2026 IEEE 26th International Symposium on Cluster, Cloud and Internet Computing (CCGrid 2026)

Sydney, Australia
18-21 May 2026

IEEE Catalog Number: CFP26276-POD
ISBN: 979-8-3315-7065-1

IEEE
Advancing Technology for Humanity

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:	CFP26276-POD
ISBN (Print-On-Demand):	979-8-3315-7065-1
ISBN (Online):	979-8-3315-7064-4
ISSN:	2376-4414

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

Kernel-Initiated One-Sided Networking for GPU-Accelerated AI Workloads 1
Khaled Hamidouche, John Bachan, Pak Markthub, Peter-Jan Gootzen, Elena Agostini, Sylvain Jeaugey Aamir Shafi, Georgios Theodorakis, Manjunath Gorentla Venkata

Mammoth: Macro-Level MPI Offloading to Off-Path Accelerator in DPU 11
Jong-Bin Lee, Pu-Rum Seo, Ki-Moon Jeong, Hyun-Wook Jin

Scalable, Topology- and Multi-HCA-Aware Hierarchical GPU Allgather Using Parallel Rings 22
Amirreza Barati Sedeh, Ryan Grant, Ahmad Afsahi

Characterizing LLM Inference Energy-Performance Tradeoffs Across Workloads and GPU Scaling 33
Paul Joe Maliakel, Shashikant Ilager, Ivona Brandic

Quicktopia: Iteration-Level GPU Frequency Control for Energy–Latency Co-Optimization in LLM Inference 44
Soyang Baek, Bodon Jeong, Hongsu Byun, Sungyong Park

ARKV: Adaptive and Resource-Efficient KV Cache Management Under Limited Memory Budget for Long-Context Inference in LLMs 54
Jianlong Lei, Shashikant Ilager

ElastiFlow: Elastic Resource Management for Iterative Scientific Workflows in Hybrid HPC-Cloud Infrastructure 64
Srishti Dasgupta, Kavitha Subramaniam, Michael Gerndt

Efficiently Reproducing Distributed Workflows in Notebook-based Systems 74
Talha Azaz, Raza Ahmad, Md Saiful Islam, Douglas Thain, Tanu Malik

Compass: Optimizing Compound AI Workflows for Dynamic Adaptation 84
Milos Gravara, Juan Luis Herrera, Stefan Nastic

ADVICE: Automatic Identification of Variables to Checkpoint Through Compiler Augmentation 94
Xin Huang, Luanzheng Guo, Nathan R. Tallent, Kento Sato

Enhanced SVM for Improving Application Performance Under GPU Memory Oversubscription 105
Bennett Cooper, Thomas R. W. Scogland, Rong Ge

Priority-Aware GPU Co-Scheduling for High Performance Computing 115
Naman Kulshreshtha, Tapasya Patki, Aniruddha Marathe, Tom Scogland, Rong Ge

EmbdC: Error-Bounded Lossy Video Embedding Compression for On-Device LLM Inference 126
Bo Jiang, Taolue Yang, Youyuan Liu, Xubin He, Sheng Di, Sian Jin

LLM-Pilot: SLO-Aware and Cost-Efficient LLM Serving on Public Cloud VM Clusters via Offloading 136
Jinwoo Kim, Kihyun Kim, Hyunsun Chung, Jihoon Yang, James J. Kim, Dong Li, Youngjae Kim

EAFAL: An Edge-Based Agentic Framework for Adaptive Selection Between SLMs and LLMs 147
Chamara Madarasingha, Prajyot Singh, Redowan Mahmud, Mahbuba Afrin, Aneesh Krishna, Salil Kanhere

Exploring Silent Data Corruption as a Reliability Challenge in LLM Training 157
Anton Altenbernd, Philipp Wiesner, Odej Kao

Noise-Aware Misclassification Attack Detection in Collaborative DNN Inference 167
Shima Yousefi, Saptarshi Debroy

SeCI: A Framework for Self-Certified Identity for Autonomous AI Agents 178
Nehal F. Al-Otaiby, Mohammad Hammoudeh, Jameleddine Hassine

GASched: Goal-Adaptive Hierarchical Reinforcement Learning for Multi-Objective HPC Job Scheduling 188
Minsol Choo, Sangyoon Oh

Reducing Backfill Failures From Workload Drift with Lightweight Uncertainty Buffers in HPC Job Scheduling 198
Jiheon Choi, Sangyoon Oh

Automated Configuration of Power-Management Knobs for Optimal HPC Job Executions 210
Francesco Antici, Andrea Proia, Ryoma Ohara, Toshihiro Hanawa, Zeynep Kiziltan, Andrea Bartolini Jens Domke

Reuse-Aware Min-Cost Flow Scheduling for Distributed LLM Training in Hybrid Optical–Electrical Networks 221
Yifan Yang, Gongming Zhao, Hongli Xu, Huihui Tang

Tula: Optimizing Time, Cost, and Generalization in Distributed Large-Batch Training 231
Sahil Tyagi, Feiyi Wang

VeX: Scaling HNSW-Based Vector Search with DPU Memory and Parallelism 243
Kihwan Kim, Hyungsun Yoo, Woojung Kim, Donghyun Min, Myungcheol Lee, Jihoon Yang, Weikuan Yu Youngjae Kim

Green or Fast? Learning to Balance Cold Starts and Idle Carbon in Serverless Computing 254
Bowen Sun, Christos Antonopoulos, Evgenia Smirni, Bin Ren, Nikolaos Bellas, Spyros Lalis

OpenDC-STEAM: Realistic Modeling and Systematic Exploration of Composable Techniques for Sustainable Datacenters 265
Dante Niewenhuis, Sacheendra Talluri, Alexandru Iosup, Tiziano De Matteis

Spanergy: Energy-Aware Distributed Tracing for Microservices 276
César Perdigão Batista, Denis Conan, Sophie Chabridon

Larger Cloud Servers, Fewer Hosts? on the Evolution of VM Sizes in IaaS Platforms 287
Pierre JacquetT, Camille Coti, Marcos Dias De Assunção

TelePod: Live Migration for Stateful Containers 297
Mingjie Yan, Atharva Ranade, Xin Zhang, Kartik Gopalan

Cost-Justified Multi-type Resource Fair Scheduling for Kubernetes 307
Bo Yan, Madhusudhan Govindaraju, Sujoy Sikdar

Xfagent: Automating Multi-Cloud Deployment of Agentic Workflows on Faas Platforms 313
Varad Kulkarni, Vaibhav Jha, Nikhil Reddy, Anand Eswaran, Praveen Jayachandran, Yogesh Simmhan

A Multi-Armed Bandit–Based Participant Selection Method for Federated Recommendation Systems 319
Jintao Liu, Mohammad Goudarzi, Adel Nadjaran Toosi

FedMO: Mobility-Aware Client Selection in Federated Learning for Drone Delivery Systems 329
Yuan Jiang, Xiao Liu, Jia Xu, Aiting Yao, Frank Jiang, Xuejun Li

Heterogeneous-Resource-Aware Federated Learning with Intelligent LoRA Allocation and Aggregation 339
Youye Xie, Yao Lian, Kevin Chen, Abdul Latif, Lingzhi Zhao, Reza Farivar

AdaSched: A Performance-Driven Cluster Scheduler for Deep Learning Workloads using Deep Reinforcement Learning 350
Han Yin, Jialun Li, Xuan Mo, Weigang Wu

Congestion-Aware Pricing for Fast and Efficient Edge-Cloud Computing 360
Polyzois Soumplis, Emmanouel Varvarigos

SD-MoE: Scenario-Driven MoE Forecasting for Intelligent Elastic Scaling in Cloud Clusters 370
Xianzhao Guo, Weipeng Cao, Minxian Xu, Dachuan Li, Chuanfei Xu, Xi Tao, Zhong Ming

Orchestrating WASM-Based MCP Tool Runtimes for AI Agents Across Edge-Cloud Continuum 380
Moohyun Song, Hayoung Kim, Kyoohyun Lee, Jae Gi Son, Kyungyong Lee

SwiftBot: A Decentralized Platform for LLM-Powered Federated Robotic Task Execution 392
YueMing Zhang, Shuai Xu, Zhengxiong Li, Fangtian Zhong, Xiaokun Yang, Hailu Xu

DriveCache: On-Board Compute Caching for Scalable Vehicular Edge Computing Networks 402
Suvarthi Sarkar, Aditya Gupta, Salil Kashyap, Aryabartta Sahu

Dual-Distilled Heterogeneous Federated Learning with Adaptive Margins for Trainable Global Prototypes 412
Fatema Siddika, Md Anwar Hossen, Wensheng Zhang, Anuj Sharma, J. Pablo Muñoz, Ali Jannesari

RDA-CAFL: Reputation-Dynamic and Distillation-based Asynchronous Conflict-Aware Federated Learning for UAV Networks 422
Jie Li, Pengyang Li, Liming Sun, Xingwei Wang, Yihang Zhang

Evaluating Federated Learning Beyond Simulation: A Deployment-Aware Methodology 431
Mathis Valli, Cédric Tedeschi, Gabriel Antoniu, Loïc Cudennec, Alexandru Costan

Making Variable-Size I/O Practical in ZNS SSDs 442
Sijie Lan, Abutalib Aghayev, Mahmut Kandemir, Umesh Maheshwari

NIO-Cache: Device-Affinitive Page Cache Placement Mechanism for NUMA Systems 452
Jiazheng Zhang, Xiaoyang Wang, Jiwu Shu

S-MSHR: A Scalable MSHR Architecture Using Cache Tag Data-Ready Bits and Index Queues 462
Shuang Wu, Xu Zhang, Yibin Xu, Yangyang Zhao, Tianyue Lu, Mingyu Chen

Benchmarking Message Brokers for IoT Edge Computing: A Comprehensive Performance Study 472
Tapajit Chandra Paul, Pawissanutt Lertpongrujikorn, Hai Duc Nguyen, Mohsen Amini Salehi

Adaptive Edge Context Caching Framework for Context-Aware Internet of Things Applications 482
Ashish Manchanda, Prem Prakash Jayaraman, Abhik Banerjee, Arkady Zaslavsky

Hierarchical Semi-Supervised Federated Learning for UAV-Enabled Fire Monitoring 488
Mark Adrian Gambito, Bahman Javadi, Lorenzo Carnevale, Massimo Villari

Synthetic Data Generation for Storage Failure Prediction in Large-Scale Systems 494
Chandranil Chakraborttii, Ana Luisa Veroneze Solórzano, Devesh Tiwari

FedOort: A Fair and Efficient Optimization Method for Federated Learning 500
Zhaohua Zheng, Junhui Du, Xin Xie, Qiquan Chen, Xu Han, Qijun Huang, Jin Zhang

Evidential Trust-Aware Model Personalization in Decentralized Federated Learning for Wearable IoT 510
Murtaza Rangwala, Richard O. Sinnott, Rajkumar Buyya

CroSatFL: Energy-Efficient Federated Learning with Cross-Aggregation for Satellite Edge Computing 520
Nan Yang, Bahman Javadi, Rodrigo Neves Calheiros, David Boland, Philip Leong

ARC: Adaptive Resource Coordination for Write Stall Mitigation in LSM-Tree 530
Guangxun Zhao, Yongjie Zhu, Suhwan Shin, Seehwan Yoo, Jongmoo Choi

ALTOCUMULUS: Enabling Efficient Erasure Coding in IPFS 540
Mohammad Rizk, Shadi Ibrahim, Thomas Lambert

BucketLSM: Breaking the Compaction Scalability Barrier in LSM-Based Key-Value Stores 550
Jaewan Park, Kyungwook Min, Sungjin Byeon, Taewan Noh, Hyungi Park, Xubin He, Hong-Yeon Kim Youngjae Kim

Multi-Objective Load Balancing for Heterogeneous Edge-Based Object Detection Systems 561
Daghash K. Alqahtani, Maria A. Rodriguez, Muhammad Aamir Cheema, Adel N. Toosi

Deep Reinforcement Learning-Driven Edge Offloading for Latency-Constrained XR Pipelines 571
Sourya Saha, Saptarshi Debroy

Decentralized Resource Sharing in Edge–Cloud Federations via Multi-Agent Hierarchical Reinforcement Learning 583
Panagiotis Kokkinakis, Polyzois Soumplis, Emmanouel Varvarigos

CloudFormer: An Attention-Based Performance Prediction for Public Clouds with Unknown Workload 593
Amirhossein Shahbazinia, Darong Huang, Luis Costero, David Atienza

PM2Lat: Highly Accurate and Generalized Prediction of DNN Execution Latency on GPUs 604
Truong-Thanh Le, Hoang-Loc La, Amir Taherkordi, Frank Eliassen, Phuong Hoai Ha, Peiyuan Guan

Towards Proactive AIOps: Transfer Learning for Unsupervised Anomaly Detection via Bi-LSTM Autoencoder 615
Danny De Novi, Lorenzo Carnevale, Massimo Villari

Efficient Concurrent GHZ State Distribution in Quantum Networks 625
Charles Cao, Weisheng Si, Jong Youl Choi, Sajal K. Das

QRAP: Adaptive Quantum-Safe Risk-Aware Prioritization for Cloud Applications 635
Surabhi Garg, Nidhi Singh, Delton Myalil, Shiv Shankar, Meena Singh Dilip Thakur, Rajan M A

BCRLSecureLink: A Blockchain, Cryptography, and Reinforcement Learning-Based Defense Against Link Discovery Attacks in SDVN 641
Patikiri Arachchige Don Shehan Nilmantha Wijesekara, Harsha S. Gardiyawasam Pussewalage Kalupahana Liyanage Kushan Sudheera, Geeth Priyankara Wijesiri

Lifting to Tensors when Compiling Scientific Computing Workloads for AI Engines 647
Nick Brown, Gabriel Rodriguez-Canal

Kumo: A Security-Focused Serverless Cloud Simulator 653
Wei Shao, Khaled N. Khasawneh, Setareh Rafatirad, Houman Homayoun, Chongzhou Fang

LogLearners: Identifying Compromised AI Functions in Serverless Systems 665
Adil Bin Bhutto, Erik Elmroth, Monowar Bhuyan

PoliFlow: Inferring Control-Flow Policies from Serverless Workflows 675
Pedro Escaleira, Vitor A. Cunha, João P. Barraca, Diogo Gomes, Rui L. Aguiar

ACNNS: A Multi-Interest Recommendation Model with Capsule Network 686
Yan Zhang, Xiaotong Cui, Nan Wang, Yingli Zhong

Inversetune: Inverse Synthetic Fine-Tuning for Reliable Structured Output in Small Language Models 695
Markus Goetz, Hojjat Baghban

SOtrain: Efficient LLM Fine-Tuning on a Single-GPU Commodity Server via User-Space I/O and Scheduling Optimizations 701
Zhengguo Liu, Hao Lan, Jiwu Shu

RapidGNN: Communication-Efficient Distributed Training on Large-Scale Graph Neural Networks 707
Arefin Niam, Tevfik Kosar, M. S. Q. Zulkar Nine