

2023 IEEE 30th International Conference on High Performance Computing, Data, and Analytics (HiPC 2023)

**Goa, India
18 – 21 December 2023**



**IEEE Catalog Number: CFP23176-POD
ISBN: 979-8-3503-8323-2**

**Copyright © 2023 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:	CFP23176-POD
ISBN (Print-On-Demand):	979-8-3503-8323-2
ISBN (Online):	979-8-3503-8322-5
ISSN:	1094-7256

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

2023 IEEE 30th International Conference on High Performance Computing, Data, and Analytics (HiPC) **HiPC 2023**

Table of Contents

Message from the HiPC 2023 General Co-chairs	xi
Message from the HiPC 2023 Program Chairs	xiii
HiPC 2023 Organization	xv
HiPC 2023 Steering Committee	xvii
HiPC 2023 Technical Program Committee	xviii
Keynote 2: Sunita Sarawagi	xxii
Keynote 3: Priyanka Sharma	xxiii
Keynote 4: Manish Parashar	xxiv
Keynote 5: Vittal Setty	xxv

Technical Session 1: High Performance Computing – Architecture

DNA-TEQ: An Adaptive Exponential Quantization of Tensors for DNN Inference	1
<i>Bahareh Khabbazan (Universitat Politècnica de Catalunya (UPC), Spain), Marc Riera Villanueva (Universitat Politècnica de Catalunya (UPC), Spain), and Antonio González (Universitat Politècnica de Catalunya (UPC), Spain)</i>	
PARAG: PIM Architecture for Real-Time Acceleration of GCNs	11
<i>Gian Singh (Arizona State University), Sanmukh R. Kuppannagari (Case Western Reserve University), and Sarma Vrudhula (Arizona State University)</i>	
Hybrid CUDA Unified Memory Management in Fully Homomorphic Encryption Workloads	21
<i>Jake Choi (Seoul National University, South Korea), Jaejin Lee (CryptoLab, South Korea), Sunchul Jung (CryptoLab, South Korea), and Heonyoung Yeom (Seoul National University, South Korea)</i>	
Mobile Gaming Experience: An Approach Based on Thread Scheduler & Thread Priority Manager .	31
<i>Jani Basha Shaik (Samsung R&D Institute India, Bangalore), Sandani Shaik (Samsung R&D Institute India, Bangalore), Nazrinbanu Nagori (Samsung R&D Institute India, Bangalore), and Veerendra Shetty (Samsung R&D Institute India, Bangalore)</i>	

Optimized All-to-all Connection Establishment for High-Performance MPI Libraries Over InfiniBand	41
<i>Shulei Xu (The Ohio State University, USA), Goutham Kalikrishna Reddy Kuncham Kuncham (The Ohio State University, USA), Mustafa Abduljabbar (The Ohio State University, USA), Hari Subramoni (The Ohio State University, USA), and Dhabaleswar K. Panda (The Ohio State University, USA)</i>	
MOSAIC: A Multi-objective Optimization Framework for Sustainable Datacenter Management	51
<i>Sirui Qi (Colorado State University, USA), Dejan Milojevic (Hewlett Packard Labs, USA), Cullen Bash (Hewlett Packard Labs, USA), and Sudeep Pasricha (Colorado State University, USA)</i>	
A 118 GOPS/mm ² 3D eDRAM TensorCore Architecture for Large-Scale Matrix Multiplication	61
<i>Mengtian Yang (University of Texas at Austin), Yipeng Wang (University of Texas at Austin), and Jaydeep P. Kulkarni (University of Texas at Austin)</i>	

Technical Session 2: Data Science – Scalable Algorithms and Analytics

Contour Algorithm for Connectivity	66
<i>Zhihui Du (New Jersey Institute of Technology, USA), Oliver Alvarado Rodriguez (New Jersey Institute of Technology, USA), Fuhuan Li (New Jersey Institute of Technology, USA), Mohammad Dindoost (New Jersey Institute of Technology, USA), and David A. Bader (New Jersey Institute of Technology, USA)</i>	
CAPTURE: Memory-Centric Partitioning for Distributed DNN Training with Hybrid Parallelism ...	76
<i>Henk Dreuning (University of Amsterdam, The Netherlands), Kees Verstoep (Vrije Universiteit Amsterdam, The Netherlands), Henri E. Bal (Vrije Universiteit Amsterdam, The Netherlands), and Rob V. van Nieuwpoort (Leiden University, The Netherlands)</i>	
MiCRO: Near-Zero Cost Gradient Sparsification for Scaling and Accelerating Distributed DNN Training	87
<i>Daegun Yoon (Ajou University) and Sangyoon Oh (Ajou University)</i>	
Understanding Patterns of Deep Learning Model Evolution in Network Architecture Search	97
<i>Robert Underwood (Argonne National Laboratory, USA), Meghana Madhyastha (Johns Hopkins University, USA), Randal Burns (Johns Hopkins University, USA), and Bogdan Nicolae (Argonne National Laboratory, USA)</i>	
Flover: A Temporal Fusion Framework for Efficient Autoregressive Model Parallel Inference	107
<i>Jinghan Yao (The Ohio State University, USA), Nawras Alnaasan (The Ohio State University, USA), Tian Chen (The Ohio State University, USA), Aamir Shafi (The Ohio State University, USA), Hari Subramoni (The Ohio State University, USA), and Dhabaleswar K. Panda (The Ohio State University, USA)</i>	
Characterization and Detection of Artifacts for Error-Controlled Lossy Compressors	117
<i>Pu Jiao (University of Kentucky, USA), Sheng Di (Argonne National Laboratory, USA), Jingyang Liu (University of California, Riverside, USA), Xin Liang (University of Kentucky, USA), and Franck Cappello (Argonne National Laboratory, USA)</i>	

Performance Characterization of Containerized DNN Training and Inference on Edge Accelerators	127
<i>Prashanthi S.K. (Indian Institute of Science, India), Vinayaka Hegde (Indian Institute of Science, India), Keerthana Patchava (Indian Institute of Science, India), Ankita Das (Indian Institute of Science, India), and Yogesh Simmhan (Indian Institute of Science, India)</i>	

Technical Session 3: Data Science – Scalable Systems and Software

SECRE: Surrogate-Based Error-Controlled Lossy Compression Ratio Estimation Framework	132
<i>Arham Khan (University of Chicago, USA), Sheng Di (Argonne National Laboratory, USA), Kai Zhao (Florida State University, USA), Jinyang Liu (University of California, USA), Kyle Chard (University of Chicago, USA), Ian Foster (University of Chicago, USA), and Franck Cappello (Argonne National Laboratory, USA; University of Illinois Urbana-Champaign, USA)</i>	
Fast Algorithms for Scientific Data Compression	143
<i>Tania Banerjee (University of Florida, USA), Jaemoon Lee (University of Florida, USA), Jong Choi (Oak Ridge National Laboratory, USA), Qian Gong (Oak Ridge National Laboratory, USA), Jieyang Chen (University of Alabama at Birmingham, USA), Scott Klasky (Oak Ridge National Laboratory, USA), Anand Rangarajan (University of Florida, USA), and Sanjay Ranka (University of Florida, USA)</i>	
CAPIO: a Middleware for Transparent I/O Streaming in Data-Intensive Workflows	153
<i>Alberto Riccardo Martinelli (University of Turin, Italy), Massimo Torquati (University of Pisa, Italy), Marco Aldinucci (University of Turin, Italy), Iacopo Colonnelli (University of Turin, Italy), and Barbara Cantalupo (University of Turin, Italy)</i>	
JASS: A Tunable Checkpointing System for NVM-Based Systems	164
<i>Akshin Singh (IIT Delhi, India) and Smruti R. Sarangi (IIT Delhi, India)</i>	
Multi-streamed Metadata-Integrity Verification for Cloud Migration In Deduplication Systems	174
<i>Shashank Khobragade (Dell Technologies), Santi Gopal Mondal (Dell Technologies), and Kalyan Gunda (Dell Technologies)</i>	
CPU-GPU Tuning to Improve Modern Scientific Applications run on Heterogeneous Nodes	179
<i>Mathialakan Thavappiragasam (Argonne National Laboratory, USA) and Vivek Kale (Sandia National Laboratory, USA)</i>	
DDIOSim: A Microarchitecture Simulator for Data Direct I/O Technology	184
<i>Hari Sharan (Indian Institute of Technology Bombay, India), Mythili Vutukuru (Indian Institute of Technology Bombay, India), and Biswabandan Panda (Indian Institute of Technology Bombay, India)</i>	
FPGA Accelerated Bi-Cubic Convolution for Image Interpolation	189
<i>Ankit Choudhary (National Remote Sensing Centre, India), S. K. Vaibhav Kodavati (National Remote Sensing Centre, India), Mythili B. (National Remote Sensing Centre, India), Anjaneyulu R. V. G. (National Remote Sensing Centre, India), and M. Manju Sarma (National Remote Sensing Centre, India)</i>	

Technical Session 4: Best Paper Nominees

DeltaSPARSE: High-Performance Sparse General Matrix-Matrix Multiplication on Multi-GPU Systems	194
<i>Shuai Yang (University of Chinese Academy of Sciences, China), Changyou Zhang (University of Chinese Academy of Sciences, China), and Ji Ma (University of Chinese Academy of Sciences, China)</i>	
Strategies for Fast I/O Throughput in Large-Scale Climate Modeling Applications	203
<i>Koushik Sen (Qualcomm, India), Sathish Vadhiyar (Indian Institute of Science, India), and Vinayachandran PN (Indian Institute of Science, India)</i>	
ME-ViT: A Single-Load Memory-Efficient FPGA Accelerator for Vision Transformers	213
<i>Kyle Marino (University of Southern California, USA), Pengmiao Zhang (University of Southern California, USA), and Viktor K. Prasanna (University of Southern California, USA)</i>	
Graph Pattern Mining Paradigms: Consolidation and Renewed Bearing	224
<i>Vinicius Dias (Universidade Federal de Lavras, Brazil), Samuel Ferraz (Universidade Federal de Mato Grosso do Sul, Brazil; Universidade Federal de Minas Gerais (UFMG), Brazil), Aditya Vadlamani (The Ohio State University, USA), Mahdi Erfanian (The Ohio State University, USA), Carlos H. C. Teixeira (Universidade Federal de Minas Gerais, Brazil), Dorgival Guedes (Universidade Federal de Minas Gerais, Brazil), Wagner Meira Jr (Universidade Federal de Minas Gerais, Brazil), and Srinivasan Parthasarathy (The Ohio State University, USA)</i>	
Accelerating Time to Science Using CRADLE: A Framework for Materials Data Science	234
<i>Arafath Nihar (Case Western Reserve University, USA), Thomas G. Ciardi (Case Western Reserve University, USA), Rounak Chawla (Case Western Reserve University, USA), Olatunde Akanbi (Case Western Reserve University, USA), Vipin Chaudhary (Case Western Reserve University, USA), Yinghui Wu (Case Western Reserve University, USA), and Roger H. French (Case Western Reserve University, USA)</i>	
Optimizing the Training of Co-Located Deep Learning Models Using Cache-Aware Staggering	246
<i>Kevin Assogba (Rochester Institute of Technology), Bogdan Nicolae (Argonne National Laboratory), and M. Mustafa Rafique (Rochester Institute of Technology)</i>	

Technical Session 5: High Performance Computing – Systems

Towards Efficient I/O Pipelines Using Accumulated Compression	256
<i>Avinash Maurya (Rochester Institute of Technology, USA), Bogdan Nicolae (Argonne National Laboratory, USA), M. Mustafa Rafique (Rochester Institute of Technology, USA), and Franck Cappello (Argonne National Laboratory, USA)</i>	

Oikonomos-II: A Reinforcement-Learning, Resource-Recommendation System for Cloud HPC	266
<i>J.L.F. Betting (Erasmus Medical Center, The Netherlands), C.I. De Zeeuw (Erasmus Medical Center, The Netherlands; Netherlands Institute for Neuroscience, The Netherlands), and C. Strydis (Erasmus Medical Center, The Netherlands; Delft University of Technology, The Netherlands)</i>	
SCoOL – Scalable Common Optimization Library	277
<i>Zainul Abideen Sayed (University at Buffalo) and Jaroslaw Zola (University at Buffalo)</i>	
Data Locality Aware Computation Offloading in Near Memory Processing Architecture for Big Data Applications	288
<i>Satanu Maity (Indian Institute of Information Technology Guwahati, India), Mayank Goel (Indian Institute of Information Technology Guwahati, India), and Manojit Ghose (Indian Institute of Information Technology Guwahati, India)</i>	
Benesh: A Framework for Choreographic Coordination of In Situ Workflows	298
<i>Philip E. Davis (University of Utah, USA), Jacob Merson (Rensselaer Polytechnic Institute, USA), Pradeep Subedi (Samsung Semiconductor Inc., USA), Lee Ricketson (Lawrence Livermore National Laboratory, USA), Cameron W. Smith (Rensselaer Polytechnic Institute, USA), Mark S. Shephard (Rensselaer Polytechnic Institute, USA), and Manish Parashar (University of Utah, USA)</i>	
Profit Maximization Using Collaborative Storage Management in Multi-tier Edge-Cloud System...	309
<i>Shubhradeep Roy (Indian Institute of Technology Guwahati, India), Suvarthi Sarkar (Indian Institute of Technology Guwahati, India), and Aryabartta Sahu (Indian Institute of Technology Guwahati, India)</i>	
Towards Enhanced I/O Performance of NVM File Systems	319
<i>Jiwoo Bang (Seoul National University, Korea), Chungyong Kim (Seoul National University, Korea), Eun-Kyu Byun (Korea Institute of Science and Technology Information, Korea), Hanul Sung (Sangmyung University, Korea), Jaehwan Lee (Korea Aerospace University, Korea), and Hyeonsang Eom (Seoul National University, Korea)</i>	

Technical Session 6: High Performance Computing – Algorithms and Applications

Fast Parallel Tensor Times Same Vector for Hypergraphs	324
<i>Shruti Shivakumar (Georgia Institute of Technology, USA), Ilya Amburg (Pacific Northwest National Laboratory, USA), Sinan G. Aksoy (Pacific Northwest National Laboratory, USA), Jiajia Li (North Carolina State University, USA), Stephen J. Young (Pacific Northwest National Laboratory, USA), and Srinivas Aluru (Georgia Institute of Technology, USA)</i>	
Reduce, Reuse and Adapt: Accelerating Graph Processing on GPUs	335
<i>Ullas A (Indian Institute of Science, India), Rupesh Nasre (Indian Institute of Science, India), and R Govindarajan (Indian Institute of Science, India)</i>	

Reduce Computational Complexity for Convolutional Layers by Skipping Zeros	347
<i>Zhiyi Zhang (University of Science and Technology of China, China), Pengfei Zhang (University of Science and Technology of China, China), Zhuopin Xu (University of Science and Technology of China, China), and Qi Wang (University of Science and Technology of China, China)</i>	
SpikeNC: An Accurate and Scalable Simulator for Spiking Neural Network on Multi-core Neuromorphic Hardware	357
<i>Lisheng Xie (Shanghai Jiao Tong University, China), Jianwei Xue (Shanghai Jiao Tong University, China), Liangshun Wu (Shanghai Jiao Tong University, China), Faquan Chen (Shanghai Jiao Tong University, China), Qingyang Tian (Shanghai Jiao Tong University, China), Yifan Zhou (Shanghai Jiao Tong University, China), Rendong Ying (Shanghai Jiao Tong University, China), and Peilin Liu (Shanghai Jiao Tong University, China)</i>	
DAGit: A Platform For Enabling Serverless Applications	367
<i>Anubhav Jana (Indian Institute of Technology Bombay, India), Purushottam Kulkarni (Indian Institute of Technology Bombay, India), and Umesh Bellur (Indian Institute of Technology Bombay, India)</i>	
Efficient GPU Implementation of Automatic Differentiation for Computational Fluid Dynamics....	377
<i>Mohammad Zubair (Old Dominion University, USA), Desh Ranjan (Old Dominion University, USA), Aaron Walden (NASA Langley Research Center, USA), Gabriel Nastac (NASA Langley Research Center, USA), Eric Nielsen (NASA Langley Research Center, USA), Boris Diskin (National Institute of Aerospace, USA), Marc Paterno (Fermi National Accelerator Laboratory, USA), Samuel Jung (Northwestern University, USA), and Joshua Hoke Davis (University of Maryland, USA)</i>	
A Lossless Compression Pipeline for Petabyte-Scale Whole Genome Sequencing Data	387
<i>Ajeya Bhat (Indian Institute of Science, India), Sai Manasa Chadalavada (Indian Institute of Science, India), Nagakishore Jammula (Synopsys Inc., USA), Chirag Jain (Indian Institute of Science, India), and Yogesh Simmhan (Indian Institute of Science, India)</i>	
Author Index	393