2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP 2021)

Virtual Conference
7 – 8 July 2021

IEEE

IEEE Catalog Number: CFP21063-POD
ISBN: 978-1-6654-2702-9

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:	CFP21063-POD
ISBN (Print-On-Demand):	978-1-6654-2702-9
ISBN (Online):	978-1-6654-2701-2
ISSN:	2160-0511

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 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP)

ASAP 2021

Table of Contents

ASAP 2021 Message from the Chairs xi
ASAP 2021 Organizing Committee xii
ASAP 2021 Technical Program Committee xiii
ASAP 2021 Sponsors xv

Paper Session 1: Heterogeneous Designs and Architectures I

To Buffer, or Not to Buffer? A Case Study on FFT Accelerators for Ultra-Low-Power Multicore Clusters 1
Luca Bertaccini (ETH Zurich, Switzerland), Luca Benini (ETH Zurich and University of Bologna, Switzerland), and Francesco Conti (University of Bologna, Switzerland)

Algorithm and Hardware Co-Design for FPGA Acceleration of Hamiltonian Monte Carlo Based No-U-Turn Sampler 9
Yu Wang (University of California, USA) and Peng Li (University of California, USA)

Improving Inference Lifetime of Neuromorphic Systems via Intelligent Synapse Mapping 17
Shihao Song (Drexel University), Twisha Titirsha (Drexel University), and Anup Das (Drexel University)

A Lightweight ISE for ChaCha on RISC-V 25
Ben Marshall (University of Bristol, UK), Daniel Page (University of Bristol, UK), and Thinh Hung Pham (University of Bristol, UK)

RFC-HyPGCN: A Runtime Sparse Feature Compress Accelerator for Skeleton-Based GCNs Action Recognition Model with Hybrid Pruning 33
Dong Wen (National University of Defense Technology, China), Jingfei Jiang (National University of Defense Technology, China), Jinwei Xu (National University of Defense Technology, China), Kang Wang (National University of Defense Technology, China), Tao Xiao (National University of Defense Technology, China), Yang Zhao (National University of Defense Technology, China), and Yong Dou (National University of Defense Technology, China)

Virtual Circuit-Switching Network with Flexible Topology for High-Performance FPGA Cluster ..41
Tomohiro Ueno (RIKEN Center for Computational Science, Japan), Atsushi Koshiba (RIKEN Center for Computational Science, Japan), and Kentaro Sano (RIKEN Center for Computational Science, Japan)

Poster Session 1

Power, Performance and Area Consequences of Multi-Context Support in CGRAs ..49
Vimal Chacko (University of Toronto, Canada) and Jason Anderson (University of Toronto, Canada)

SPNC: Accelerating Sum-Product Network Inference on CPUs and GPUs ..53
Lukas Sommer (TU Darmstadt, Germany), Michael Halkenhäuser (TU Darmstadt, Germany), Cristian Axenie (Huawei Munich Research Center, Germany), and Andreas Koch (TU Darmstadt, Germany)

NEMO-CNN: An Efficient Near-Memory Accelerator for Convolutional Neural Networks ..57
Grant Brown (University of Utah, USA), Valerio Tenace (University of Utah, USA), and Pierre-Emmanuel Gaillardon (University of Utah, USA)

Edge-Disjoint Spanning Trees in the Line Graph of Hypercubes ..61
Yu Qian (Soochow University, China), Baolei Cheng (Soochow University, China), Jianxi Fan (Soochow University, China), Yifeng Wang, and Ruofan Jiang (Soochow University, China)

Customized Instruction on RISC-V for Winograd-Based Convolution Acceleration ..65
Shihang Wang (Southern University of Science and Technology, China), Jianghan Zhu (Southern University of Science and Technology, China), Qi Wang (Southern University of Science and Technology, China), Can He (Southern University of Science and Technology, China), and Terry Tao Ye (Southern University of Science and Technology, China)

Paper Session 2: Special Session on Real-Time AI

Real-Time Super-Resolution System of 4K-Video Based on Deep Learning (Invited Paper) ..69
Yanpeng Cao (Southeast University, China), Chengcheng Wang (Southeast University, China), Changjun Song (Southeast University, China), Yongming Tang (Southeast University, China), and He Li (University of Cambridge, UK)

An Efficient Real-Time Object Detection Framework on Resource-Constricted Hardware Devices via Software and Hardware Co-Design ..77
Mingshuo Liu (California State University, Fullerton), Shiyi Luo (California State University, Fullerton), Kevin Han (California State University Fullerton), Bo Yuan (Rutgers University), Ronald F. DeMara (University of Central Florida), and Yu Bai (California State University, Fullerton)

Binary Complex Neural Network Acceleration on FPGA ..85
Hongwu Peng (University of Connecticut, USA), Shanglin Zhou (University of Connecticut, USA), Scott Weitze (Stevens Institute of Technology, USA), Jiaxin Li (University of Connecticut, USA), Sahidul Islam (University of Texas at San Antonio, USA), Tong Geng (Pacific Northwest National Laboratory, USA), Ang Li (Pacific Northwest National Laboratory, USA), Wei Zhang (University of Connecticut, USA), Minghu Song (University of Connecticut, USA), Mimi Xie (University of Texas at San Antonio, USA), Hang Liu (Stevens Institute of Technology, USA), and Caiwen Ding (University of Connecticut, USA)

How to Reach Real-Time AI on Consumer Devices? Solutions for Programmable and Custom Architectures ..93
Stylianos Venieris (Samsung AI Center-Cambridge, UK), Ioannis Panopoulos (National Technical University of Athens,Greece), Ilias Leontiadis (Samsung AI Center-Cambridge, UK), and Iakovos Venieris (National Technical University of Athens, Greece)

Paper Session 3: ML Algorithms and Tools

Talos: A Weighted Speedup-Aware Device Placement of Deep Learning Models ..101
Yuanjia Xu (University of Chinese Academy of Sciences, China), Heng Wu (Institute of Software, Chinese Academy of Sciences, China), Wenbo Zhang (Institute of Software, Chinese Academy of Sciences, China), Chen Yang (University of Chinese Academy of Sciences, China), Yuewen Wu (Institute of Software, Chinese Academy of Sciences, China), Heran Gao (University of Chinese Academy of Sciences, China), and Tao Wang (Institute of Software, Chinese Academy of Sciences, China)

Hodgkin-Huxley-Based Neural Simulation with Networks Connecting to Near-Neighbor Neurons .. 109
Masashi Ogaki (Toyohashi University of Technology, Japan) and Yukinori Sato (Toyohashi University of Technology, Japan)

Accelerating Recurrent Neural Networks for Gravitational Wave Experiments ..117
Zhiqiang Que (Imperial College London, UK), Erwei Wang (Imperial College London, UK), Umar Marikar (Imperial College London, UK), Eric Moreno (California Institute of Technology, USA), Jennifer Ngadiuba (California Institute of Technology, USA), Hamza Javed (European Organization for Nuclear Research, Switzerland), Bartlomiej Borzyszkowski (European Organization for Nuclear Research, Switzerland), Thea Aarrestad (European Organization for Nuclear Research, Switzerland), Vladimir Loncar (European Organization for Nuclear Research, Switzerland), Sioni Summers (European Organization for Nuclear Research, Switzerland), Maurizio Pierini (European Organization for Nuclear Research, Switzerland), Peter Y Cheung (Imperial College London, UK), and Wayne Luk (Imperial College London, UK)

Array-Aware Neural Architecture Search ..125
Krishna Teja Chitty-Venkata (Iowa State University, USA) and Arun K. Somani (Iowa State University, USA)

TwinDNN: A Tale of Two Deep Neural Networks .133
Hyunmin Jeong (University of Illinois Urbana-Champaign, USA) and Deming Chen (University of Illinois Urbana-Champaign, USA)

Image Caption Generation Method Based on an Interaction Mechanism and Scene Concept Selection Module .141
Liping Zhang (Qilu University of Technology, China) and Qin Lu (Qilu University of Technology, China)

Panel: Coarse-Grained Reconfigurable Arrays and Their Opportunities as Application Accelerators

OpenCGRA: Democratizing Coarse-Grained Reconfigurable Arrays .149
Cheng Tan (Pacific Northwest National Laboratory), Nicolas Bohm Agostini (Pacific Northwest National Laboratory), Jeff Zhang (Harvard University), Marco Minutoli (Pacific Northwest National Laboratory), Vito Giovanni Castellana (Pacific Northwest National Laboratory), Chenhao Xie (Pacific Northwest National Laboratory), Tong Geng (Pacific Northwest National Laboratory), Ang Li (Pacific Northwest National Laboratory), Kevin Barker (Pacific Northwest National Laboratory), and Antonino Tumeo (Pacific Northwest National Laboratory)

CGRA-ME: An Open-Source Framework for CGRA Architecture and CAD Research .156
Jason Anderson (University of Toronto), Rami Beidas (University of Toronto), Vimal Chacko (University of Toronto), Hsuan Hsiao (University of Toronto), Xiaoyi Ling (University of Toronto), Omar Ragheb (University of Toronto), Xinyuan Wang (University of Toronto), and Tianyi Yu (University of Toronto)

Paper Session 4: Green Designs and Security

Number Theoretic Transform Architecture Suitable to Lattice-Based Fully-Homomorphic Encryption .163
Rogério Paludo (Universidade de Lisboa, Portugal) and Leonel Sousa (Universidade de Lisboa, Portugal)

ABACa: Access Based Allocation on Set Wise Multi-Retention in STT-RAM Last Level Cache .171
Sukarn Agarwal (IIT (BHU), India) and Shounak Chakraborty (NTNU, Norway)

DARM: A Low-Complexity and Fast Modular Multiplier for Lattice-Based Cryptography .175
Xiao Hu (Nanjing University, China), Minghao Li (Nanjing University, China), Jing Tian (Nanjing University, China), and Zhongfeng Wang (Nanjing University, China)

XDIVINSA: eXtended DIVersifying INStruction Agent to Mitigate Power Side-Channel Leakage .179
Thinh Hung Pham (University of Bristol, UK), Ben Marshall (University of Bristol, UK), Alexander Fell (Barcelona Supercomputing Center, Spain), Siew-Kei Lam (Nanyang Technological University, Singapore), and Daniel Page (University of Bristol, UK)

Memory-Aware Efficient Deep Learning Mechanism for IoT Devices 187
Jishnu Banerjee (The University of Texas at San Antonio), Sahidul Islam (The University of Texas at San Antonio), Wei Wei (The University of Texas at San Antonio), Chen Pan (Texas A&M University-Corpus Christi), Dakai Zhu (The University of Texas at San Antonio), and Mimi Xie (The University of Texas at San Antonio)

AERO: Towards Energy-Efficient Autonomous Flight in MAVs Using Approximate Execution 195
Ben Li (Jilin University, China), Jingweijia Tan (Jilin University, China), and Kaige Yan (Jilin University, China)

Poster Session 2

A Low Power Branch Prediction for Deep Learning on RISC-V Processor 203
Mingjian Sun (University of Science and Technology of China), Yuan Li (University of Science and Technology of China), Song Chen (University of Science and Technology of China), and Yi Kang (University of Science and Technology of China)

Parallel Construction of Independent Spanning Trees on Folded Crossed Cubes 207
Huanwen Zhang (Soochow University, China), Yan Wang (Soochow University, China), Jianxi Fan (Soochow University, China), and Ruyan Guo (Soochow University, China)

Paper Session 5: Special Session on Design Automation of Robust and Secure Machine Intelligence

Assessing Robustness of Hyperdimensional Computing Against Errors in Associative Memory 211
Sizhe Zhang (Villanova University), Ruixuan Wang (Villanova University), Jeff Jun Zhang (Harvard University), Abbas Rahimi (IBM Research-Zurich), and Xun Jiao (Villanova University)

Towards Automatic and Agile AI/ML Accelerator Design with End-to-End Synthesis 218
Jeff Zhang (Harvard University, USA), Nicolas Bohm Agostini (Pacific Northwest National Laboratory, USA), Shihao Song (Pacific Northwest National Laboratory, USA), Cheng Tan (Pacific Northwest National Laboratory, USA), Ankur Limaye (Pacific Northwest National Laboratory, USA), Vinay Amatya (Pacific Northwest National Laboratory, USA), Joseph Manzano (Pacific Northwest National Laboratory, USA), Marco Minutoli (Pacific Northwest National Laboratory, USA), Vito Giovanni Castellana (Pacific Northwest National Laboratory, USA), Antonino Tumeo (Pacific Northwest National Laboratory, USA), Gu-Yeon Wei (Harvard University, USA), and David Brooks (Harvard University, USA)

Paper Session 6: Heterogeneous Designs and Architectures II

ASBNN: Acceleration of Bayesian Convolutional Neural Networks by Algorithm-Hardware Co-Design 226
Yoshiki Fujiwara (The University of Tokyo, Japan) and Shinya Takamaeda-Yamazaki (The University of Tokyo, Japan)

A Novel Ring-Based Small-World NoC for Neuromorphic Processor 234
Yuchen Qiu (National University of Defense Technology, China), Chao Xiao (National University of Defense Technology, China), Linghui Peng (National University of Defense Technology, China), Junhui Wang (National University of Defense Technology, China), Ziyang Kang (National University of Defense Technology, China), Shiming Li (National University of Defense Technology, China), and Lei Wang (National University of Defense Technology, China)

Double-Pumping the Interconnect for Area Reduction in Coarse-Grained Reconfigurable Arrays 242
Xinyuan Wang (University of Toronto, Canada), Tianyi Yu (University of Toronto, Canada), Hsuan Hsiao (University of Toronto, Canada), and Jason Anderson (University of Toronto, Canada)

An Efficient Hardware Architecture for Sparse Convolution Using Linear Feedback Shift Registers 250
Murad Qasaimeh (Iowa State University, USA), Joseph Zambreno (Iowa State University, USA), and Phillip H. Jones (Iowa State University, USA)

WinoCNN: Kernel Sharing Winograd Systolic Array for Efficient Convolutional Neural Network Acceleration on FPGAs 258
Xinheng Liu (University of Illinois at Urbana-Champaign, USA), Yao Chen (Advanced Digital Sciences Center, Singapore), Cong Hao (Advanced Digital Sciences Center, Singapore), Ashutosh Dhar (University of Illinois at Urbana-Champaign, USA), and Deming Chen (University of Illinois at Urbana-Champaign, USA)

FlexACC: A Programmable Accelerator with Application-Specific ISA for Flexible Deep Neural Network Inference 266
En-Yu Yang (Harvard University), Tianyu Jia (Harvard University), David Brooks (Harvard University), and Gu-Yeon Wei (Harvard University)

Author Index 275