2026 IEEE 19th International Conference on Cloud Computing (CLOUD 2026)

Sydney, Australia
13-18 July 2026

IEEE Catalog Number: CFP26CLO-POD
ISBN: 979-8-3195-1197-3

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:	CFP26CLO-POD
ISBN (Print-On-Demand):	979-8-3195-1197-3
ISBN (Online):	979-8-3195-1196-6
ISSN:	2159-6182

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

A Hybrid Approach for Intent-Driven Recommendation of CEI Orchestration Frameworks 1
Chaimaa Zidane, Hayet Brabra, Imen Jerbi, Elimane Yassine Seidou, Ali Nour Eldin, Mohamed Sellami Walid Gaaloul

Intent-Aware Multi-Domain Cloud Continuum Coordination through Hierarchical Deep Reinforcement Learning 12
Anna Lackinger, Andrea Morichetta, Schahram Dustdar

Aura: A Type-Safe, Multi-Paradigm Cloud Simulation Framework in Scala 3 23
Andrea Cappelletti

A Neurosymbolic Prolog Skill for LLM-Driven Service Placement 29
Jacopo Massa, Giuseppe Bisicchia, Patrizio Dazzi, Antonio Brogi

REACH: Reinforcement Learning for Adaptive Microservice Rescheduling in the Cloud–Edge Continuum 32
Xu Bai, Adel N. Toosi, Rajkumar Buyya, Muhammed Tawfiqul Islam

BridgeScheduler: Non-Intrusive Sensing and Collaborative Scheduling for Hybrid Edge-Cloud Environments 43
Song He, Yuehao Xu, Jiasheng Wang, Jian Yu, Zhijun Ding

Enhancing Task-based Workflows Deployment Across the Compute Continuum with Kubernetes Multiclusters 54
Alba Cañete-Garrucho, Bernat Xandri-Zaragoza, Kemalcan Bora, Elli Kartsakli, Eduardo Quiñones

LLM-Based Multi-Agent Collaboration for Constrained Multi-Objective Container Placement 61
Suvarthi Sarkar, Abhishek Hazra, Akshat Mittal, Aryabartta Sahu

An SLO-Driven Feedback Controller for Kubernetes Horizontal Pod Autoscaling 68
Angelo Marchese, Orazio Tomarchio

Data-Parallel and Heterogeneity-Aware Scheduling for Geo-Distributed Serverless Scientific Workflows 78
Shubhradeep Roy, Suvarthi Sarkar, Koushik Mukka, Tanish Verma, Aryabartta Sahu

MLOPSHER: A Lightweight MLOps Usher with a Minimal Operational Contract for DVC–MLflow–FastAPI on Kubernetes 89
Junsu Park, Hyoungwoo Park

SpotWhisk: Transferring the Spot Instances Cost-Efficiency to Serverless Data Analytics Pipelines 100
Enrique Molina-Giménez, Pedro García-López, Javier Fabra

Ding-Dong Ditch: Peeking Into Spot Instance Availability 111
Kyumin Kim, Moohyun Song, Taeyoon Kim, Kyungyong Lee

Understanding Cross-Cloud Interconnects: Hands-On Measurements and Cost Optimization 122
Eitan Eliav, Isaac Keslassy, David Breitgand, Dean H. Lorenz, Avi Weit

Federated Querying of Heterogeneous Spatiotemporal Data over Polyglot Stores 133
Everett Lewark, Nathan Orwick, Paige Hansen, Ayush Adhikari, Sangmi Lee Pallickara Shrideep Pallickara

Coded Redundancy for Reducing Task Average Latency in Distributed Stream Processing Systems 145
Yuxuan Lin, Bin Tang, Shengjing Wang, Lei Chen, Xutong Jiang, Rajkumar Buyya

Topology-Aware Multi-Agent Reinforcement Learning for Efficient Resource Allocation in Cloud-Native Stream Processing 156
Sunday J. Awine, Jinwei Liu

Predicting Lakehouse Performance in Clouds: An Empirical Exploration of Query Runtime Variance 167
James Nurdin, Wei Liu, Richard Mccreadie, Lauritz Thamsen

Semantic-Aware Graph Schema Discovery for Heterogeneous Cloud Metadata 178
Yann Ramusat, Jeremy Mechouche, Nada El Mansouri, Zakaria Maamar

Small-Sample Tail Latency Prediction for Cloud Time-Series Database Services via Censoring-Aware Extreme Value Theory 189
Fuming Fu

Regime-Aware Resource Demand Forecasting for Cloud Scheduling: When History Beats ML and When ML Matters 196
Kalab M. Kiros, Jinwei Liu

Two-level Network Bandwidth Allocation for Multi-tenant AI Clouds in Kubernetes 203
Sunyanan Choochotkaew, Lionel Jouin, Tatsuhiro Chiba

Augur: Pre-Execution Energy Prediction for Workflow Tasks in Heterogeneous Clusters 214
Kathleen West, Vasilis Bountris, Philipp Thamm, Ulf Leser, Yehia Elkhatib, Lauritz Thamsen

Nf-PEAK: Process-Based Energy Attribution for Nextflow Workflows on Kubernetes Clusters 225
Philipp Thamm, Somayeh Mohammadi, Kathleen West, Knut Reinert, Lauritz Thamsen, Ulf Leser

Estimating Execution Progress of Black-Box Tasks with Resource-Consumption Models 235
Sami Kharma, Tobias Wies, Florian Schintke, Björn Scheuermann

Ditto: Online Memory Rightsizing and Value-Aware Eviction for Reducing Cold Starts in Serverless Platforms 246
Dongjae Lee, Kyuli Park, Yeonwoo Jeong, Hongsu Byun, Sungyong Park

CODA: Breaking CPU Bottlenecks for Scalable Ceph with DPU-Based OSD Decoupling 257
Kyuli Park, Sungmin Yoon, Awais Khan, Sungyong Park, Youngjae Kim

BYSTANDER: State-Aware End-to-End Latency Prediction for Heterogeneous LLM Inference Scheduling 268
Minjae Jung, Hyungjun Kim, Hokun Park, Hwamin Lee, Heonchang Yu

Can Byte-Level and Bit-Level Optimizations Enhance The Parameter Lossless Compression of Large Language Models? 279
Miaowei Liu, Hongcai He, Chengda Wu, Hui Xu, Zetao Zheng, Jie Shao

Reducing Memory Requirements of LLM Inference Through Online rANS Decompression 290
Väinö-Waltteri Granat, Jakub Žádník, Markku Mäkitalo, Joonas Multanen, Pekka Jääskeläinen

Energy Efficiency and Fairness of Local Large Language Models in Data Stream Processing 297
Hessel Tjeerdsma, Roya Nasiri, Joshua Owotogbe, Indika Kumara, Damian Andrew Tamburri Willem-Jan Van Den Heuvel

NetCAS: Dynamic Cache and Backend Device Management in Networked Environments 304
Joon Yong Hwang, Chanseo Park, Younghoon Kim

FATE: Fairness-Aware Memory Allocation for Multi-Tenant Tiered Memory Environments 315
Minwoo Jo, Binwon Song, Hayong Jeong, Heeseung Jo

Tail Erase Count: A Wear Metric for Lifetime-Aware Management in Heterogeneous SSD Clusters 327
Ilhoon Shin

Input-Agnostic Runtime Metric Based Memory Optimization for Serverless Functions 334
Seunghun Kim, Youngjoo Shin

PRISM: Workload-Aware Autonomous Network Fault Recovery for Hyperscale AI Training Fabrics 341
Rohan Bhosle

ContaTOE: FPGA SmartNIC-Based TCP Offloading for Scalable Multi-Container Networking 352
Tatsuya Kawano, Ryuichi Sakamoto, Gongyu Deng, Takaaki Fukai, Takahiro Hirofuchi

Penny: Reducing Storage Costs for Ethereum Synchronization with Hybrid Storage Devices 364
Yunhyeong Seo, Inhwi Hwang, Hyeonsang Eom, Jin Yong Ha

Design Methodology and Performance Trade-offs Management for Distributed and Compound AI Systems 376
Milos Gravara, Andrija Stanisic, Stefan Nastic

Pick and Spin: Cold-Start-Aware Routing for Self-Hosted LLM Serving 388
Bhanu Prakash Vangala, Tanu Malik

Task-Aware Offloading for Heterogeneous DC-HAP Systems 395
Jichen Lu, Balsam Alkouz, Osama Amin, Basem Shihada

A Cooperative MADRL Approach for Joint Computing Offloading and Route Optimization in Cloud-Edge Networks 402
Yi Yue, Shuai Zhang, Zhengxin Han, Ying Liu, Chang Cao, Xiongyan Tang

Fairness-Aware Dynamic Pricing and Service Routing for Cloud-Edge Systems via Heterogeneous Multi-Agent Learning 408
Yun Xia, Gang Zhou, Lichuan Pan, Ming Zhang, Jian Yang, Yang Yang, Zhipeng Gao

InferScale: Kubernetes-Native Predictive Autoscaling for SLO-Aware AI Inference 414
Wen Huang

Dynamic Pipeline Inference Optimization for LLMs: Load-Aware Partitioning, Updating, and Migration 422
Yingxin Wang, Binbin Feng, Zhijun Ding

WVA: A Global Optimization Control Plane for llmd 433
Abhishek Malvankar, Lionel Villard, Mohammed Abdi, Tommaso Sgreccia, Evgeny Shindin, Braulio Dumba Vishakha Ramani, Asser Tantawi, Tamar Eilam

AnchorDL: Dual-Locality-based Scheduling for Serverless Inference Workflows 443
Min Chang Kim, HyungJun Kim, Hokun Park, Minjae Kang, Hwamin Lee, Heonchang Yu

Ada-MGAD: Adaptive Multi-Modal Graph Anomaly Detection 454
Linlin Zhang, Xiaoran Zhao, Chuan Luo

QCOEM: Quantum Cloud Orchestration with Evolutionary Multi-Objective Optimization 460
Tam N. Pham, Hoa T. Nguyen, Quan Le-Trung

Operational Intelligence for Incident Management through OSS MCP-Based Agentic Orchestration 467
Ba Hung Nguyen, Hitoshi Yabusaki

SplitPGV: Separating Neighbors from Element for Efficient HNSW Indexing 470
Beomjoon Shin, Jonghyeok Park

Detecting False Assurance in Cloud Control Inheritance Using a Federated Compliance Integrity Graph 473
Arun Chowdary Sanna

Kernel-Anchored Zero-Trust: A Sidecar-less Service Mesh for Low-Latency Financial Edge Applications 483
Shrikant Sonparote, Sudhanshu Maurya

Secure and Efficient Monitoring of Confidential VMs Using Programmable Read-ahead with eBPF 493
Kanta Uesugi, Kenichi Kourai

C-PCP: Correlation-Aware Proportional Critical Path Scheduling for Input-Sensitive Serverless Workflows 500
G Vasantha Reddy, Chandrashekhar Dharmarajan, Suvarthi Sarkar, Aryabartta Sahu

GreenGNN: Energy-Aware Windowed Communication Optimization for Distributed GNN Training 507
Arefin Niam, Tevfik Kosar, M. S. Q. Zulkar Nine

HYADES: Adaptive Consumer Tuning in Redis Streams via Hybrid Expert Selection 518
Kemalcan Bora, Elli Kartsakli, Eduardo Quiñones Moreno

Flare: Leveraging Serverless Elasticity to Absorb Microservice Load Spikes 529
Dilina Dehigama, Shyam Jesalpura, David Schall, Antonios Katsarakis, Marios Kogias, Rakesh Kumar Boris Grot

Probabilistic Marking based Secure Cross-user Deduplication in Cloud Storage 539
Lele Rao, Jiaxin Li, Xin Tang

A Cloud and IoT-Based Secure Data Sharing Model for Smart Homes and Cyber-Physical Systems 549
Weerasak Chiangsoi, Maribel Fernández, Jenjira Jaimunk, Bhavani Thuraisingham

Grouping based Double Obfuscation Strategy for Secure Cross-user Deduplication in Cloud Storage 561
Ruoxi Zhao, Ranshang Sun, Chuqiao Song, Xin Tang

Closing the Agent Blind Spot: Zero Trust Behavioral Verification for Human and Non-Human Identities in Cloud Storage 568
Venkata Saikumar Thalupuru, Vishal Gupta