

SC24: International Conference for High Performance Computing, Networking, Storage and Analysis

**Atlanta, Georgia, USA
17-22 November 2024**

Pages 1-571



**IEEE Catalog Number: CFP24SUP-POD
ISBN: 979-8-3503-5292-4**

**Copyright © 2024 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:	CFP24SUP-POD
ISBN (Print-On-Demand):	979-8-3503-5292-4
ISBN (Online):	979-8-3503-5291-7

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

TABLE OF CONTENTS

ACM GORDON BELL CLIMATE MODELING FINALISTS

ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability.....	1
<i>Xiao Wang, Siyan Liu, Aristeidis Tsaris, Jong-Youl Choi, Ashwin M. Aji, Ming Fan, Wei Zhang, Junqi Yin, Moetasim Ashfaq, Dan Lu, Prasanna Balaprakash</i>	
Boosting Earth System Model Outputs and Saving PetaBytes in Their Storage using Exascale Climate Emulators.....	12
<i>Sameh Abdulah, Allison H. Baker, George Bosilca, Qinglei Cao, Stefano Castruccio, Marc G. Genton, David E. Keyes, Zubair Khalid, Hatem Ltaief, Yan Song, Georgiy L. Stenchikov, Ying Sun</i>	
A Performance-Portable Kilometer-Scale Global Ocean Model on ORISE and New Sunway Heterogeneous Supercomputers	24
<i>Junlin Wei, Xiang Han, Jiangfeng Yu, Jinrong Jiang, Hailong Liu, Pengfei Lin, Maoxue Yu, Kai Xu, Lian Zhao, Pengfei Wang, Weipeng Zheng, Jingwei Xie, Yanzhi Zhou, Tao Zhang, Feng Zhang, Yehong Zhang, Yue Yu, Yuzhu Wang, Yidi Bai, Chen Li, Zipeng Yu, Haoyu Deng, Yaxin Li, Xuebin Chi</i>	

ACM GORDON BELL FINALISTS

Democratizing AI: Open-Source Scalable LLM Training on GPU-Based Supercomputers	36
<i>Siddharth Singh, Prajwal Singhania, Aditya Ranjan, John Kirchenbauer, Jonas Geiping, Yuxin Wen, Neel Jain, Abhimanyu Hans, Manli Shu, Aditya Tomar, Tom Goldstein, Abhinav Bhatele</i>	
Pushing the Limit of Quantum Mechanical Simulation to the Raman Spectra of a Biological System with 100 Million Atoms.....	50
<i>Honghui Shang, Ying Liu, Zhikun Wu, Zhenchuan Chen, Jinfeng Liu, Meiyue Shao, Yingzhou Li, Bowen Kan, Huimin Cui, Xiaobing Feng, Yunquan Zhang, Donald G. Truhlar, Hong An, Xiao He, Jinlong Yang</i>	
Toward Capturing Genetic Epistasis from Multivariate Genome-Wide Association Studies using Mixed-Precision Kernel Ridge Regression	62
<i>Hatem Ltaief, Rabab Alomairy, Qinglei Cao, Jie Ren, Lotfi Slim, Thorsten Kurth, Benedikt Dorschner, Salim Bougouffa, Rached Abdelkhalak, David E. Keyes</i>	
MProt-DPO: Breaking the ExaFLOPS Barrier for <u>M</u> ultimodal <u>P</u> rotein Design Workflows with <u>D</u> irect <u>P</u> reference <u>O</u> ptimization	74
<i>Gautham Dharuman, Kyle Hippe, Alexander Brace, Sam Foreman, Väinö Hatanpää, Varuni K. Sastry, Huihuo Zheng, Logan Ward, Servesh Muralidharan, Archit Vasan, Bharat Kale, Carla M. Mann, Heng Ma, Yun-Hsuan Cheng, Yuliana Zamora, Shengchao Liu, Chaowei Xiao, Murali Emani, Tom Gibbs, Mahidhar Tatineni, Deepak Canchi, Jerome Mitchell, Koichi Yamada, Maria Garzaran, Michael E. Papka, Ian Foster, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, Arvind Ramanathan</i>	
Breaking the Molecular Dynamics Timescale Barrier using a Wafer-Scale System	87
<i>Kylee Santos, Stan Moore, Tomas Oppelstrup, Amirali Sharifian, Ilya Sharapov, Aidan Thompson, Delyan Z. Kalchev, Danny Perez, Robert Schreiber, Scott Pakin, Edgar A. Leon, James H. Laros, Michael James, Sivasankaran Rajamanickam</i>	

Breaking the Million-Electron and 1 EFLOP/s Barriers: Biomolecular-Scale <i>Ab Initio</i> Molecular Dynamics using MP2 Potentials.....	100
<i>Ryan Stocks, Jorge L. G. Vallejo, Fiona C. Y. Yu, Calum Snowdon, Elise Palethorpe, Jakub Kurzak, Dmytro Bykov, Giuseppe M. J. Barca</i>	

TECHNICAL PAPERS

ENERGY AND CARBON-EFFICIENT ARCHITECTURES

M ³ XU: Achieving High-Precision and Complex Matrix Multiplication with Low-Precision MXUs.....	112
<i>Dongho Ha, Yunan Zhang, Chen-Chien Kao, Christopher J. Hughes, Won W. Ro, Hung-Wei Tseng</i>	
Hydrogen: Contention-Aware Hybrid Memory for Heterogeneous CPU-GPU Architectures	128
<i>Yiwei Li, Mingyu Gao</i>	
EcoLife: Carbon-Aware Serverless Function Scheduling for Sustainable Computing.....	143
<i>Yankai Jiang, Rohan B. Roy, Baolin Li, Devesh Tiwari</i>	

HIGH-PERFORMANCE COMPRESSION AND I/O MANAGEMENT

CUSZ- <i>i</i> : High-Ratio Scientific Lossy Compression on GPUs with Optimized Multi-Level Interpolation	158
<i>Jinyang Liu, Jiannan Tian, Shixun Wu, Sheng Di, Boyuan Zhang, Robert Underwood, Yafan Huang, Jiajun Huang, Kai Zhao, Guanpeng Li, Dingwen Tao, Zizhong Chen, Franck Cappello</i>	
Tango: A Cross-Layer Approach to Managing I/O Interference Over Local Ephemeral Storage.....	173
<i>Zhenbo Qiao, Qirui Tian, Zhenlu Qin, Jinzhen Wang, Qin G. Liu, Norbert Podhorszki, Scott Klasky, Hongjian Zhu</i>	
CUSZP2: A GPU Lossy Compressor with Extreme Throughput and Optimized Compression Ratio	188
<i>Yafan Huang, Sheng Di, Guanpeng Li, Franck Cappello</i>	

WORKFLOW CHARACTERIZATION AND OPTIMIZATION

LLM-Pilot: Characterize and Optimize Performance of Your LLM Inference Services	206
<i>Malgorzata Lazuka, Andreea Anghel, Thomas Parnell</i>	
DFTracer: An Analysis-Friendly Data Flow Tracer for AI-Driven Workflows	224
<i>Hariharan Devarajan, Loïc Pottier, Kaushik Velusamy, Huihuo Zheng, Izzet Yildirim, Olga Kogiou, Weikuan Yu, Anthony Kougkas, Xian-He Sun, Jae S. Yeom, Kathryn Mohror</i>	
Efficient Weighted Graph Matching on GPUs	248
<i>Michael Mandulak, Sayan Ghosh, S. M. Ferdous, Mahantesh Halappanvar, George Slota</i>	

COMPILER ANALYSIS AND CODE GENERATION

Automated Code Generation of High-Order Stencils for a Dataflow Architecture	264
<i>Ryuichi Sai, John Mellor-Crummey, Jinfan Xu, Mauricio Araya-Polo</i>	

Moirae: Generating High-Performance Composite Stencil Programs with Global Optimizations	277
<i>Xiaoyan Liu, Xinyu Yang, Kejie Ma, Shanghao Liu, Kaige Zhang, Hailong Yang, Yi Liu, Zhongzhi Luan, Depei Qian</i>	

autoGEMM: Pushing the Limits of Irregular Matrix Multiplication on Arm Architectures.....	292
<i>Du Wu, Jintao Meng, Wenxi Zhu, Minwen Deng, Xiao Wang, Tao Luo, Mohamed Wahib, Yanjie Wei</i>	

POWER MANAGEMENT AND COOLING

Accurate and Convenient Energy Measurements for GPUs: A Detailed Study of NVIDIA GPU's Built-In Power Sensor	307
<i>Zeyu Yang, Karel Adamek, Wesley Armour</i>	

A Digital Twin Framework for Liquid-Cooled Supercomputers as Demonstrated at Exascale.....	324
<i>Wesley Brewer, Matthias Maiterth, Vineet Kumar, Rafal Wojda, Sedrick Bouknight, Jesse Hines, Woong Shin, Scott Greenwood, David Grant, Wesley Williams, Feiyi Wang</i>	

Toward Sustainable HPC: In-Production Deployment of Incentive-Based Power Efficiency Mechanism on the Fugaku Supercomputer	342
<i>Ana L. V. Solórzano, Kento Sato, Keiji Yamamoto, Fumiyoshi Shoji, Jim M. Brandt, Benjamin Schwaller, Sara P. Walton, Jennifer Green, Devesh Tiwari</i>	

SCHEDULING

Towards Highly Compatible I/O-Aware Workflow Scheduling on HPC Systems	358
<i>Yiqin Dai, Ruibo Wang, Yong Dong, Kai Lu</i>	

PAL: A Variability-Aware Policy for Scheduling ML Workloads in GPU Clusters.....	373
<i>Rutwik Jain, Brandon Tran, Keting Chen, Matthew D. Sinclair, Shivaram Venkataraman</i>	

Toward High-Performance Blockchain System by Blurring the Line Between Ordering and Execution.....	391
<i>Donghyeon Ryu, Chanik Park</i>	

ADVANCED COMPUTATIONAL METHODS AND ARCHITECTURES

Matrix-Free Finite Volume Kernels on a Dataflow Architecture.....	407
<i>Ryuichi Sai, François P. Hamon, John Mellor-Crummey, Mauricio Araya-Polo</i>	

Rapid GPU-Based Pangenome Graph Layout.....	418
<i>Jiajie Li, Jan-Niklas Schmelzle, Yixiao Du, Simon Heumos, Andrea Guarracino, Giulia Guidi, Pjotr Prins, Erik Garrison, Zhiru Zhang</i>	

Scaling Molecular Dynamics with <i>ab initio</i> Accuracy to 149 Nanoseconds Per Day	437
<i>Jianxiong Li, Boyang Li, Zhuoqiang Guo, Mingzhen Li, Enji Li, Lijun Liu, Guojun Yuan, Zhan Wang, Guangming Tan, Weile Jia</i>	

LEADERSHIP-CLASS SUPERCOMPUTERS

An Evaluation of the Effect of Network Cost Optimization for Leadership Class Supercomputers	452
<i>Awais Khan, John R. Lange, Nick Hagerty, Edwin F. Posada, John Holmen, James B. White, Austin Harris, Verónica M. Vergara, Christopher Zimmer, Scott Atchley</i>	

Application-Driven Exascale: The JUPITER Benchmark Suite.....	468
<i>Andreas Herten, Sebastian Achilles, Damian Alvarez, Jayesh Badwaik, Eric Behle, Mathis Bode, Thomas Breuer, Daniel Caviedes-Voullième, Mehdi Cherti, Adel Dabah, Salem El Sayed, Wolfgang Frings, Ana Gonzalez-Nicolas, Eric B. Gregory, Kaveh H. Mood, Thorsten Hater, Jenia Jitsev, Chelsea M. John, Jan H. Meinke, Catrin I. Meyer, Pavel Mezentsev, Jan-Oliver Mirus, Stepan Nassyr, Carolin Penke, Manoel Römmer, Ujjwal Sinha, Benedikt Von St. Vieth, Olaf Stein, Estela Suarez, Dennis Willsch, Ilya Zhukov</i>	

Exploring GPU-to-GPU Communication: Insights into Supercomputer Interconnects	513
<i>Daniele De Sensi, Lorenzo Pichetti, Flavio Vella, Tiziano De Matteis, Zebin Ren, Luigi Fusco, Matteo Turisini, Daniele Cesarini, Kurt Lust, Animesh Trivedi, Duncan Roweth, Filippo Spiga, Salvatore Di Girolamo, Torsten Hoefler</i>	

PARALLEL PROGRAM ANALYSIS AND CODE OPTIMIZATION

MCFuser: High-Performance and Rapid Fusion of Memory-Bound Compute-Intensive Operators	528
<i>Zheng Zhang, Donglin Yang, Xiaobo Zhou, Dazhao Cheng</i>	
Static Generation of Efficient OpenMP Offload Data Mappings	543
<i>Luke Marzen, Akash Dutta, Ali Jannesari</i>	
HiRace: Accurate and Fast Data Race Checking for GPU Programs	558
<i>John Jacobson, Martin Burtcher, Ganesh Gopalakrishnan</i>	

SERVERLESS COMPUTING AND DISAGGREGATED MEMORY

Boosting Data Center Performance via Intelligently Managed Multi-Backend Disaggregated Memory	572
<i>Jing Wang, Hanzhang Yang, Chao Li, Yiming Zhuansun, Wang Yuan, Cheng Xu, Xiaofeng Hou, Minyi Guo, Yang Hu, Yaqian Zhao</i>	
SMIless: Serving DAG-Based Inference with Dynamic Invocations Under Serverless Computing.....	590
<i>Chengzhi Lu, Huanle Xu, Yudan Li, Wenyan Chen, Kejiang Ye, Chengzhong Xu</i>	
Stellaris: Staleness-Aware Distributed Reinforcement Learning with Serverless Computing	607
<i>Hanfei Yu, Hao Wang, Devesh Tiwari, Jian Li, Seung-Jong Park</i>	

GPU OPTIMIZATIONS FOR ML

PipeInfer: Accelerating LLM Inference using Asynchronous Pipelined Speculation.....	624
<i>Branden Butler, Sixing Yu, Arya Mazaheri, Ali Jannesari</i>	
RecFlex: Enabling Feature Heterogeneity-Aware Optimization for Deep Recommendation Models with Flexible Schedules.....	643
<i>Zaifeng Pan, Zhen Zheng, Feng Zhang, Bing Xie, Ruofan Wu, Shaden Smith, Chuanjie Liu, Olatunji Ruwase, Xiaoyong Du, Yufei Ding</i>	
ParvaGPU: Efficient Spatial GPU Sharing for Large-Scale DNN Inference in Cloud Environments	658
<i>Munkyu Lee, Sihoon Seong, Minki Kang, Jihyuk Lee, Gap-Joo Na, In-Geol Chun, Dimitrios Nikolopoulos, Cheol-Ho Hong</i>	

PARALLEL PROGRAMMING FRAMEWORKS, LIBRARIES AND RUNTIMES

CUDASTF: Bridging the Gap Between CUDA and Task Parallelism	672
<i>Cédric Augonnet, Andrei Alexandrescu, Albert Sidelnik, Michael Garland</i>	
KaMPIng: Flexible and (Near) Zero-Overhead C++ Bindings for MPI	689
<i>Tim N. Uhl, Matthias Schimek, Lukas Hübner, Demian Hespe, Florian Kurpicz, Daniel Seemaier, Christoph Stelz, Peter Sanders</i>	
NetCL: A Unified Programming Framework for In-Network Computing	710
<i>George Karlos, Henri Bal, Lin Wang</i>	

SPARSE MATRIX COMPUTATIONS

Distributed-Memory Parallel Algorithms for Sparse Matrix and Sparse Tall-And-Skinny Matrix Multiplication	730
<i>Isuru Ranawaka, M. Taufique Hussain, Charles Block, Gerasimos Gerogiannis, Josep Torrellas, Ariful Azad</i>	
A Sparsity-Aware Distributed-Memory Algorithm for Sparse-Sparse Matrix Multiplication	747
<i>Yuxi Hong, Aydin Buluç</i>	
A Conflict-Aware Divide-And-Conquer Algorithm for Symmetric Sparse Matrix-Vector Multiplication	761
<i>Haozhong Qiu, Chuanfu Xu, Jianbin Fang, Jian Zhang, Liang Deng, Yue Ding, Qingsong Wang, Shizhao Chen, Yonggang Che, Jie Liu</i>	

MACHINE LEARNING APPLICATIONS

Accelerating Distributed DLRM Training with Optimized TT Decomposition and Micro-Batching	776
<i>WeiHu Wang, Yaqi Xia, Donglin Yang, Xiaobo Zhou, Dazhao Cheng</i>	
Scaling New Heights: Transformative Cross-GPU Sampling for Training Billion-Edge Graphs	791
<i>Yaqi Xia, Donglin Yang, Xiaobo Zhou, Dazhao Cheng</i>	
A Scalable Algorithm for Active Learning	806
<i>Youguang Chen, Zheyu Wen, George Biros</i>	

MATRIX COMPUTATIONS ON TENSOR CORES

AmgT: Algebraic Multigrid Solver on Tensor Cores	823
<i>Yuechen Lu, Lijie Zeng, Tengcheng Wang, Xu Fu, Wenxuan Li, Helin Cheng, Dechuang Yang, Zhou Jin, Marc Casas, Weifeng Liu</i>	
LoRAStencil: Low-Rank Adaptation of Stencil Computation on Tensor Cores	839
<i>Yiwei Zhang, Kun Li, Liang Yuan, Jiawen Cheng, Yunquan Zhang, Ting Cao, Mao Yang</i>	
High Performance Unstructured SpMM Computation using Tensor Cores	856
<i>Patrik Okanovic, Grzegorz Kwasniewski, Paolo S. Labini, Maciej Besta, Flavio Vella, Torsten Hoeftler</i>	

PERFORMANCE ANALYSIS

Versatile Datapath Soft Error Detection on the Cheap for HPC Applications	870
<i>Yafan Huang, Sheng Di, Zhaorui Zhang, Xiaoyi Lu, Guanpeng Li</i>	
MCBound: An Online Framework to Characterize and Classify Memory/Compute-Bound HPC Jobs	885
<i>Francesco Antici, Andrea Bartolini, Zeynep Kiziltan, Ozalp Babaoglu, Yuetsu Kodama</i>	
GVARP: Detecting Performance Variance on Large-Scale Heterogeneous Systems	900
<i>Xin You, Zhibo Xuan, Hailong Yang, Zhongzhi Luan, Yi Liu, Depei Qian</i>	

HIGH-PERFORMANCE SOLVERS

Mille-Feuille: A Tile-Grained Mixed Precision Single-Kernel Conjugate Gradient Solver on GPUs	916
<i>Dechuang Yang, Yuxuan Zhao, Yiduo Niu, Weile Jia, En Shao, Weifeng Liu, Guangming Tan, Zhou Jin</i>	
DBSR: An Efficient Storage Format for Vectorizing Sparse Triangular Solvers on Structured Grids	932
<i>Xiaojian Yang, Shengguo Li, Fan Yuan, Dezun Dong</i>	
Many-Body Electronic Correlation Energy using Krylov Subspace Linear Solvers	946
<i>Shikhar Shah, Boqin Zhang, Hua Huang, John E. Pask, Phanish Suryanarayana, Edmond Chow</i>	

HPC FOR PHYSICS AND MATERIAL SCIENCE

Enabling 13K-Atom Excited-State GW Calculations via Low-Rank Approximations and HPC on the New Sunway Supercomputer	961
<i>Wentiao Wu, Zhengbang Zhou, Qingcai Jiang, Junwei Feng, Xinming Qin, Huanhuan Ma, Zhenwei Cao, Junshi Chen, Sheng Chen, Xinyong Meng, Bingkun Hou, Yuanfan Xiong, Linhao Wang, Yixuan Sun, Hong An, Jinlong Yang, Wei Hu</i>	
Reshaping High Energy Physics Applications for Near-Interactive Execution using TaskVine	975
<i>Barry Sly-Delgado, Ben Tovar, Jin Zhou, Douglas Thain</i>	
Towards Exascale Simulations of Nanoelectronic Devices in the GW Approximation	988
<i>Leonard Deuschle, Alexander Maeder, Vincent Maillou, Nicolas Vetsch, Anders Winka, Jiang Cao, Alexandros N. Ziogas, Mathieu Luisier</i>	

PERFORMANCE MODELING

LLAMP: Assessing Network Latency Tolerance of HPC Applications with Linear Programming	1004
<i>Siyuan Shen, Langwen Huang, Marcin Chrapek, Timo Schneider, Jai Dayal, Manisha Gajbe, Robert Wisniewski, Torsten Hoefer</i>	
A Workflow Roofline Model for End-To-End Workflow Performance Analysis	1022
<i>Nan Ding, Brian Austin, Yang Liu, Neil Mehta, Steven Farrell, Johannes P. Blaschke, Leonid Oliker, Hai A. Nam, Nicholas J. Wright, Samuel Williams</i>	
Learning Generalizable Program and Architecture Representations for Performance Modeling	1037
<i>Lingda Li, Thomas Flynn, Adolfo Hoisie</i>	

QUANTUM AND APPROXIMATE COMPUTING I

LEXIQL: Quantum Natural Language Processing on NISQ-Era Machines	1052
<i>Daniel Silver, Aditya Ranjan, Rakesh Achutha, Tirthak Patel, Devesh Tiwari</i>	
Optimizing Quantum Fourier Transformation (QFT) Kernels for Modern NISQ and FT Architectures	1067
<i>Yuwei Jin, Xiangyu Gao, Minghao Guo, Henry Chen, Fei Hua, Chi Zhang, Eddy Z. Zhang</i>	
On the Efficacy of Surface Codes in Compensating for Radiation Events in Superconducting Devices	1082
<i>Marzio Vallero, Gioele Casagrande, Flavio Vella, Paolo Rech</i>	

ANALYSIS OF HPC SYSTEMS

Revisiting Computation for Research: Practices and Trends.....	1097
<i>Jeremiah Giordani, Ziyang Xu, Ella Colby, August Ning, Bhargav R. Godala, Ishita Chaturvedi, Shaowei Zhu, Yebin Chon, Greg Chan, Zujun Tan, Galen Collier, Jonathan D. Halverson, Enrico A. Deiana, Jasper Liang, Federico Sossai, Yian Su, Atmn Patel, Bangyen Pham, Nathan Greiner, Simone Campanoni, David I. August</i>	
Understanding Data Movement Patterns in HPC: A NERSC Case Study.....	1111
<i>Anna Giannakou, Damian Hazen, Bjoern Enders, Lavanya Ramakrishnan, Nicholas J. Wright</i>	

QUANTUM AND APPROXIMATE COMPUTING II

HPAC-ML: A Programming Model for Embedding ML Surrogates in Scientific Applications.....	1128
<i>Zane Fink, Konstantinos Parasyris, Praneet Rathi, Giorgis Georgakoudis, Harshitha Menon, Peer-Timo Bremer</i>	
PARALLAX: A Compiler for Neutral Atom Quantum Computers Under Hardware Constraints	1144
<i>Jason Ludmir, Tirthak Patel</i>	

SPARSITY AND QUANTIZATION IN ML

MixQ: Taming Dynamic Outliers in Mixed-Precision Quantization by Online Prediction.....	1161
<i>Yidong Chen, Chen Zhang, Rongchao Dong, Haoyuan Zhang, Yonghua Zhang, Zhonghua Lu, Jidong Zhai</i>	
LONG EXPOSURE: Accelerating Parameter-Efficient Fine-Tuning for LLMs Under Shadowy Sparsity	1176
<i>Tuowei Wang, Kun Li, Zixu Hao, Donglin Bai, Ju Ren, Yaoyue Zhang, Ting Cao, Mao Yang</i>	

EFFICIENT TRANSFORMERS

Adaptive Patching for High-Resolution Image Segmentation with Transformers	1194
<i>Enzhi Zhang, Isaac Lyngaas, Peng Chen, Xiao Wang, Jun Igarashi, Yuankai Huo, Masaharu Munetomo, Mohamed Wahib</i>	
TorchGT: A Holistic System for Large-Scale Graph Transformer Training	1210
<i>Meng Zhang, Jie Sun, Qinghao Hu, Peng Sun, Zeke Wang, Yonggang Wen, Tianwei Zhang</i>	

Exploring Efficient Partial Differential Equation Solution using Speed Galerkin Transformer.....	1227
<i>Xun Wang, Zeyang Zhu, Xiangyu Meng, Tao Song</i>	

QUANTUM AND APPROXIMATE COMPUTING III

Surpassing Sycamore: Achieving Energetic Superiority Through System-Level Circuit Simulation	1241
<i>Rong Fu, Zhongling Su, Han-Sen Zhong, Xiti Zhao, Jianyang Zhang, Feng Pan, Pan Zhang, Xianhe Zhao, Ming-Cheng Chen, Chao-Yang Lu, Jian-Wei Pan, Zhilin Pei, Xingcheng Zhang, Wanli Ouyang</i>	
Realizing Quantum Kernel Models at Scale with Matrix Product State Simulation	1261
<i>Mekena Metcalf, Pablo Andrés-Martínez, Nathan Fitzpatrick</i>	
Atlas: Hierarchical Partitioning for Quantum Circuit Simulation on GPUs.....	1281
<i>Mingkuan Xu, Shiyi Cao, Xupeng Miao, Umut A. Acar, Zhihao Jia</i>	

RESOURCE UTILIZATION AND PACKAGE MANAGEMENT

Unlocking High Performance with Low-Bit NPUs and CPUs for Highly Optimized HPL-MxP on Cloud Brain II.....	1298
<i>Weicheng Xue, Kai Yang, Yongxiang Liu, Dengdong Fan, Pengxiang Xu, Yonghong Tian</i>	
Fire-Flyer AI-HPC: A Cost-Effective Software-Hardware Co-Design for Deep Learning	1314
<i>Wei An, Xiao Bi, Guanting Chen, Shanhuang Chen, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Wenjun Gao, Kang Guan, Jianzhong Guo, Yongqiang Guo, Zhe Fu, Ying He, Panpan Huang, Jiashi Li, Wenfeng Liang, Xiaodong Liu, Xin Liu, Yiyuan Liu, Yuxuan Liu, Shanghai Lu, Xuan Lu, Xiaotao Nie, Tian Pei, Junjie Qiu, Hui Qu, Zehui Ren, Zhangli Sha, Xuecheng Su, Xiaowen Sun, Yixuan Tan, Minghui Tang, Shiyu Wang, Yaohui Wang, Yongji Wang, Ziwei Xie, Yiliang Xiong, Yanhong Xu, Shengfeng Ye, Shuiping Yu, Yukun Zha, Liyue Zhang, Haowei Zhang, Mingchuan Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Yuheng Zou</i>	
A Probabilistic Approach to Selecting Build Configurations in Package Managers	1337
<i>Daniel Nichols, Harshitha Menon, Todd Gamblin, Abhinav Bhatele</i>	

SCIENTIFIC DATA PROCESSING AND VISUALIZATION

A High-Quality Workflow for Multi-Resolution Scientific Data Reduction and Visualization.....	1350
<i>Daoce Wang, Pascal Grosset, Jesus Pulido, Tushar M. Athawale, Jiannan Tian, Kai Zhao, Zarija Lukic, Axel Huebl, Zhe Wang, James Ahrens, Dingwen Tao</i>	
Error-Controlled Progressive Retrieval of Scientific Data Under Derivable Quantities of Interest	1368
<i>Xuan Wu, Qian Gong, Jieyang Chen, Qing Liu, Norbert Podhorszki, Xin Liang, Scott Klasky</i>	
CARP: Range Query-Optimized Indexing for Streaming Data.....	1384
<i>Ankush Jain, Charles D. Cranor, Qing Zheng, Bradley W. Settlemyer, George Amvrosiadis, Gary A. Grider</i>	

COMMUNICATION OPTIMIZATION FOR ML

Optimizing Distributed ML Communication with Fused Computation-Collective Operations	1403
<i>Kishore Punniyamurthy, Khaled Hamidouche, Bradford M. Beckmann</i>	

Accelerating Communication in Deep Learning Recommendation Model Training with Dual-Level Adaptive Lossy Compression.....	1420
<i>Hao Feng, Boyuan Zhang, Fanjiang Ye, Min Si, Ching-Hsiang Chu, Jiannan Tian, Chunxing Yin, Summer Deng, Yuchen Hao, Pavan Balaji, Tong Geng, Dingwen Tao</i>	

APTMoE: Affinity-Aware Pipeline Tuning for MoE Models on Bandwidth-Constrained GPU Nodes.....	1436
<i>Yuanxin Wei, Jiangsu Du, Jiazhi Jiang, Xiao Shi, Xianwei Zhang, Dan Huang, Nong Xiao, Yutong Lu</i>	

COMPUTATIONAL EFFICIENCY AND LEARNING TECHNIQUES

Accelerated Atomistic Kinetic Monte Carlo Simulations of Resistive Memory Arrays.....	1450
<i>Manasa Kaniselman, Alexander Maeder, Marko Mladenovic, Mathieu Luisier, Alexandros N. Ziogas</i>	

Large Language Models for Anomaly Detection in Computational Workflows: From Supervised Fine-Tuning to In-Context Learning.....	1466
<i>Hongwei Jin, George Papadimitriou, Krishnan Raghavan, Pawel Zuk, Prasanna Balaprakash, Cong Wang, Anirban Mandal, Ewa Deelman</i>	

Designing a GPU-Accelerated Communication Layer for Efficient Fluid-Structure Interaction Computations on Heterogeneous Systems.....	1483
<i>Aristotle Martin, Geng Liu, Balint Joo, Runxin Wu, Mohammed S. Kabir, Erik W. Draeger, Amanda Randles</i>	

SCALE-UP INTERCONNECTS

Efficient Tensor Offloading for Large Deep-Learning Model Training Based on Compute Express Link	1503
<i>Dong Xu, Yuan Feng, Kwangsik Shin, Daewoo Kim, Hyeran Jeon, Dong Li</i>	

COAXIAL: A CXL-Centric Memory System for Scalable Servers	1521
<i>Albert Cho, Anish Saxena, Moinuddin Qureshi, Alexandros Daglis</i>	

Switch-Less Dragonfly on Wafers: A Scalable Interconnection Architecture Based on Wafer-Scale Integration	1536
<i>Yinxiao Feng, Kaisheng Ma</i>	

SCALING AND CHECKPOINTING

Fast and Efficient Scaling for Microservices with SurgeGuard.....	1553
<i>Anyesha Ghosh, Neeraja J. Yadwadkar, Mattan Erez</i>	

Realizing Joint Extreme-Scale Simulations on Multiple Supercomputers—Two Superfacility Case Studies	1568
<i>Theresa Pollinger, Alexander Van Craen, Philipp Offenhäuser, Dirk Pflüger</i>	

AutoCheck: Automatically Identifying Variables for Checkpointing by Data Dependency Analysis	1585
<i>Xiang Fu, Weiping Zhang, Shiman Meng, Xin Huang, Wubiao Xu, Luanzheng Guo, Kento Sato</i>	

GRAPH ALGORITHMS AND COMPUTATION ON GRAPHS

Enumeration of Billions of Maximal Bicliques in Bipartite Graphs Without using GPUs.....	1601
<i>Zhe Pan, Shuibing He, Xu Li, Xuechen Zhang, Yanlong Yin, Rui Wang, Lidan Shou, Mingli Song, Xian-He Sun, Gang Chen</i>	
Doubling Graph Traversal Efficiency to 198 TeraTEPS on the Supercomputer Fugaku.....	1616
<i>Junya Arai, Masahiro Nakao, Yuto Inoue, Kanto Teranishi, Koji Ueno, Keiichiro Yamamura, Mitsuhsa Sato, Katsuki Fujisawa</i>	
Asynchronous Distributed-Memory Parallel Algorithms for Influence Maximization	1630
<i>Shubhendra P. Singhal, Souvadra Hati, Jeffrey Young, Vivek Sarkar, Akihiro Hayashi, Richard Vuduc</i>	

SCALE-OUT INTERCONNECTS

Network-Offloaded Bandwidth-Optimal Broadcast and Allgather for Distributed AI	1649
<i>Mikhail Khalilov, Salvatore Di Girolamo, Marcin Chrapek, Rami Nudelman, Gil Bloch, Torsten Hoefler</i>	
hZCCL: Accelerating Collective Communication with Co-Designed Homomorphic Compression.....	1666
<i>Jiajun Huang, Sheng Di, Xiaodong Yu, Yujia Zhai, Jinyang Liu, Zizhe Jian, Xin Liang, Kai Zhao, Xiaoyi Lu, Zizhong Chen, Franck Cappello, Yanfei Guo, Rajeev Thakur</i>	
UNR: Unified Notifiable RMA Library for HPC	1681
<i>Guangnan Feng, Jiabin Xie, Dezun Dong, Yutong Lu</i>	

STORAGE MANAGEMENT

EXO: Accelerating Storage Paravirtualization with eBPF	1696
<i>Shi Qiu, Li Wang, Yiming Zhang</i>	
CoRD: Combining Raid and Delta for Fast Partial Updates in Erasure-Coded Storage Clusters	1711
<i>Hai Zhou, Dan Feng, Yuchong Hu, Wei Wang, Huadong Huang</i>	
MegaMmap: Blurring the Boundary Between Memory and Storage for Data-Intensive Workloads	1725
<i>Luke Logan, Anthony Kougkas, Xian-He Sun</i>	

Author Index