Korea), Eunbok Lee (Seoul National University, Republic of Korea), Seungho Lee (Seoul National University, Republic of Korea), and Jangwoo Kim (Seoul National University, Republic of Korea)

HERTI: A Reinforcement Learning-Augmented System for Efficient Real-Time Inference on Heterogeneous Embedded Systems 90
Myeonggyun Han (UNIST) and Woongki Baek (UNIST)

X-Layer: Building Composable Pipelined Dataflows for Low-Rank Convolutions 103
Naveen Vedula (Simon Fraser University, Canada), Reza Hojabr (Simon Fraser University, Canada; University of Tehran, Iran), Ahmad Khonsari (University of Tehran, Iran; IPM, Tehran, Iran), and Arrvindh Shriraman (Simon Fraser University, Canada)

InnerSP: A Memory Efficient Sparse Matrix Multiplication Accelerator with Locality-Aware Inner Product Processing 116
Daehyeon Baek (KAIST, South Korea), Soojin Hwang (KAIST, South Korea), Taekyung Heo (KAIST, South Korea), Daehoon Kim (DGIST, South Korea), and Jaehyuk Huh (KAIST, South Korea)

Precision Batching: Bitserial Decomposition for Efficient Neural Network Inference on GPUs 129
Maximilian Lam (Harvard University, USA), Zachary Yedidia (Harvard University, USA), Colby R. Banbury (Harvard University, USA), and Vijay Janapa Reddi (Harvard University, USA)

Session 3: Characterization and Near-Memory Computing

AIBench Scenario: Scenario-Distilling AI Benchmarking 142
Wanling Gao (Institute of Computing Technology, Chinese Academy of Sciences; BenchCouncil (International Open Benchmark Council); University of Chinese Academy of Sciences), Fei Tang (Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences), Jianfeng Zhan (Institute of Computing Technology, Chinese Academy of Sciences; BenchCouncil (International Open Benchmark Council); University of Chinese Academy of Sciences), Xu Wen (Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences), Lei Wang (Institute of Computing Technology, Chinese Academy of Sciences; BenchCouncil (International Open Benchmark Council); University of Chinese Academy of Sciences), Zheng Cao (Alibaba), Chuanxin Lan (Institute of Computing Technology, Chinese Academy of Sciences), Chunjie Luo (Institute of Computing Technology, Chinese Academy of Sciences; BenchCouncil (International Open Benchmark Council); University of Chinese Academy of Sciences), Xiaoli Liu (Alibaba), and Zihan Jiang (Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences)

Google Neural Network Models for Edge Devices: Analyzing and Mitigating Machine Learning Inference Bottlenecks 159
Amirali Boroumand (Carnegie Mellon University; Stanford University), Saugata Ghose (University of Illinois Urbana-Champaign), Berkin Akin (Google), Ravi Narayanaswami (Google), Geraldo F. Oliveira (ETH Zurich), Xiaoyu Ma (Google), Eric Shiu (Google), and Onur Mutlu (ETH Zurich; Carnegie Mellon University)

SEER: A Time Prediction Model for CNNs from GPU Kernel's View 173
Guodong Liu (Institute of Computing Technology, Chinese Academy of Sciences, China), Sa Wang (Institute of Computing Technology, Chinese Academy of Sciences, China), and Yungang Bao (Institute of Computing Technology, Chinese Academy of Sciences, China)

PIM-DL: Boosting DNN Inference on Digital Processing In-Memory Architectures via Data Layout Optimizations 186
Minxuan Zhou (University of California San Diego, USA), Guoyang Chen (Alibaba Group, USA), Mohsen Imani (University of California Irvine, USA), Saransh Gupta (University of California San Diego, USA), Weifeng Zhang (Alibaba Group, USA), and Tajana Rosing (University of California San Diego, USA)

Ultra Efficient Acceleration for De Novo Genome Assembly via Near-Memory Computing 199
Minxuan Zhou (University of California, San Diego, USA), Lingxi Wu (University of Virginia, USA), Muzhou Li (University of California, San Diego, USA), Niema Moshiri (University of California, San Diego, USA), Kevin Skadron (University of Virginia, USA), and Tajana Rosing (University of California, San Diego, USA)

Session 4: Memory Hierarchy

CBP: Coordinated Management of Cache Partitioning, Bandwidth Partitioning and Prefetch Throttling 213
Nadja Ramhöj Holtryd (Chalmers University of Technology, Sweden), Madhavan Manivannan (Chalmers University of Technology, Sweden), Per Stenström (Chalmers University of Technology, Sweden), and Miquel Pericás (Chalmers University of Technology, Sweden)

Invalidate or Update? Revisiting Coherence for Tomorrow’s Cache Hierarchies 226
Mingcan Zhu (The University of Edinburgh, United Kingdom), Amna Shahab (The University of Edinburgh, United Kingdom), Antonios Katsarakis (The University of Edinburgh, United Kingdom), and Boris Grot (The University of Edinburgh, United Kingdom)

Write Prediction for Persistent Memory Systems 242
Suyash Mahar (UC San Diego, USA), Sihang Liu (University of Virginia, USA), Korakit Seemakhupt (University of Virginia, USA), Vinson Young (Microsoft, USA), and Samira Khan (University of Virginia, USA)

nuKSM: NUMA-Aware Memory De-Duplication on Multi-socket Servers 258
Akash Panda (Indian Institute of Science, India), Ashish Panwar (Indian Institute of Science, India), and Arkaprava Basu (Indian Institute of Science, India)

CoPlace: Effectively Mitigating Cache Conflicts in Modern Clouds .274......................................
Xiaowei Shang (New Jersey Institute of Technology, USA), Weiwei Jia (New Jersey Institute of Technology, USA), Xiaoning Ding (New Jersey Institute of Technology, USA), and Jianchen Shan (Hofstra University, USA)

Session 5: Graphs and Applications

Dryadic: Flexible and Fast Graph Pattern Matching at Scale .289...
Daniel Mawhirter (Colorado School of Mines; Katana Graph Inc.), Sam Reinehr (Colorado School of Mines; Katana Graph Inc.), Wei Han (Colorado School of Mines), Noah Fields (Colorado School of Mines), Miles Claver (Colorado School of Mines), Connor Holmes (Colorado School of Mines), Jedidiah McClurg (Colorado School of Mines), Tongping Liu (University of Massachusetts Amherst), and Bo Wu (Colorado School of Mines; Katana Graph Inc.)

Skywalker: Efficient Alias-Method-Based Graph Sampling and Random Walk on GPUs .304.........
Pengyu Wang (Shanghai Jiao Tong University, China), Chao Li (Shanghai Jiao Tong University, China), Jing Wang (Shanghai Jiao Tong University, China), Taolei Wang (Shanghai Jiao Tong University, China), Lu Zhang (Shanghai Jiao Tong University, China), Jingwen Leng (Shanghai Jiao Tong University, China), Quan Chen (Shanghai Jiao Tong University, China), and Minyi Guo (Shanghai Jiao Tong University, China)

SumPA: Efficient Pattern-Centric Graph Mining with Pattern Abstraction .318............................
Chuangyi Gui (Huazhong University of Science and Technology, China), Xiaofei Liao (Huazhong University of Science and Technology, China), Long Zheng (Huazhong University of Science and Technology, China), Pengcheng Yao (Huazhong University of Science and Technology, China), Qinggang Wang (Huazhong University of Science and Technology, China), and Hai Jin (Huazhong University of Science and Technology, China)

SURFNet: Super-Resolution of Turbulent Flows with Transfer Learning using Small Datasets .331..
Octavi Obiols-Sales (University of California, Irvine), Abhinav Vishnu (Advanced Micro Devices, Inc.), Nicholas P. Malaya (Adavanced Micro Devices, Inc.), and Aparna Chandramowlishwaran (University of California, Irvine)

Accelerating Fourier and Number Theoretic Transforms using Tensor Cores and Warp Shuffles .345
Sultan Durrani (University of Illinois at Urbana-Champaign), Muhammad Saad Chughtai (Georgia Institute of Technology), Mert Hidayetoglu (University of Illinois at Urbana-Champaign), Rashid Tahir (University of Prince Mugrin), Abdul Dakkak (University of Illinois at Urbana-Champaign), Lawrence Rauchwerger (University of Illinois at Urbana-Champaign), Fareed Zaffar (Lahore University of Management Sciences), and Wen-mei Hwu (University of Illinois at Urbana-Champaign)

Author Index 357. . . . .