

2026 IEEE International Parallel and Distributed Processing Symposium (IPDPS 2026)

**New Orleans, Louisiana, USA
25-29 May 2026**

Pages 1–706

IEEE Catalog Number: CFP26023-POD
ISBN: 979-8-3195-0603-0



Copyright © 2026, IEEE

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:	CFP26023-POD
ISBN (Print-On-Demand):	979-8-3195-0603-0
ISBN (Online):	979-8-3195-0602-3
ISSN:	1530-2075

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

TABLE OF CONTENTS

TC_SpGEMM: High Performance Sparse General Matrix Multiplication with Tensor Core-Accelerated	1
<i>FuKai Sun, Xing Cong, Chenhao Xie, Rui Wang, Zhongzhi Luan, Yi Liu, Depei Qian</i>	
STTID: High-Performance Sparse Tensor-Train Interpolative Decomposition	15
<i>Zhaonan Meng, Miles Stoudenmire, Karl Pierce, Frank Mueller, Jiajia Li</i>	
DFA-SpTRSV: A Depth-First Asynchronous Algorithm for Sparse Triangular Solver	32
<i>Zhimeng Han, Chuanfu Xu, Haozhong Qiu, Sen Yan, Yue Ding, Qingsong Wang, Yue Wang, Wenjing Zhang</i>	
<i>Ying Jin, Yonggang Che</i>	
cuHPX: GPU-Accelerated Differentiable Spherical Harmonic Transforms on HEALPix Grids	45
<i>Xiaopo Cheng, Akshay Subramaniam, Shixun Wu, Noah Brenowitz</i>	
GPU Algorithms for Biconnected Components on Large Graphs	59
<i>Abhijeet Sahu, Andaluri S P V M Aditya, G Ramakrishna, Kishore Kothapalli, Dip Sankar Banerjee</i>	
Parallel Louvain Algorithms with Convergence Guarantee	73
<i>Jason Niu, M. Yusuf Özkaya, Ahmet Erdem Sartyüce, Ümit V. Çatalyürek</i>	
GPU-Accelerated Multilevel Graph Clustering: A Parallel Perspective on Louvain and Leiden	87
<i>Michael S. Gilbert, Kamesh Madduri</i>	
UVVs: Identifying Unchanged Vertex Values in Evolving Graphs via Intersection-Union Analysis	100
<i>Mahbod Afarin, Chao Gao, Xizhe Yin, Zhijia Zhao, Nael Abu-Ghazaleh, Rajiv Gupta</i>	
SLEEK: Compressing Memory Copies for Floating-Point Data on GPUs	115
<i>Anju Mongandampulath Akathoott, Andrew Rodriguez, Martin Burtscher</i>	
FFCz: Fast Fourier Correction for Spectrum-Preserving Lossy Compression of Scientific Data	130
<i>Congrong Ren, Robert Underwood, Sheng Di, Emre Can Kutay, Zarija Lukić, Aylin Yener, Franck Cappello</i>	
<i>Hanqi Guo</i>	
Mitigating Artifacts in Pre-quantization Based Scientific Data Compressors with Quantization-aware Interpolation	144
<i>Pu Jiao, Sheng Di, Jiannan Tian, Mingze Xia, Xuan Wu, Yang Zhang, Xin Liang, Franck Cappello</i>	
pMSz: A Distributed Parallel Algorithm for Correcting Extrema and Morse-Smale Segmentations in Lossy Compression	159
<i>Yuxiao Li, Mingze Xia, Xin Liang, Bei Wang, Robert Underwood, Sheng Di, Hemant Sharma</i>	
<i>Dishant Beniwal, Franck Cappello, Hanqi Guo</i>	
Parallel Integer and Learning-Augmented Sorting Algorithms in the Binary-Forking Model	173
<i>Michael T. Goodrich, Yan Gu</i>	
Evaluation of Dynamic Vector Bin Packing for Virtual Machine Placement	184
<i>Zong Yu Lee, Xueyan Tang</i>	
I Like To Move It – Computation Instead of Data in the Brain	198
<i>Fabian Czappa, Marvin Kaster, Felix Wolf</i>	

Organic Mergesort and Finger Buffer-Tree Sort: Adaptive Sorting Algorithms with External-Memory or Parallel Implementations	212
<i>Gerth Stølting Brodal, Michael T. Goodrich, Ryuto Kitagawa, Nodari Sitchinava, Rolf Svenning</i>	
Communication-Avoiding Linear Algebraic Kernel K-Means on GPUs	223
<i>Julian Bellavita, Matthew Rubino, Nakul Iyer, Andrew Chang, Aditya Devarakonda, Flavio Vella Giulia Guidi</i>	
Synchronized Dual-Ring: A Synergistic Algorithm for Bandwidth-Efficient Collective Communication Leveraging Die-to-Die Direct Interconnects	238
<i>Dongxiang Zhang, Erkun Zhang, Yijia Zhang, Shixun Zhang, Bingqiang Wang, FangJiong Chen</i>	
GPU-Accelerated Approximate Nearest Neighbor Search via PCA-Augmented Graph Indexing for Vector Databases	251
<i>Yuanpeng Wang, MohammadReza HoseinyFarahabady, Albert Y. Zomaya</i>	
MetaPaper: Massively Scaling Minecraft Worlds With Dynamic Load Balancing	264
<i>Marina López-Alet, Stepan Klymonchuk, Daniel Barcelona-Pons, Pedro García-López</i>	
Scalable Spike Transmission in Large-Scale Brain Network Simulations	278
<i>Mario Ibáñez, Marvin Kaster, Borja Pérez, Han Lu, Fabian Czappa, Sandra Díaz Pier, Jose Luis Bosque Felix Wolf, Thorsten Hater</i>	
Looking for (Genomic) Needles in a Haystack: Sparsity-Driven Search for Identifying Correlated Genetic Mutations in Cancer	293
<i>Ritvik Prabhu, Emil Vatai, Bernard Moussad, Emmanuel Jeannot, Ramu Anandakrishnan, Wu-Chun Feng Mohamed Wahib</i>	
Fast Algorithms for Scheduling Many-body Correlation Functions on Accelerators	307
<i>Oguz Selvitopi, Emin Ozturk, Jie Chen, Ponnuswamy Sadayappan, Robert G. Edwards, Aydın Buluç</i>	
VoLLM: Smoothness-aware Serving of LLM-powered Voice Q&A via Adaptive Preemption	321
<i>Xiang Wang, Yuanchun Li, Ju Ren, Lichen Pang, Shansong Yang, Yunxin Liu</i>	
X4-MATCH: Sustainable Prediction-based Distribution of Video Encoding on Cloud and Edge	334
<i>Samira Afzal, Narges Mehran, Andrew C. Freeman, Manuel Hoi, Armin Lachini, Christian Timmerer Radu Prodan</i>	
Pixel-Resolved Long-Context Learning for Turbulence at Exascale: Resolving Small-scale Eddies Toward the Viscous Limit	347
<i>Junqi Yin, Mijanur Palash, M. Paul Laiu, Muralikrishnan Gopalakrishnan Meena, John Gounley Stephen M. de Bruyn Kops, Feiyi Wang, Ramanan Sankaran, Pei Zhang</i>	
High-Order-Preserving Acceleration of a Shallow-Water Dynamical Core Using Tensor Units	363
<i>Jienan Yao, Lilong Zhou, Heng Li, Wei Xue</i>	
ESCHER: Efficient and Scalable Hypergraph Evolution Representation with Application to Triad Counting	377
<i>S. M. Shovan, Arindam Khanda, Sanjukta Bhowmick, Sajal K. Das</i>	
GPU-Native Compressed Neighbor Lists with a Space-Filling-Curve Data Layout	390
<i>Felix Thaler, Sebastian Keller</i>	
DSTree: Data-Driven Synchronous Traversals for Decision Forests on GPUs	401
<i>Bi Zeng, Zhenlin Wu, Hongyuan Liu, Xinyu Chen</i>	

Fast Entropy Decoding for Sparse MVM on GPUs	416
<i>Emil Schätzle, Tommaso Pegolotti, Markus Püschel</i>	
COMPAT: A Compression Framework with Fine-Grain and Object-Level Compression Units	433
<i>Qi Shao, Per Stenström</i>	
I/O-Aware PIM Acceleration for Long-Sequence LLM Inference with Hybrid Sparse Attention	445
<i>Xiaoyang Lu, Lihan Hu, Hongrui Huang, Peng Jiang, Xian-He Sun</i>	
FitCache: A Transparent Drop-In Framework for Multi-Tier Caching to Accelerate Distributed Deep Learning Workloads	458
<i>Guangxing Hu, Awais Khan, Christopher Zimmer, Michael J. Brim, Frank Mueller</i>	
Accelerating Sparse-Sparse-Dense Graph Matrix Operator using Asynchronous Execution in GPUs	473
<i>Hanan Khan, Omer Khan</i>	
FastTT: Accelerating Shift-XOR Erasure Coding for Data Storage	485
<i>Duo Sun, Ximing Fu, Qiang Wang</i>	
SciWIND: Effectively Exploiting Node-Local Storage for Data-Intensive High-Energy Physics Workflows	498
<i>Jin Zhou, Colin Thomas, Barry Sly-Delgado, Connor Moore, Benjamin Tovar, Kevin Lannon Douglas Thain</i>	
FlowKV: A Parallel and IO-Friendly Key-Value Store Exploiting High-Bandwidth Solid State Drives	513
<i>Wanting Zhou, Yubo Zhang, Jiuyue Yan, Shaohua Wang, Shuhe Xia, Jie Yao, Qiang Cao</i>	
Spritz: Path-Aware Load Balancing in Low-Diameter Networks	527
<i>Tommaso Bonato, Ales Kubicek, Abdul Kabbani, Ahmad Ghalayini, Maciej Besta, Torsten Hoeftler</i>	
CACWS: Congestion-Aware Coordinated Warp Scheduler for Partitioned GPGPU	544
<i>Bo Yuan, Sheng Liu, Yang Guo, Jianfeng Cui, Zekun Jiang</i>	
MGS: Markov Greedy Sums for Low-Power DNN Accumulation	557
<i>Vikas Natesh, H.T. Kung, David Kong</i>	
TrioSeq: A Novel Approach to Accelerate Triplet Sequence Alignment on GPUs	570
<i>Miguel Graça, Aleksandar Ilic</i>	
Accelerating AI Compression through Lightweight Lossless Encoding and Pipelined Workflows	584
<i>Boyuan Zhang, Luanzheng Guo, Jiannan Tian, Jinyang Liu, Daoce Wang, Chengming Zhang, Bo Fang Fengguang Song, Jan Strube, Nathan Tallent, Dingwen Tao</i>	
PackKV: Reducing KV Cache Memory Footprint through LLM-Aware Lossy Compression	598
<i>Bo Jiang, Taolue Yang, Youyuan Liu, Xubin He, Sheng Di, Sian Jin</i>	
Near-Zero Cost KV Cache Compression for Large Language Model Inference	612
<i>Boyuan Zhang, Ding Zhou, Yafan Huang, Shihui Song, Hao Feng, Jinda Jia, Chengming Zhang, Zhi Zhang</i>	
Quantifying the Impact of Lossy Compression on Neural Generative Surrogate Modeling	626
<i>Zhimin Li, Harshitha Menon, Charles Jekel, Valerio Pascucci, Peter Lindstrom</i>	
Accelerating High-Frequency Electromagnetic Scattering Prediction with KNN-Augmented Radial Basis Function Networks	640
<i>Huajian Zhang, Rui Xia, Chao Li, Xinhai Chen, Tiaojie Xiao, Xiao-Wei Guo, Jie Liu</i>	

Lumos: Democratizing SciML Workflows with L0-Regularized Learning for Unified Feature and Parameter Adaptation	654
<i>Shouwei Gao, Xu Zheng, Dongsheng Luo, Sheng Di, Wenqian Dong</i>	
From Skew to Symmetry: Node-Interconnect Multi-Path Balancing with Execution-time Planning for Modern GPU Clusters	669
<i>Jinghan Yao, Kaushik Kandadi, Bharath Ramesh, Hari Subramoni, Dhabaleswar K. Panda</i>	
Optimizing When, How, and What to Communicate in Shared ML Clusters	683
<i>Tianxiang Huang, Waixi Liu, Jun Cai, Shipeng Fan, Wenli Shang</i>	
MosaicUFL: A Fine-grained Unlabeled Client Selection Framework for Efficient Federated Learning	693
<i>Nan Yang, Zikang Wen, Yuning Zhang, Yanli Li, Huaming Chen, Dong Yuan</i>	
FedACT: Concurrent Federated Intelligence across Heterogeneous Data Sources	707
<i>Md Sirajul Islam, Isabelle G Chapman, N I Md Ashafuddula, Xu Yuan, Li Chen, Nian-Feng Tzeng Klara Nahrstedt</i>	
FedDAAA: Federated Dynamic Parameter Alignment and Adaptive Aggregation	721
<i>Yuhao Chen, Ping Li, Ying He</i>	
Clust-PSI-PFL: A Population Stability Index Approach for Clustered Non-IID Personalized Federated Learning	732
<i>Daniel M. Jimenez-Gutierrez, Mehrdad Hassanzadeh, David Solans, Mohammed Elbamby Nicolas Kourtellis, Aris Anagnostopoulos, Ioannis Chatzigiannakis, Andrea Vitaletti</i>	
MeCache: Communication-Efficient Multi-GPU Heterogeneous Graph Neural Network Training	746
<i>Gongqingjian Jiang, Lizhi Zhang, Menghan Jia, Zhiquan Lai, Dongsheng Li</i>	
HarMoEny: Efficient Inference of MoE Models	760
<i>Zachary Doucet, Rishi Sharma, Martijn de Vos, Rafael Pires, Anne-Marie Kermarrec, Oana Balmau</i>	
Achieving Low Latency Inference on High Resolution Images by Exploiting Sparsity in Vision Transformers	775
<i>Changxin Li, Sanmukh Kuppannagari</i>	
Reforge: Low-Latency Distributed GNN Serving with Selective Embedding Recomputation	788
<i>Geon-Woo Kim, Donghyun Kim, Jeongyoon Moon, Henry Liu, Tarannum Khan, Anand Iyer, Daehyeok Kim Aditya Akella</i>	
EvalNet: A Practical Toolchain for Generation and Analysis of Extreme-Scale Interconnects	804
<i>Maciej Besta, Patrick Iff, Marcel Schneider, Nils Blach, Alessandro Maissen, Salvatore Di Girolamo Jens Domke, Jascha Krattenmacher, Kartik Lakhotia, Laura Monroe, Fabrizio Petrini Robert Gerstenberger, Torsten Hoefler</i>	
Beyond Thread States: Diagnosing Performance Degradation with eBPF and Thread Dynamics	819
<i>Diogo Landau, Jorge G. Barbosa, Nishant Saurabh</i>	
UCTRACE: A Multi-Layer Profiling Tool for UCX-driven Communication	835
<i>Emir Gencer, Mohammad Kefah Taha Issa, Ilyas Turimbetov, James D. Trotter, Didem Unat</i>	
SecPerf: Demystifying Cost of Confidential HPC	850
<i>Marcin Chrapek, Siyuan Shen, Patrick Iff, Tiancheng Chen, Mikhail Khalilov, Marcin Copik, Maciej Besta Torsten Hoefler</i>	

Characterizing Dataflow for I/O-Aware Scheduling in HPC Workflows	868
<i>Meng Tang, Luanzheng Guo, Anthony Kougkas, Xian-He Sun, Nathan R. Tallent</i>	
No One-Size-Fits-All: A Workload-Driven Characterization of Bit-Parallel vs. Bit-Serial Data Layouts for Processing-using-Memory	885
<i>Jingyao Zhang, Elaheh Sadredini</i>	
Characterizing Production GPU Workloads using System-wide Telemetry Data	899
<i>Onur Cankur, Brian Austin, Dhruva Kulkarni, Abhinav Bhatele</i>	
The Case of the Elusive Application Performance on Production GPU Supercomputers	913
<i>Cunyang Wei, Keshav Pradeep, Abhinav Bhatele</i>	
High-performance Vector-length Agnostic Quantum Circuit Simulations on ARM Processors	928
<i>Ruimin Shi, Gabin Schieffer, Pei-Hung Lin, Maya Gokhale, Andreas Hertel, Ivy Peng</i>	
Qontrast: Contrast Filters Mitigate Quantum Noise	942
<i>Wladimir Silva, Frank Mueller</i>	
Toward Energy-Efficient HPC: Insights from Power Profiling a Cloud-Resolving Earth System Model	956
<i>Zhengji Zhao, Noel Keen, Oscar Antepara, Samuel Williams, Luca Bertagna, Naser Mahfouz James B. White, Leonid Oliker, Brian Austin, Nicholas J. Wright</i>	
Beyond Throughput: Performance and Energy Insights of LLM Inference Across AI Accelerators	970
<i>Giacomo Brunetta, Varuni Sastry, Xingfu Wu, Valerie Taylor, Michael E. Papka, Zhiling Lan</i>	
SMART-MIG: A Learning Framework for Scalable and Energy-Efficient GPU Scheduling	983
<i>Wenqing Yu, Neel Karia, Tanvi Hisaria, Clifford Stein, Olivier Tardieu, Asser Tantawi</i>	
KORAL: Knowledge Graph Guided LLM Reasoning for SSD Operational Analysis	996
<i>Mayur Akewar, Sandeep Madireddy, Dongsheng Luo, Janki Bhimani</i>	
One Memory-Many Paths: Early Experiences with Allocation and Data Copy Strategies on MI300A	1012
<i>Goutham Kalikrishna Reddy Kuncham, Siyuan Zhang, Dhabaleswar K. Panda</i>	
Microbenchmarking NVIDIA's Blackwell Architecture: An in-depth Architectural Analysis	1026
<i>Aaron Jarmusch, Sunita Chandrasekaran</i>	
Rebo: Locality-Aware Graph Processing via Reordering and Blocking	1037
<i>YuAng Chen, Yeh-Ching Chung</i>	
Hiding Page Fault Latencies in Graph Processing Applications that Cannot Fit in Memory	1050
<i>Alireza Haddadi, David Black-Schaffer, Chang Hyun Park</i>	
Locality-Aware Distributed Allocators for High-Performance Global Data Structures	1064
<i>Ian Di Dio Lavore, Beatrice Branchini, Vito Giovanni Castellana, Marco Santambrogio</i>	
Dynamic NUMA-Aware Data Structure Replication	1077
<i>Erika Hunhoff, Zack McKeivitt, Ankit Bhardwaj, Reto Achermann, Gerd Zellweger, Marcos K. Aguilera Eric Keller</i>	
Designing Domain-Specific Compilers for Lossy Compression: A Case Study on Wafer-Scale Engine	1091
<i>Shihui Song, Robert Underwood, Sheng Di, Peng Jiang, Franck Cappello</i>	
Just-in-Time Convolution and GEMM Code Generation for SIMD Architectures	1105
<i>Alexandre de Limas Santana, Adrià Armejach, Marc Casas</i>	

DaggerFFT: A Distributed FFT Framework Using Task Scheduling in Julia	1120
<i>Sana Taghipour Anvari, Julian Samaroo, Matin Raayai Ardakani, David Kaeli</i>	
EdgeWeaver: Accelerating IoT Application Development Across Edge-Cloud Continuum	1134
<i>Pawissanutt Lertpongjukorn, Juahn Kwon, Hai Duc Nguyen, Mohsen Amini Salehi</i>	
PowerMapper: Power-Optimized Mapping of SNNs onto ReRAM Crossbars coupled via Packet-Switched NoCs	1149
<i>Devin Pohl, Kazi Asifuzzaman, Aaron R. Young, Narasinga Rao Miniskar, Jeffrey S. Vetter</i>	
LibraryX: A Framework for Cross-Library-Call Optimization	1161
<i>Sanil Rao, Anant Prakash, Naifeng Zhang, Het Mankad, Franz Franchetti</i>	
BAAS: A Bidirectional Aggregation and Affinity-Aware Scheduling Framework for Parallelizing Sparse Matrix Computations	1176
<i>Sen Yan, Qingyang Zhang, Chuanfu Xu, Chun Huang, Zhimeng Han, Jie Liu</i>	
APEX: Asynchronous Parallel CPU-GPU Execution for Online LLM Inference on Constrained GPUs	1190
<i>Jiakun Fan, Yanglin Zhang, Xiangchen Li, Dimitrios S. Nikolopoulos</i>	
PowerMorph: Shaping LLM Training for Data Center Demand Response	1204
<i>Boqiang Li, Luanzheng Guo, Buxin She, Nathan R. Tallent, Veronica Adetola, Rong Ge</i>	
Efficient MoE Inference on Single Consumer-grade GPU with Dynamic Expert Caching	1219
<i>Rui Zhang, Boxuan Yang, Rongji Wang, Xuemei Peng, Zeyi Wen</i>	
Demystifying ARM SME to Optimize General Matrix Multiplications	1233
<i>Chencheng Deng, Weiling Yang, Jianbin Fang, Dezun Dong</i>	
DynaGene: An Adaptive Scheduling Framework for AI-Coupled Workflows on Heterogeneous Resources	1246
<i>Xuxuan Yuan, Haifeng Ni, Ran Cheng, Zhuozhao Li</i>	
Enhancing HPC Batch Job Scheduling via Imitation Learning-Based Search	1260
<i>Zechun Zhou, Jingwei Sun, Mingfei Ye, Guangzhong Sun</i>	
PEACE: Preemptive and Efficient Cluster Scheduling for LLM Inference with Mixed Prompts	1272
<i>Zeyu Zhang, Haiying Shen</i>	
Shipyard: A Multi-Leader Consensus Protocol with Auto-Balanced Leadership of the Sharded Keyspace	1287
<i>Xincheng Yang, Kyle C. Hale</i>	
Design and Implementation of Casting Compression for GPU-Aware MPI Collectives	1301
<i>Chen-Chun Chen, Nicholas Contini, Lang Xu, Jake Queiser, Hari Subramoni, Dhabaleswar K. Panda</i>	
The Big Send-off: Scalable and Performant Collectives for Deep Learning	1314
<i>Siddharth Singh, Keshav Pradeep, Mahua Singh, Cunyang Wei, Abhinav Bhatele</i>	
QFSync: A Queue-Free Approach to Improve the Performance of Synchronization Primitives of Multithreaded Applications	1329
<i>Pierre Sens, Luciana Arantes, Julien Sopena</i>	
CARAT: Client-Side Adaptive RPC and Cache Co-Tuning for Parallel File Systems	1343
<i>Md Hasanur Rashid, Nathan R. Tallent, Forrest Sheng Bao, Dong Dai</i>	

Rubix: Adaptive and Fast-Performant Scaling for Serverless-Enabled ML Platforms	1358
<i>Amit Samanta, Andrea Pinto, Flavio Esposito</i>	
QoSFlow: Ensuring Service Quality of Distributed Workflows Using Interpretable Sensitivity Models	1372
<i>Md Hasanur Rashid, Jesun Firoz, Nathan R. Tallent, Luanzheng Guo, Meng Tang, Dong Dai</i>	
WaterSplit: Coordinated On-Site and Off-Site Water Allocation For Sustainable Datacenter Cooling	1388
<i>Yankai Jiang, Raghavendra Kanakagiri, Rohan Basu Roy, Devesh Tiwari</i>	
Empowering Scientific Workflows with Federated Agents	1403
<i>Alok Kamatar, J. Gregory Pauloski, Yadu Babuji, Ryan Chard, Mansi Sakarvadia, Daniel Babnigg</i>	
<i>Ian Foster, Kyle Chard</i>	