



Association for
Computing Machinery

Advancing Computing as a Science & Profession

June 8 – 11, 2025
Salt Lake City, USA



ACM ICS '25

Proceedings of the 39th ACM

International Conference on Supercomputing

Sponsored by:

ACM SIGARCH & SIGHPC



**Association for
Computing Machinery**

Advancing Computing as a Science & Profession

The Association for Computing Machinery

1601 Broadway, 10th Floor
New York, New York 10019, USA

ACM COPYRIGHT NOTICE. Copyright © 2025 by the Association for Computing Machinery, Inc. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Publications Dept., ACM, Inc., fax +1 (212) 869-0481, or permissions@acm.org.

For other copying of articles that carry a code at the bottom of the first or last page, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, +1-978-750-8400, +1-978-750-4470 (fax).

ISBN: 979-8-4007-1537-2

Sponsors



Association for
Computing Machinery



ACM SIGARCH



Message from the General Chair

Welcome to the 39th ACM International Conference on Supercomputing (ICS 2025), held June 8-11 in Salt Lake City, USA. ICS has been a premiere forum for presentation of research in all aspects of high-performance computing, including architecture, algorithms/applications, compilers and runtime systems, performance modeling/measurement, programming languages, and systems software. The key to the success of the conference is the creation of a high-quality program covering a wide range of topics with results that advance the state-of-the-art. I thank the Program Chairs Michael Ferdman and Xipeng Shen for their tireless efforts in managing the dual submissions (for the first time at ICS 2025) and the creation of a very strong technical program with 83 papers.

Grateful thanks to ACM and the co-sponsorship of ICS by SIGARCH and SIGHPC, without whose support the conference would not be possible. Thanks to SIGHPC for providing funds for travel grants to student and early career attendees.

I am very thankful for all the heavy lifting by the rest of the team that made ICS 2025 possible: Hari Sundar, the Finance and Registration Chair; Varun Shankar, the Local Arrangements Chair; Yuke Wang, the Publications Chair; Wenqian Dong, the Publicity Chair; Rohan Basu Roy, the Travel Grants Chair; Yufan Xu, the Web Liaison; and Mehmet Belviranli, the Workshops Chair.

I wish to express my gratitude to the current and former steering committee chairs, Dimitrios Nikolopoulos and Alex Veidenbaum, for their valuable advice and help on so many organizational matters.

I hope that you will enjoy the program and interactions at ICS 2025!

P. (Saday) Sadayappan
University of Utah
USA

Message from the Program Chairs

Welcome to the 39th ACM International Conference on Supercomputing (ICS 2025), held June 8–11, 2025, in Salt Lake City, USA. ICS continues its tradition as a premier forum for high-performance computing research. We are pleased to again convene *in person*, fostering vibrant discussion and collaboration across the community.

This year we used a two-cycle submission model, with back-to-back deadlines in January and February 2025. We received 320 submissions—178 in the first cycle and 142 in the second. The deadline for the second cycle was intentionally set before the decisions for the first cycle were released to prevent re-submission of the same papers to both cycles. In each cycle, the authors had a chance to submit a written rebuttal to the reviews prior to the commencement of the discussion period among the reviewers. We followed a strict double-blind review process, where the reviewers were unaware of the author identities and vice versa. Moreover, we retained anonymity throughout the online discussion period, where the reviewer names were not disclosed to the other reviewers.

Decisions were made through a combination of online discussion and virtual PC meetings via Zoom. During online discussion, unanimity among all of a paper’s reviewers was needed to decide if a paper was rejected or accepted. Moreover, for a paper to be accepted during the online discussion phase, the paper had to have at least one reviewer act as champion. Papers that had a champion, but for which a unanimous decision was not reached during online discussion, were discussed at the virtual PC meeting. At the PC meeting, each of the undecided papers was discussed by the present reviewers, followed by a vote where the vote of any missing reviewers was taken from their latest review score. If a unanimous decision was not reached at that time, each paper was brought to a vote of all Program Committee members that were attending the PC meeting, with a simple majority of the attendees determining the final outcome. We accepted 83 papers in total (50 from the first cycle and 33 from the second), yielding an overall acceptance rate of 26%.

The accepted papers span a broad array of topics—from architectures and systems to programming models and applications—reflecting the dynamic nature of supercomputing research. This largest-ever ICS technical program comprises 19 sessions, including a *Best Papers session* that highlights work selected based on the highest review scores and reviewer endorsements for best paper consideration.

We express our gratitude to the 122 members of the Program Committee and the nine external reviewers for their expert evaluations. We are especially thankful to the 12 Program Committee members who served as shepherds for 14 conditionally accepted papers and the seven Program Committee members who agreed to participate on the best-paper selection committee. We are grateful to Saday Sadayappan (University of Utah), our General Chair, for taking care of all the conference organization details beyond paper selection. We thank Yuke Wang (Rice University) for liaising between the authors of the large number of accepted papers and our publisher. Finally, we would like to thank Xinning Hui and Ruifeng Zhang, graduate students from NCSU, who assisted us in running the Zoom virtual PC meetings.

We hope you find ICS 2025 intellectually rewarding and that it inspires your future work. Thank you for joining us in Salt Lake City.

Michael Ferdman (Stony Brook University)
Xipeng Shen (North Carolina State University)
Program Co-Chairs, ICS 2025

ICS 2025 Organizing Committee

General Chair

P. (Saday) Sadayappan University of Utah

Program co-chairs

Michael Ferdman Stony Brook University

Xipeng Shen North Carolina State University

Finance chair

Hari Sundar Tufts University

Publication chair

Yuke Wang Rice University

Publicity chair

Wenqian Dong Oregon State University

Web liaison

Yufan Xu Uber Technologies Inc.

Workshops and tutorials co-chairs

Mehmet Belviranli Colorado School of Mines

Local Arrangement

Varun Shankar University of Utah

ICS 2025 Program Committee

Almutaz Adileh	Huawei
Varun Agrawal	Advanced Micro Devices Inc
Alaa Alameldeen	Simon Fraser University
Rachata Ausavarungnirun	Mangoboost Inc.
Amro Awad	University of Oxford
Saurabh Bagchi	Purdue University
Rajeev Balasubramonian	University of Utah
Kevin Barker	Pacific Northwest National Laboratory
Tekin Bicer	Argonne National Laboratory
Mehmet Belviranli	Colorado School of Mines
Martin Bartscher	Texas State University
Suren Byna	The Ohio State University
Franck Cappello	Argonne National Laboratory
Trevor E. Carlson	National University of Singapore
Vito Giovanni Castellana	Pacific Northwest National Laboratory
Sunita Chandrasekaran	University of Delaware
Mainak Chaudhuri	Indian Institute of Technology Kanpur
Guoyang Chen	Apple Inc.
Jou-Ann Chen	Qualcomm
Wenguang Chen	Tsinghua University
Quan Chen	Shanghai Jiao Tong University
Zizhong Chen	UC Riverside
Preyesh Dalmia	NVIDIA
Sina Darabi	Università della Svizzera italiana
Sheng Di	Argonne National Laboratory
Yufei Ding	UCSD
Chen Ding	University of Rochester
Christian Engelmann	Oak Ridge National Laboratory
Simon Garcia de Gonzalo	Sandia National Laboratories
Jayesh Gaur	Intel Labs
Rong Ge	Clemson University
R Govindarajan	Indian Institute of Science
Hui Guan	UMass Amherst
Kyle Hale	Oregon State University
Martin Herbordt	Boston University
Jin Huang	Brookhaven National Lab
Nikhil Jain	Nvidia
Ali Jannesari	Iowa State University
Peng Jiang	University of Iowa
Daniel Jiménez	Texas A&M University

Mahmut Kandemir	Pennsylvania State University
David Kaeli	Northeastern University
Sudarsun Kannan	Rutgers University
Ahmad Maroof Karimi	Oak Ridge National Laboratory
Omer Khan	University of Connecticut
John Kim	KAIST
Youngsok Kim	Yonsei University
Milind Kulkarni	Purdue University
Hyounjun Kwon	University of California, Irvine
Ignacio Laguna	Lawrence Livermore National Laboratory
Zhiling Lan	University of Illinois Chicago
Dongyoon Lee	Stony Brook University
Gushu Li	University of Pennsylvania
Jiajia Li	NCSU
Dong Li	UC Merced
Lingda Li	Brookhaven National Laboratory
Ang Li	PNNL and UW
Chao Li	Shanghai Jiao Tong University
Guanpeng Li	University of Iowa
Chunhua Liao	Lawrence Livermore National Laboratory
Mieszko Lis	University of British Columbia
Xu Liu	Google
Jinyang Liu	University of Houston
Liu Liu	Rensselaer Polytechnic Institute
Pejman Lotfi-Kamran	Institute for Research in Fundamental Sciences
Jason Lowe-Power	University of California, Davis
Jason Mars	University of Michigan
Marco Minutoli	Pacific Northwest National Laboratory
William Moses	UIUC/Google
Abdullah Muzahid	Texas A&M University
Bogdan Nicolae	Argonne National Laboratory
Dimitrios S. Nikolopoulos	Virginia Tech
Soner Onder	Michigan Tech
Dhabaleswar K Panda	The Ohio State University
Prashant Pandey	Northeastern University
Tirthak Patel	Rice University
Suchita Pati	AMD Research & Advanced Development
Arnab K. Paul	BITS Pilani, Goa Campus
Jacques Pienaar	Google
Aleksandar Prokopec	Oracle Labs
Moinuddin Qureshi	Georgia Tech
Yihui Ren	Brookhaven National Laboratory

Tahsin Reza	University of Waterloo
Bin Ren	William & Mary
John (Jack) Sampson	Penn State
Bertil Schmidt	JGU Mainz
Joseph Schuchart	Stony Brook University
Martin Schulz	Technical University of Munich
Prateek Sharma	Indiana University
Seunghee Shin	State University of New York at Binghamton
Arrvindh Shriraman	Simon Fraser University
Yan Solihin	UCF
Ravi Soundararajan	VMware by Broadcom
Aravind Sukumaran Rajam	Meta
Hsin-Hsuan Sung	Qualcomm
Guangming Tan	Chinese Academy of Sciences
Kenjiro Taura	The University of Tokyo
Antonino Tumeo	Pacific Northwest National Laboratory
Nilay Vaish	Google
Hans Vandierendonck	Queens University Belfast
Ashish Venkat	University of Virginia
Guru Venkataramani	George Washington University
Chen Wang	Lawrence Livermore National Laboratory
Bo Wu	Colorado School of Mines
Yuanchao Xu	University of California, Santa Cruz
Helen Xu	Georgia Tech
Wei Xue	Tsinghua University
Feng Yan	University of Houston
Jun Yang	University of Pittsburgh
Yifan Yang	Nvidia
Chencheng Ye	Huazhong University of Science and Technology
Pen-Chung Yew	University of Minnesota at Twin Cities
Youngmin Yi	Sogang University
Cliff Young	Google DeepMind
Xiaodong Yu	Stevens Institute of Technology
Jianping Zeng	Samsung MSL
Jidong Zhai	Tsinghua University
Youtao Zhang	University of Pittsburgh
Zheng Zhang	Rutgers University
Feng Zhang	Renmin University of China
Zhijia Zhao	University of California, Riverside
Keren Zhou	George Mason University

ICS 2025 External Reviewers

Aamer Jaleel	Nvidia
Aparna Chandramowliswaran	UC Irvine
Bilge Acun	Meta
Boris Grot	University of Edinburgh
Frank Mueller	North Carolina State University
Jiesong Liu	North Carolina State University
Michela Becchi	North Carolina State University
Nael Abu-Ghazaleh	University of California, Riverside
Natalie Enright Jerger	University of Toronto
Nengkun Yu	Stony Brook University
Nikos Hardavellas	Northwestern
Rajiv Gupta	University of California, Riverside
Robert Underwood	Argonne National Laboratory
Swamit Tannu	University of Wisconsin–Madison
Tony (Tong) Geng	University of Rochester
Wei Niu	University of Georgia
Zhenhua Liu	Stony Brook University

ICS 2025 Steering Committee

Alex Nicolau	University of California Irvine
Alex Veidenbaum	University of California Irvine
Ana Lucia Varbanescu	University of Twente
Avi Mendelson	Technion
Bill Gropp	University of Illinois
Chen Ding	University of Rochester
Constantine Polychronopoulos	Juniper Networks
Dimitrios Nikolopoulos (Chair)	VTU
Dionisios Pnevmatikatos	NTUA
Eduard Ayguade	Barcelona Supercomputing Center
Fran Cazorla	Technical University of Catalonia/BSC
Frank Mueller	NCSU
Fred Chong	University of Chicago
Harry Wijshoff	Leiden University
Huiyang Zhou	NCSU
James Goodman	University of Wisconsin–Madison
Jose Moreira	IBM
Kenji Kise	Tokyo Institute of Technology

Kirk Cameron	VTU
Kyle Gallivan	Florida State University
Lawrence Rauchwerger	University of Illinois
Mahmut Kandemir	Pennsylvania State University
Mateo Valero	Technical University of Catalonia/BSC
Michael Gschwind	Facebook
Murali Annavaram	University of Southern California
Pete Beckman	Argonne Natl Lab
Peter Hofstee	IBM and TU Delft
Rosa Badia	Barcelona Supercomputing Center
Rudi Eigenman	University of Delaware
Sally A. McKee	Clemson
Valentina Salapura	Google
John Sopka	Independent consultant
Wen-mei Hwu	Nvidia
Yoav Etsion	Technion
Zhiyuan Li	Purdue

Contents

Sponsors	iii
Message from the General Chair	iv
Message from the Program Chairs	v
ICS 2024 Organizing Committee	vii
ICS 2024 Program Committee	viii
ICS 2024 Steering Committee	xi

Session: Approximation

[SYprox: Combining Host and Device Perforation with Mixed Precision Approximation on Heterogeneous Architectures](#)

Lorenzo Carpentieri (University of Salerno), Biagio Cosenza (University of Salerno)

[BitWeaver: Read-Time Truncation in Memory](#)

Garrett Gagnon (Samsung Semiconductor US, Rensselaer Polytechnic Institute), Srikanth Malla (Samsung Semiconductor US), Yangwook Kang (Samsung Semiconductor US), Liu Liu (Rensselaer Polytechnic Institute)

[NeurLZ: An Online Neural Learning-based Method to Enhance Scientific Lossy Compression](#)

Wenqi Jia (University of Texas at Arlington), Zhewen Hu (Texas A&M University), Youyuan Liu (Temple University), Boyuan Zhang (Indiana University), Jinzhen Wang (UNC Charlotte), Jinyang Liu (University of Houston), Wei Niu (University of Georgia), Stavros Kalafatis (Texas A&M University), Junzhou Huang (University of Texas at Arlington), Sian Jin (Temple University), Daoce Wang (Indiana University), Jiannan Tian (University of Kentucky), Miao Yin (University of Texas at Arlington)

[ghZCCL: Advancing GPU-aware Collective Communications with Homomorphic Compression](#)

Jiajun Huang (University of South Florida), Sheng Di (Argonne National Laboratory), Yafan Huang (University of Iowa), Zizhong Chen (University of California Riverside), Franck Cappello (Argonne National Laboratory), Yanfei Guo (Argonne National Laboratory), Rajeev Thakur (Argonne National Laboratory)

Session: Graph Neural Networks

[Scaling Large-scale GNN Training to Thousands of Processors on CPU-based Supercomputers](#)

Chen Zhuang (Institute of Science Tokyo, Riken Center for Computational Science), Lingqi Zhang (RIKEN Center for Computational Science), Du Wu (Institute of Science Tokyo, RIKEN Center for Computational Science), Peng Chen (RIKEN Center for Computational Science), Jiajun Huang (University of South Florida), Xin Liu (National Institute of Advanced Industrial Science & Technology), Rio Yokota (Institute of Science Tokyo), Nikoli Dryden (Lawrence Livermore National Laboratory), Toshio Endo (Institute of Science Tokyo), Satoshi Matsuoka (RIKEN Center for Computational Science, Institute of Science Tokyo), Mohamed Wahib (RIKEN Center for Computational Science)

[CoLa: Towards Communication-efficient Distributed Sparse Matrix-Matrix Multiplication on GPUs](#)

Lixing Zhang (Beijing University of Posts and Telecommunications), Yingxia Shao (Beijing University of Posts and Telecommunications), Shigang Li (Beijing University of Posts and Telecommunications)

Cherry: Breaking the GPU Memory Wall for Large-Scale GNN Training via Micro-Batching

Yan Wang (Guangzhou Institute of Technology, Xidian University), Qinghua Guo (Guangzhou Institute of Technology, Xidian University), Haoran Kong (Chinese Academy of Sciences), Kai Sheng (Guangzhou Institute of Technology, Xidian University), Zhen Xie (Binghamton University), Hao Chen (College of Computer Science and Electronic Engineering, Hunan University), Weile Jia (Chinese Academy of Sciences), Dingwen Tao (Chinese Academy of Sciences), Xin He (Guangzhou Institute of Technology, Xidian University)

Fused3S: Fast Sparse Attention on Tensor Cores

Zitong Li (University of California, Irvine), Aparna Chandramowlishwaran (University of California, Irvine)

Session: Sparse Linear Algebra

StructLU: Dependency-Preserving Incomplete LU with Hierarchical Parallelism for Structured Grid PDEs on GPUs

Hao Luo (Peking University), Qianchao Zhu (Peking University), Xiaochen Hao (School of Computer Science, Peking University), Chunxi Lei (Peking University), Chengdi Ma (Peking University), Chenchen Zhang (Peking University), Yun Liang (School of Integrated Circuits, Peking University), Chao Yang (Peking University, PKU-Changsha Institute for Computing and Digital Economy)

IA-Chol: Input-Aware Cholesky Decomposition on CPU and GPU

Jixiao Deng (National University of Defense Technology), Qinglin Wang (National University of Defense Technology), Lin Chen (National University of Defense Technology), Bo Yang (National University of Defense Technology), Xinhai Chen (National University of Defense Technology), Jie Liu (National University Of Defense Technology)

CB-SpMV: A Data Aggregating and Balance Algorithm for Cache-Friendly Block-Based SpMV on GPUs

Xing Cong (Beihang University), FuKai Sun (Beihang University), YiFan Chen (Beihang University), Chenhao Xie (Beihang University), Yi Liu (Beihang University), Depei Qian (Beihang University)

HR-SpMM: Adaptive Row Partitioning and Hybrid Kernel Design for Sparse Matrix Multiplication

Qi Wang (Southwest University of Science and Technology), Yaobin Wang (Southwest University of Science and Technology), Yi Luo (Southwest University of Science and Technology), Rong Luo (Southwest University of Science and Technology), Pingping Tang (Southwest University of Science and Technology)

Session: Acceleration

G^3SA: A GPU-Accelerated Gold Standard Genomics Library for End-to-End Sequence Alignment

Yeejoo Han (Seoul National University), Sunwoo Kim (Seoul National University), Seongyeon Park (Seoul National University), Jinho Lee (Seoul National University)

Graph Convolutional Network Acceleration Using Adiabatic Superconductor Josephson Devices

Zhengang Li (Northeastern University), Hongwu Peng (University of Connecticut), Xuan Shen (Northeastern University), Masoud Zabihi (Northeastern University), Xi Xie (University of Connecticut), Geng Yuan (University of Georgia), Yanzhi Wang (Northeastern University), Olivia Chen (Kyushu University), Caiwen Ding (University of Minnesota Twin Cities)

TMMModel: Modeling Texture Memory and Mobile GPU Performance to Accelerate DNN Computations

Jiexiong Guan (University of Thessaly, William & Mary), Zhenqing Hu (William & Mary), Christos D. Antonopoulos (University of Thessaly), Nikolaos Bellas (University of Thessaly), Spyros Lalis (University of Thessaly), Evgenia Smirni (William & Mary), Gang Zhou (William & Mary), Gagan Agrawal (University of Georgia), Bin Ren (William & Mary)

DR-CircuitGNN: Training Acceleration of Heterogeneous Circuit Graph Neural Network on GPUs

Yuebo Luo (University of Minnesota Twin Cities), Shiyang Li (University of Minnesota Twin Cities, Nankai University), Junran Tao (Stevens Institute of Technology), Kiran Gautam Thorat (University of Connecticut), Xi Xie (University of Connecticut), Hongwu Peng (University of Connecticut), Nuo Xu (University of Minnesota Twin Cities), Caiwen Ding (University of Minnesota Twin Cities), Shaoyi Huang (Stevens Institute of Technology)

Session: Applications

CLOVER: Spatio-graph-based kNN on the GPU

Victor Kamel (University of Toronto), Hanxueyu Yan (University of Victoria), Sean Chester (University of Victoria)

Efficient Locality-aware Instruction Stream Scheduling for Stencil Computation on ARM Processors

Shanghao Liu (Beihang University), Hailong Yang (Beihang University), Xin You (Beihang University), Zhongzhi Luan (Beihang University), Yi Liu (Beihang University), Depei Qian (Beihang University)

Accelerating Complex Stencil Computations with Adaptive Fusion Strategy

Siqi Wang (Beihang University), Hailong Yang (Beihang University), Pengbo Wang (Beihang University), Shaokang Du (Beihang University), Yufan Xu (Independent Researcher), Qingxiao Sun (China University of Petroleum, Beijing), Xiaoyan Liu (Beihang University), Xuezhu Wang (Beihang University), Xuning Liang (Beihang University), Zhongzhi Luan (Beihang University), Yi Liu (Beihang University), Depei Qian (Beihang University)

A3FR: Agile 3D Gaussian Splatting with Incremental Gaze Tracked Foveated Rendering in Virtual Reality

Shuo Xin (Physics, Stanford University), Haiyu Wang (Tandon School of Engineering, New York University), Sai Qian Zhang (Tandon School of Engineering, New York University)

EPIClear: Exploiting Domain-Specific Features for Epistasis Detection Acceleration on Tensor Cores

Ricardo Nobre (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa), Miguel Graça (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa), Leonel Sousa (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa), Aleksandar Ilic (INESC-ID, Instituto Superior Técnico, Universidade de Lisboa)

Statistical Treatment of Variable MPI Latencies and MPI-Communication Hiding for Matrix-Free Finite Element Operators

Max Heldman (Virginia Tech), Johann Rudi (Virginia Tech), Julie Bessac (Computational Science Center, National Renewable Energy Laboratory)

Session: GPU Scheduling

Fast and Fair Training for Deep Learning in Heterogeneous GPU Clusters

Zizhao Mo (University of Macau), Huanle Xu (University of Macau), Wing Cheong Lau (The Chinese University of Hong Kong)

SortingHat: System Topology-aware Scheduling of Deep Neural Network Models on Multi-GPU Systems

Seok Namkoong (Yonsei University), Taehyeong Park (Yonsei University), Kiung Jung (Yonsei University), Jinyoung Kim (Yonsei University), Yongjun Park (Yonsei University)

CTCCL: Cost-Efficient Joint Device-Network Load Balancing for LLM Training in RoCE-based Intelligent Computing Network

Zhuotong Li (State Cloud, China Telecom), Liang Xu (State Cloud, China Telecom), Ziqi Huang (State Cloud, China Telecom), Shuyun Qian (State Cloud, China Telecom), Hongwei Bu (State Cloud, China Telecom), Ming Yang (State Cloud, China Telecom), Mengyun Luan (State Cloud, China Telecom), Weiguo Chen (State Cloud, China Telecom), Xu Wen (State Cloud, China Telecom)

Cephalo: Harnessing Heterogeneous GPU Clusters for Training Transformer Models

Runsheng Benson Guo (Cheriton School of Computer Science, University of Waterloo), Utkarsh Anand (Cheriton School of Computer Science, University of Waterloo), Arthur Chen (Cheriton School of Computer Science, University of Waterloo), Khuzaima Daudjee (Cheriton School of Computer Science, University of Waterloo)

A Device-Side Execution Model for Multi-GPU Task Graphs

Ilyas Turimbetov (Koç University), Mohamed Wahib (RIKEN Center for Computational Science), Didem Unat (Koç University)

Session: Solvers & Sparsity

CRAMG: A Communication-Reduced Algebraic Multigrid Method

Fan Yuan (School of Mathematics and Computer Science, Xiangtan University), Xiaojian Yang (National University of Defense Technology), Yunqing Huang (School of Mathematics and Computer Science, Xiangtan University), Dezun Dong (National University of Defense Technology), Chuanfu Xu (National University of Defense Technology), Jie Liu (National University of Defense Technology), Xiaoqiang Yue (School of Mathematics and Computer Science, Xiangtan University), Shengguo Li (National University of Defense Technology), Hongxia Wang (National University of Defense Technology)

An Efficient 2D Fusion Method for High-Performance Two-Stage Eigensolvers on Modern Heterogeneous Architectures

Yongxiao Zhou (Tsinghua University), Yi Zong (Tsinghua University), Yuyang Jin (Tsinghua University), Heng Li (Tsinghua University), Wei Xue (Tsinghua University, Qinghai University)

SnuSolver: Optimizing Sparse Direct Solvers for Heterogeneous Systems

Chaewon Kim (Seoul National University), Jaehwan Lee (Seoul National University), Jinpyo Kim (Seoul National University), Dohyun Kim (Seoul National University), Kyusu Ahn (Seoul National University), Hyung Uk Cho (Samsung Display Co., Ltd.), Seungin Baek (Samsung Display Co., Ltd.), Jaejin Lee (Seoul National University)

MAGNUS: Generating Data Locality to Accelerate Sparse Matrix-Matrix Multiplication on CPUs

Jordi Wolfson-Pou (Intel Labs), Jan Laukemann (Friedrich-Alexander-Universität Erlangen-Nürnberg), Fabrizio Petrini (Intel Labs)

Session: Processing-in-Memory

PIM-CARE: A Compiler-Assisted Dynamic Resource Allocation Framework for Real-world DRAM PIM

Inyong Hwang (Yonsei University), Donghyeon Kim (Hanyang University), Seokwon Kang (Yonsei University), Taehyeong Park (Yonsei University), Taehoon Kim (Hanyang University), Jiwon Seo (Seoul National University), Hanjun Kim (Yonsei University), Youngsok Kim (Yonsei University), Yongjun Park (Yonsei University)

Proteus: Achieving High-Performance Processing-Using-DRAM via Dynamic Precision Bit-Serial Arithmetic

Geraldo Francisco de Oliveira Junior (ETH Zurich), mayank kabra (International institute of information technology Bangalore), Yuxin Guo (Cambridge University), Kangqi Chen (ETH Zurich), Abdullah Giray Yaglikci (ETH Zurich), Melina Soysal (ETH Zurich), Mohammad Sadrosadati (ETH Zurich), Joaquin Olivares Bueno (Universidad de Córdoba), Saugata Ghose (University of Illinois Urbana-Champaign), Juan Gomez Luna (NVIDIA), Onur Mutlu (ETH Zurich)

SparsePIM: An Efficient HBM-Based PIM Architecture for Sparse Matrix-Vector Multiplications

Taewoon Kang (Korea University), Geonwoo Choi (Korea University), Taeweon Suh (Korea University), Gunjae Koo (Korea University)

MARS: Processing-In-Memory Acceleration of Raw Signal Genome Analysis Inside the Storage Subsystem

Melina Soysal (ETH Zurich), Konstantina Koliogeorgi (ETH Zurich), Can Firtina (ETH Zurich), Nika Mansouri Ghiasi (ETH Zurich), Rakesh Nadig (ETH Zurich), Haiyu Mao (ETH Zurich), Geraldo Francisco de Oliveira Junior (ETH Zurich), Yu Liang (ETH Zurich), Klea Zambaku (ETH Zurich Bilkent University), Mohammad Sadrosadati (ETH Zürich), Onur Mutlu (ETH Zurich)

Session: Efficiency

DALdex: A DPU-Accelerated Persistent Learned Index via Incremental Learning

Aoyang Tong (Huazhong University of Science and Technology), Yu Hua (Huazhong University of Science and Technology), Menglei Chen (Huazhong University of Science and Technology)

From Islands to Archipelago: Towards Collaborative and Adaptive Burst Buffer for HPC Systems

Mingtian Shao (National University of Defense Technology), Ruibo Wang (National University of Defense Technology), Wenzhe Zhang (National University of Defense Technology), Kai Lu (National University of Defense Technology), Yiqin Dai (National University of Defense Technology), Huijun Wu (National University of Defense Technology)

PIE: Enabling Fast and Scalable Incremental Evolving Graph Analytics on Persistent Memory

Yunmo Zhang (City University of Hong Kong), Jiacheng Huang (City University of Hong Kong), Xizhe Yin (University of California Riverside), Junqiao Qiu (City University of Hong Kong), Hong Xu (The Chinese University of Hong Kong), Chun Jason Xue (MBZUAI)

DEDUPKV: A Space-Efficient and High-Performance Key-Value Store via Fine-Grained Deduplication

Safdar Jamil (Sogang University), Awais Khan (Oak Ridge National Lab), Xubin He (Temple University), Youngjae Kim (Sogang University)

Session: Optimizing Compilation

ConTraPh: Contrastive Learning for Parallelization and Performance Optimization

Quazi Ishtiaque Mahmud (Iowa State University), Ali TehraniJamsaz (Iowa State University), Nesreen K. Ahmed (Cisco AI Research), Theodore L. Willke (DataStax), Ali Jannesari (Iowa State University)

UJOpt: Heuristic Approach for Applying Unroll-and-Jam Optimization and Loop Order Selection

Shilpa Babalad (Indian Institute of Science), Shirish K Shevade (Indian Institute of Science), Matthew Jacob Thazhuthaveetil (Indian Institute of Science), R Govindarajan (Indian Institute of Science)

Loop Fusion in Matrix Multiplications with Sparse Dependence

Mohammad Mehdi Salehi (McMaster University), Kazem Cheshmi (McMaster University)

ConCo: Optimizing Compilation of Concurrent Tensor Programs on Shared GPU

Jiamin Lu (University of Science and Technology of China), Jingwei Sun (University of Science and Technology of China), Yunlong Xu (Independent Researcher), Peng Sun (Independent Researcher), Guangzhong Sun (University of Science and Technology of China)

Session: Best Papers

Pushing the Limits of GPU Lossy Compression: A Hierarchical Delta Approach

Boyuan Zhang (Indiana University), Yafan Huang (University of Iowa), Sheng Di (Argonne National Laboratory), Fengguang Song (Indiana University), Guanpeng Li (University of Iowa), Franck Cappello (Argonne National Laboratory)

Parallel Contraction Hierarchies Can Be Efficient and Scalable

Zijin Wan (University of California, Riverside), Xiaojun Dong (University of California, Riverside), Letong Wang (University of California, Riverside), Enzuo Zhu (University of California, Davis), Yan Gu (University of California, Riverside), Yihan Sun (University of California, Riverside)

BMQSim: Overcoming Memory Constraints in Quantum Circuit Simulation with a High-Fidelity Compression Framework

Boyuan Zhang (Indiana University), Bo Fang (Pacific Northwest National Laboratory), Fanjiang Ye (Indiana University), Luanzheng Guo (Pacific Northwest National Laboratory), Fengguang Song (Indiana University), Nathan Tallent (Pacific Northwest National Laboratory), Dingwen Tao (Indiana University)

DIV: An Index & Value compression method for SpMV on large matrices

Dimitrios Galanopoulos (National Technical University of Athens), Panagiotis Mpakos (National Technical University of Athens), Petros Anastasiadis (National Technical University of Athens), Nectarios Koziris (National Technical University of Athens), Georgios Goumas (National Technical University of Athens)

DIMPLES: Distributed Influence Maximization for Pandemic pLanning on Exascale Systems

Marco Minutoli (Pacific Northwest National Laboratory), Reece Neff (North Carolina State University), Naw Safrin Sattar (Oak Ridge National Laboratory), Hao Lu (Oak Ridge National Laboratory), John Feo (Pacific Northwest National Laboratory), Henning Mortveit (University of Virginia), Anil Vullikanti (University of Virginia), Dawen Xie (University of Virginia), Mandy L Wilson (University of Virginia), Gregor von Laszewski (University of Virginia), Parantapa Bhattacharya (University of Virginia), S M Ferdous (Pacific Northwest National Laboratory), Ananth Kalyanaraman (Washington State University), Michela Becchi (North Carolina State University), Madhav Marathe (University of Virginia), Mahantesh Halappanavar (Pacific Northwest National Laboratory)

Light-FP: Analyze Floating-Point Error in a Highly Condensed Approach

Jiazhi Mi (Chinese Academy of Sciences), Li Chen (Chinese Academy of Sciences), Haoyu Wang (Chinese Academy of Sciences), Ruixiang Gao (Shandong University of Science and Technology), Hongze Zhang (Shandong University of Science and Technology), Ronghong Shen (Chinese Academy of Sciences), Kai Lin (Beijing Institute of Technology), You Fu (Shandong University of Science and Technology), Huimin Cui (Chinese Academy of Sciences)

Session: Performance Analysis

WisIO: Automated I/O Bottleneck Detection with Multi-Perspective Views for HPC Workflows

Izzet Yildirim (Illinois Institute of Technology), Hariharan Devarajan (Lawrence Livermore National Laboratory), Anthony Kougkas (Illinois Institute of Technology), Xian-He Sun (Illinois Institute of Technology), Kathryn Mohror (Lawrence Livermore National Laboratory)

Efficient Server Consolidation through a balanced mix of Transformer-based and Conventional Applications

Pablo Abad (Universidad de Cantabria), Pablo Prieto (Universidad de Cantabria), Valentin Puente (Universidad de Cantabria), Jose Angel Gregorio (Universidad de Cantabria)

Taking GPU Programming Models to Task for Performance Portability

Joshua Hoke Davis (University of Maryland), Pranav Sivaraman (University of Maryland), Joy Kitson (University of Maryland), Konstantinos Parasyris (Lawrence Livermore National Laboratory), Harshitha Menon (Lawrence Livermore National Laboratory), Isaac Minn (University of Maryland), Giorgis Georgakoudis (Lawrence Livermore National Laboratory), Abhinav Bhatele (University of Maryland)

Analyzing the Performance of Applications at Exascale

Dragana Grbic (Rice University), John Mellor-Crummey (Rice University)

Session: Heterogeneity

Understanding the Idiosyncrasies of Emerging BlueField DPUs

Arjun Kashyap (University of California, Merced), Yuke Li (University of California, Merced), Darren Ng (University of California, Merced), Xiaoyi Lu (University of California, Merced)

Multi-node Multi-GPU Datalog

Ahmedur Rahman Shovon (University of Illinois Chicago), Yihao Sun (Syracuse University), Kristopher Micinski (Syracuse University), Thomas Gilray (Washington State University), Sidharth Kumar (University of Illinois Chicago)

SmartNIC-GPU-CPU Heterogeneous System for Large Machine Learning Model with Software-Hardware Codesign

Anqi Guo (Boston University), Yuchen Hao (Meta Platforms), Xiteng Yao (Boston University), Shining Yang (Boston University), Jianyu Huang (Meta Platforms), Tony (Tong) Geng (University of Rochester), Martin Herbordt (Boston University)

D-Rex: Heterogeneity-Aware Reliability Framework and Adaptive Algorithms for Distributed Storage

Maxime Gonthier (University of Chicago, Argonne National Laboratory), Dante D. Sanchez-Gallegos (University Carlos III of Madrid), Haochen Pan (University of Chicago), Bogdan Nicolae (Argonne National Laboratory), Sicheng Zhou (Southern University of Science and Technology), Hai Duc Nguyen (University of Chicago, Argonne National Laboratory), Valerie Hayot-Sasson (University of Chicago, Argonne National Laboratory), J. Gregory Pauloski (University of Chicago), Jesus Carretero (University Carlos III of Madrid), Kyle Chard (University of Chicago, Argonne National Laboratory), Ian Foster (University of Chicago, Argonne National Laboratory)

Session: Resource Management

ORION: Optimizing OLAP Query Execution with Proactive Caching and Separate Operators

Zhixin Tong (Shanghai Jiao Tong University), Jiuchen Shi (Shanghai Jiao Tong University, The Hong Kong Polytechnic University), Quan Chen (Shanghai Jiao Tong University), Pu Pang (Shanghai Jiao Tong University), Shixuan Sun (Shanghai Jiao Tong University), Jie Meng (Huawei Cloud), Jiang Liu (Huawei Cloud), En Shao (Chinese Academy of Sciences), Minyi Guo (Shanghai Jiao Tong University)

ORA: Job Runtime Prediction for High-Performance Computing Platforms Using the Online Retrieval-Augmented Language Model

Hongyi Liu (Peking University), Yinping Ma (Peking University), Xiaosong Huang (Peking University), Lingzhe Zhang (Peking University), Tong Jia (Peking University, National Key Laboratory of Data Space Technology and System), Ying Li (Peking University)

Generating Microservice Graphs with Production Characteristics for Efficient Resource Scaling

Fanrong Du (Shanghai Jiao Tong University), Jiuchen Shi (Shanghai Jiao Tong University, The Hong Kong Polytechnic University), Quan Chen (Shanghai Jiao Tong University), Pu Pang (Shanghai Jiao Tong University), Li Li (Shanghai Jiao Tong University), Minyi Guo (Shanghai Jiao Tong University)

HARNESS: Holistic Resource Management for Diversely Scaled Edge Cloud Systems

Ismet Dagli (Colorado School of Mines), Justin Davis (Colorado School of Mines), Mehmet Esat Belviranli (Colorado School of Mines)

Session: Code Optimization

Leonid: Exploring Automated Kernel Fusion in Performance-Portable Programming Models for Scientific Computation

Chenchen Zhang (Peking University), Hao Luo (Peking University), Chao Yang (Peking University)

DeCOS: Data-Efficient Reinforcement Learning for Compiler Optimization Selection Ignited by LLM

Tianming Cui (University of Minnesota), Pen-Chung Yew (University of Minnesota), Stephen McCamant (University of Minnesota), Antonia Zhai (University of Minnesota)

Pearl: Automatic Code Optimization Using Deep Reinforcement Learning

Djamel Rassem Lamouri (New York University Abu Dhabi), Iheb Nassim Aouadj (New York University Abu Dhabi), Smail Kourta (New York University Abu Dhabi), Riyadh Baghdadi (New York University Abu Dhabi)

CIEplorer: Microarchitecture-Aware Exploration for Tightly Integrated Custom Instruction

Xiaoyu Hao (University of Science and Technology of China), Sen Zhang (University of Science and Technology of China), Liang Qiao (University of Science and Technology of China), Qingcai Jiang (University of Science and Technology of China), Jun Shi (University of Science and Technology of China), Junshi Chen (University of Science and Technology of China, Laoshan Laboratory), Hong An (University of Science and Technology of China, Laoshan Laboratory), Xulong Tang (University of Pittsburgh), Hao Shu (NIO), Honghui Yuan (NIO)

Session: Energy & Servers

EVeREST-C: An Effective and Versatile Runtime Energy Saving Tool for CPUs

Anna Yue (University of Minnesota), Pen-Chung Yew (University of Minnesota), Sanyam Mehta (Hardware-software Codesign, Hewlett-Packard Labs)

EDAN: Towards Understanding Memory Parallelism and Latency Sensitivity in HPC

Siyuan Shen (ETH Zürich), Mikhail Khalilov (ETH Zürich), Lukas Gianinazzi (ETH Zürich), Timo Schneider (ETH Zürich), Marcin Chrapek (ETH Zürich), Jai Dayal (Cerebras Systems), Manisha Gajbe, Robert Wisniewski (Hewlett Packard Enterprise), Torsten Hoeftler (ETH Zürich)

ROCKET: An RNS-based Photonic Accelerator for High-Precision and Energy-Efficient DNN Training

Hao Zhang (University of Otago), Haibo Zhang (University of Otago), Chengpeng Xia (University of Otago), Zhiyi Huang (University of Otago), Yawen Chen (University Of New South Wales), Amanda Barnard (Australian National University)

A Global Perspective on Supercomputer Power Provisioning: Case Studies from United States and Europe

Tapasya Patki (Lawrence Livermore National Laboratory), Barry Rountree (Lawrence Livermore National Laboratory), Torsten Wilde (Hewlett-Packard Enterprise), Andrea Bartolini (University of Bologna), Stephanie Brink (Lawrence Livermore National Laboratory), Esa Heiskanen (CSC IT Center for Science Ltd.), Sachin Idgunji (NVIDIA Corporation), Matthias Maiterth (Oak Ridge National Laboratory), James Rogers (Oak Ridge National Laboratory), Ermal Rrapaj (Lawrence Berkeley National Laboratory), Ralf Schneider (HLRS High Performance Computing Center Stuttgart), Woong Shin (Oak Ridge National Laboratory), Kathleen Shoga (Lawrence Livermore National Laboratory), Christian Simmendinger (Hewlett-Packard Enterprise), Nicholas J. Wright (Lawrence Berkeley National Laboratory), Zhengji Zhao (Lawrence Berkeley National Laboratory)

Session: Potpourri

PortFC: Designing High-performance Deadlock-free BCube Networks

Peirui Cao (Nanjing University), Rui Ning (Nanjing University), Hongwei Yang (China Mobile), Zhaochen Zhang (Nanjing University), Chang Liu (Nanjing University), Rui Li (Nanjing University), Yongqi Yang (Nanjing University),

Yunzhuo Liu (Nanjing University), Chengyuan Huang (Nanjing University), Tao Sun (China Mobile), Xiaodong Duan (China Mobile), Guihai Chen (Nanjing University), Chen Tian (Nanjing University)

Auto-Healer: Self-Healing Hardware for Perception Stage Faults in Autonomous Driving Systems

Ali Suvizi (George Washington University), Guru Venkataramani (George Washington University)

OpaQue: Program Output Obfuscation for Quantum Software Circuits in Quantum Clouds

Tirthak Patel (Rice University), Aditya Ranjan (Northeastern University), Daniel Silver (Northeastern University), Harshitta Gandhi (QBit Solutions Research), William Cutler (Oxford University), Devesh Tiwari (Northeastern University)

JBSA: A Bit-Serial Accelerator for Deep Neural Networks Using Superconducting SFQ Logic

Yang Su (ShanghaiTech University), Sheng Li (ShanghaiTech University), Huilong Jiang (Chinese Academy of Sciences), Hao-fei Yin (ShanghaiTech University), Rongliang Fu (The Chinese University of Hong Kong), Junying Huang (Chinese Academy of Sciences), Xiaochun Ye (Chinese Academy of Sciences), Zhimin Zhang (Chinese Academy of Sciences), Jie Ren (Chinese Academy of Sciences), Xiaoping Gao (Chinese Academy of Sciences), Tsung-Yi Ho (The Chinese University of Hong Kong), Dongrui Fan (Chinese Academy of Sciences)

Session: Graph Algorithms

YH-Light: Yielding Hierarchy-aware Partitioner for Large-scale Graph Processing

Xinbiao Gan (National University of Defense Technology), Tiejun Li (National University of Defense Technology), Chunye Gong (National University of Defense Technology), Jie Liu (National University of Defense Technology), Kai Lu (National University of Defense Technology)

MG- α GCD: Accelerating Graph Community Detection on Multi-GPU Platforms

Shuai Yang (Chinese Academy of Sciences), Changyou Zhang (Chinese Academy of Sciences)

GraCFL: A Versatile Vertex-Centric Graph System for High-Performance CFL Reachability Analysis

Sakib Fuad (University of California, Riverside), Amir Hossein Nodehi Sabet (University of California, Riverside), Umar Farooq (University of California, Riverside), Zhijia Zhao (University of California, Riverside)

OPMOS: Ordered Parallel Algorithm for Multi-Objective Shortest-Paths

Leo Gold (University of Connecticut), Adam Bienkowski (University of Connecticut), David Sidoti (US Naval Research Laboratory), Krishna Pattipati (University of Connecticut), Omer Khan (University of Connecticut)

A Multi-GPU Algorithm for Computing Maximal Independent Sets in Large Graphs

Anju Mongandampulath Akathoot (Texas State University), Benila Virgin Jerald Xavier (Texas State University), Martin Burtscher (Texas State University)

Session: Memory Systems

A Cost-Effective Dueling Framework for Set-Associative Cache Indexing

Kevin Weston (Texas A&M University), Vahid Janfaza (Texas A&M University), Avery Johnson (Texas A&M University), Abdullah Muzahid (Texas A&M University)

DREAM: Device-Driven Efficient Access to Virtual Memory

Nurlan Nazaraliyev (University of California, Riverside), Elaheh Sadredini (UC Riverside), Nael Abu-Ghazaleh (Computer Science and Engineering, University of California, Riverside)

Page Migration for Hardware Memory Disaggregation Across a Network

Archit Patke (University of Illinois at Urbana-Champaign), Christian Pinto (IBM Research Europe), Saurabh Jha (IBM Research), Haoran Qiu (University of Illinois at Urbana-Champaign), Zbigniew Kalbarczyk (University of Illinois at Urbana-Champaign), Ravishankar K. Iyer (University of Illinois at Urbana-Champaign)

MEMPLEX: A Memory System with Replication and Migration of Data for Multi-Chiplet NUMA Architectures

Neethu Bal Mallya (Chalmers University of Technology and University of Gothenburg), Bhavishya Goel (Chalmers University of Technology and University of Gothenburg), Ioannis Sourdis (Chalmers University of Technology and University of Gothenburg)

Persistent Memory Objects on the Cheap

Derrick Greenspan (University of Central Florida), Naveed Ul Mustafa (New Mexico State University), Jongouk Choi (University of Central Florida), Mark Heinrich (University of Central Florida), Yan Solihin (University of Central Florida)