

# **2024 33rd International Conference on Parallel Architectures and Compilation Techniques (PACT 2024)**

**Long Beach, California, USA  
13 – 16 October 2024**



**IEEE Catalog Number: CFP24073-POD  
ISBN: 979-8-3315-3398-4**

**Copyright © 2024, ACM  
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:	CFP24073-POD
ISBN (Print-On-Demand):	979-8-3315-3398-4
ISBN (Online):	979-8-4007-0631-8

**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](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

CURRAN ASSOCIATES INC.  
**proceedings**  
.com

# Contents

<b>PipeGen: Automated Transformation of a Single-Core Pipeline into a Multicore Pipeline for a Given Memory Consistency Model .....</b>	<b>1</b>
An Qi Zhang ( <i>University of Utah</i> ); Andrés Goens ( <i>University of Amsterdam</i> ); Nicolai Oswald ( <i>Nvidia</i> ); Tobias Grosser ( <i>University of Cambridge</i> ); Daniel Sorin ( <i>Duke University</i> ); Vijay Nagarajan ( <i>University of Utah</i> )	
<b>vSPACE: Supporting Parallel Network Packet Processing in Virtualized Environments through Dynamic Core Management.....</b>	<b>14</b>
Gyeongseo Park ( <i>DGIST/ETRI</i> ), Minhoo Kim ( <i>DGIST</i> ); Ki-Dong Kang ( <i>ETRI</i> ); Yunhyeong Jeon, Sungju Kim, Hyosang Kim ( <i>DGIST</i> ); Daehoon Kim ( <i>Yonsei University</i> )	
<b>MORSE: <u>M</u>emory <u>O</u>verwrite Time Guided <u>S</u>oft Writes to Improve ReRAM Energy and Endurance .....</b>	<b>26</b>
Devesh Singh, Donald Yeung ( <i>University of Maryland</i> )	
<b>Optimizing Tensor Computation Graphs with Equality Saturation and Monte Carlo Tree Search .....</b>	<b>40</b>
Jakob Hartmann, Guoliang He, Eiko Yoneki ( <i>University of Cambridge</i> )	
<b>Toast: A Heterogeneous Memory Management System .....</b>	<b>53</b>
Maurice Bailleu ( <i>Huawei Research</i> ); Dimitrios Stavrakakis ( <i>TU Munich / The University of Edinburgh</i> ); Rodrigo Rocha ( <i>Huawei Research</i> ); Soham Chakraborty ( <i>TU Delft</i> ); Deepak Garg ( <i>Max Planck Institute for Software Systems (MPI-SWS)</i> ); Pramod Bhatotia ( <i>TU Munich / The University of Edinburgh</i> )	
<b>A Transducers-based Programming Framework for Efficient Data Transformation.....</b>	<b>66</b>
Tri Nguyen, Michela Becchi ( <i>North Carolina State University</i> )	
<b>Activation Sequence Caching: High-Throughput and Memory-Efficient Generative Inference with a Single GPU .....</b>	<b>78</b>
Sowoong Kim, Eunyong Sim ( <i>UNIST</i> ); Youngsam Shin, YeonGon Cho ( <i>Samsung Advanced Institute of Technology</i> ); Woongki Baek ( <i>UNIST</i> )	
<b>GraNNDis: Fast Distributed Graph Neural Network Training Framework for Multi-Server Clusters ....</b>	<b>91</b>
Jaeyong Song, Hongsun Jang, Hunseong Lim, Jaewon Jung ( <i>Seoul National University</i> ); Youngsok Kim ( <i>Yonsei University</i> ); Jinho Lee ( <i>Seoul National University</i> )	
<b>Trimma: Trimming Metadata Storage and Latency for Hybrid Memory Systems.....</b>	<b>108</b>
Yiwei Li, Boyu Tian ( <i>Tsinghua University</i> ); Mingyu Gao ( <i>Tsinghua University / Shanghai Qi Zhi Institute</i> )	
<b>BoostCom: Towards Efficient Universal Fully Homomorphic Encryption by Boosting the Word-wise Comparisons .....</b>	<b>121</b>
Ardhi Wiratama Baskara Yudha ( <i>University of Central Florida / Advanced Micro Devices, Inc.</i> ); Jiaqi Xue, Qian Lou ( <i>University of Central Florida</i> ); Huiyang Zhou ( <i>North Carolina State University</i> ); Yan Solihin ( <i>University of Central Florida</i> )	
<b>Leveraging Difference Recurrence Relations for High-Performance GPU Genome Alignment.....</b>	<b>133</b>
Alberto Zeni ( <i>Politecnico di Milano / NVIDIA Corporation</i> ); Seth Onken ( <i>NVIDIA Corporation</i> ); Marco Domenico Santambrogio ( <i>Politecnico di Milano</i> ); Mehrzad Samadi ( <i>NVIDIA Corporation</i> )	

<b>Chimera: Leveraging Hybrid Offsets for Efficient Data Prefetching .....</b>	<b>144</b>
Shuiyi He, Zicong Wang, Xuan Tang, Qiyao Sun, Dezun Dong ( <i>National University of Defense Technology</i> )	
<b>MIREncoder: Multi-modal IR-based Pretrained Embeddings for Performance Optimizations.....</b>	<b>156</b>
Akash Dutta, Ali Jannesari ( <i>Iowa State University</i> )	
<b>NavCim: Comprehensive Design Space Exploration for Analog Computing-in-Memory Architectures .....</b>	<b>168</b>
Juseong Park, Boseok Kim ( <i>Pohang University of Science and Technology</i> ); Hyojin Sung ( <i>Seoul National University</i> )	
<b>Mozart: Taming Taxes and Composing Accelerators with Shared-Memory.....</b>	<b>183</b>
Vignesh Suresh, Bakshree Mishra, Ying Jing, Zeran Zhu, Naiyin Jin, Charles Block ( <i>University of Illinois at Urbana-Champaign</i> ); Paolo Mantovani, Davide Giri, Joseph Zuckerman, Luca P. Carloni ( <i>Columbia University</i> ); Sarita V. Adve ( <i>University of Illinois at Urbana-Champaign</i> )	
<b>PIM-Opt: Demystifying Distributed Optimization Algorithms on a Real-World Processing-In-Memory System .....</b>	<b>201</b>
Steve Rhyner, Haocong Luo ( <i>ETH Zurich</i> ); Juan Gómez-Luna ( <i>NVIDIA</i> ); Mohammad Sadrosadati ( <i>ETH Zurich</i> ); Jiawei Jiang ( <i>Wuhan University</i> ); Ataberk Olgun, Harshita Gupta ( <i>ETH Zurich</i> ); Ce Zhang ( <i>University of Chicago</i> ); Onur Mutlu ( <i>ETH Zurich</i> )	
<b>Parallel Loop Locality Analysis for Symbolic Thread Counts.....</b>	<b>219</b>
Fangzhou Liu, Yifan Zhu, Shaotong Sun, Chen Ding, Wesley Smith, Kaave Seyed Hosseini ( <i>University of Rochester</i> )	
<b>Improving Throughput-oriented LLM Inference with CPU Computations .....</b>	<b>233</b>
Daon Park, Bernhard Egger ( <i>Seoul National University</i> )	
<b>ZeD: A Generalized Accelerator for Variably Sparse Matrix Computations in ML.....</b>	<b>246</b>
Pranav Dangi, Zhenyu Bai, Rohan Juneja, Dhananjaya Wijerathne, Tulika Mitra ( <i>National University of Singapore</i> )	
<b>ACE: Efficient GPU Kernel Concurrency for Input-Dependent Irregular Computational Graphs.....</b>	<b>258</b>
Sankeerth Durvasula, Adrian Zhao, Raymond Kiguru, Yushi Guan, Zhonghan Chen, Nandita Vijaykumar ( <i>University of Toronto</i> )	
<b>SZKP: A Scalable Accelerator Architecture for Zero-Knowledge Proofs .....</b>	<b>271</b>
Alhad Daftardar, Brandon Reagen, Siddharth Garg ( <i>New York University</i> )	
<b>BOOM: Use your Desktop to Accurately Predict the Performance of Large Deep Neural Networks .....</b>	<b>284</b>
Qidong Su ( <i>University of Toronto / Vector Institute / CentML</i> ); Jiacheng Yang ( <i>University of Toronto / Vector Institute</i> ); Gennady Pekhimenko ( <i>University of Toronto / Vector Institute / CentML</i> )	
<b>A Parallel Hash Table for Streaming Applications .....</b>	<b>297</b>
Magnus Östgren, Ioannis Sourdis ( <i>Chalmers University of Technology</i> )	
<b>Recompiling QAOA Circuits on Various Rotational Directions .....</b>	<b>309</b>
Enhyeok Jang, Dongho Ha, Seungwoo Choi, Youngmin Kim, Jaewon Kwon, Yongju Lee, Sungwoo Ahn, Hyungseok Kim, Won Woo Ro ( <i>Yonsei University</i> )	

<b>Rethinking Page Table Structure for Fast Address Translation in GPUs: A Fixed-Size Hashed Page Table .....</b>	<b>325</b>
Sunghbin Jang, Junhyeok Park, Osang Kwon, Yongho Lee, Seokin Hong ( <i>Sungkyunkwan University</i> )	
<b>FriendlyFoe: Adversarial Machine Learning as a Practical Architectural Defense against Side Channel Attacks .....</b>	<b>338</b>
Hyoungwook Nam ( <i>University of Illinois at Urbana-Champaign</i> ); Raghavendra Pradyumna Pothukuchi ( <i>Yale University</i> ); Bo Li, Nam Sung Kim, Josep Torrellas ( <i>University of Illinois at Urbana-Champaign</i> )	
<b>Faster and More Reliable Quantum SWAPs via Native Gates.....</b>	<b>351</b>
Pranav Gokhale, Teague Tomesh ( <i>Inflection</i> ); Martin Suchara ( <i>Microsoft</i> ); Fred Chong ( <i>University of Chicago</i> )	