2025 IEEE 18th International Conference on Cloud Computing (CLOUD 2025)

Helsinki, Finland
7-12 July 2025

IEEE

IEEE Catalog Number: CFP25CLO-POD
ISBN: 979-8-3315-5558-0

****** 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:	CFP25CLO-POD
ISBN (Print-On-Demand):	979-8-3315-5558-0
ISBN (Online):	979-8-3315-5557-3
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

2025 IEEE 18th International Conference on Cloud Computing (CLOUD)

CLOUD 2025

Table of Contents

Message from the 2025 Congress Steering Committee Chair
Message from the 2025 Congress General Chairs
Message from the 2025 Congress Program Chairs
Message from the CLOUD 2025 Chairs
CLOUD 2025 Organizers
CLOUD 2025 Program Committee

Session I: AI for Cloud Optimization and Deployment

Accelerating RL-Based Scheduler Adaptation with Transfer Learning in Evolving HPC Architectures 1
Lingfei Wang (The University of Melbourne, Australia), Maria A. Rodriguez (The University of Melbourne, Australia), and Nir Lipovetzky (The University of Melbourne, Australia)

LLM-Powered Automated Cloud Forensics: From Log Analysis to Investigation 12
Dalal Alharthi (University of Arizona, USA) and Rozhin Yasaei (University of Arizona, USA)

Korel: Mitigating Stragglers via Real-Time Automatic Mixed Precision in Distributed Deep Learning Environments 23
Hyunseung Jung (Korea University), HyungJun Kim (Korea University), and Heonchang Yu (Korea University)

Session II: AI for Cloud Optimization and Deployment

Multi-Agent Reinforcement Learning-Based In-Place Scaling Engine for Edge-Cloud Systems 32
Jovan Prodanov (Jožef Stefan Institute, Slovenia), Blaž Bertalanic (Jožef Stefan Institute, Slovenia), Carolina Fortuna (Jožef Stefan Institute, Slovenia), Shih-Kai Chou (Jožef Stefan Institute, Slovenia), Matjaž Branko Jurič (University of Ljubljana, Slovenia), Ramon Sanchez-Iborra (University of Murcia, Spain), and Jernej Hribar (Jožef Stefan Institute, Slovenia)

Streamlining Resilient Kubernetes Autoscaling with Multi-Agent Systems via an Automated Online Design Framework 43
Julien Soulé (Univ. Grenoble Alpes, France), Jean-Paul Jamont (Univ. Grenoble Alpes, France), Michel Occello (Univ. Grenoble Alpes, France), Louis-Marie Traonouez (BU IAS, France), and Paul Theron (AICA IWG, France)

The IoT Whisperer: A Framework for Intelligent IoT Service Composition through LLMs 54
Ewan Warburton (Lancaster University, United Kingdom), Abdessalam Elhabbash (Lancaster University, United Kingdom), Saad Ezzini (King Fahd University of Petroleum and Minerals, Saudi Arabia), and Yehia Elkhatib (University of Glasgow, United Kingdom)

Session III: Resource Management, Scheduling, and Orchestration

Dynamic In-Node Group-Aware Scheduling for Multi-Tenant Machine Learning Services on Kubernetes 65
Peini Liu (Barcelona Supercomputing Center, Spain) and Jordi Guitart (Universitat Politècnica de Catalunya, Spain)

ESTHER: Application-First Hardware-Level QoS-Enforcement for Cloud Native Environments 75
Oliver Larsson (Umeå University, Sweden), Thijs Metsch (Intel Corporation, Germany), Cristian Klein (Umeå University, Sweden), and Erik Elmroth (Umeå University, Sweden)

Towards Secure Cloud-Native Computing: Unveiling Kubernetes Misconfigurations with Large Language Models 86
Mostafa Anouar Ghorab (Laval University, Canada) and Mohamed Aymen Saied (Laval University, Canada)

Session IV: Resource Management, Scheduling, and Orchestration

Is Your Cluster Truly Fully Loaded? Exploring Shadow Resources in Host State Synchronization 97
Jiawen Liu (Tongji University, China), Yuehao Xu (Tongji University, China), and Zhijun Ding (Tongji University, China; The Shanghai Artificial Intelligence Laboratory, China)

Helm-ET: Reducing Exposure to Lateral Movement in Kubernetes Artifacts 109
Jacopo Bufalino (CNAM, France; Aalto University, Finland), Jose Luiz Martin Navarro (Universitat de València, Spain; Aalto University, Finland), Aleksi Peltonen (CISPA Helmholtz Center for Information Security, Germany), and Tuomas Aura (Aalto University, Finland)

HeteroScheduler: Dynamic Task Scheduling for CPU-GPU Optimization and Contention Mitigation in Cloud Data Centers 121
Seokwon Choi (Seoul National University) and Hyeonsang Eom (Seoul National University)

Session V: Serverless & Function-as-a-Service

MOBOS: Co-Optimizing Cost and Execution Time in Serverless Workflow with Multi-Objective Bayesian Optimization 132
Minjae Kang (Korea University, Republic of Korea) and Heonchang Yu (Korea University, Republic of Korea)

Causal Latency Modelling for Cloud Microservices 143
Christopher Lohse (University of Dublin Trinity College, Ireland), Diego Tsutsumi (IBM Research Europe, Ireland), Amadou Ba (IBM Research Europe, Ireland), Pavithra Harsha (IBM T. J Watson Research Center, USA), Chitra Subramanian (IBM T. J Watson Research Center, USA), Martin Straesser (University of Würzburg, Germany), and Marco Ruffini (University of Dublin Trinity College, Ireland)

HotSwap: Enabling Live Dependency Sharing in Serverless Computing 152
Rui Li (Northeastern University, USA), Devesh Tiwari (Northeastern University, USA), and Gene Cooperman (Northeastern University, USA)

Session VI: Infrastructure for ML & AI

Speeding up Model Loading with Fastsafetensors 163
Takeshi Yoshimura (IBM Research - Tokyo, Japan), Tatsuhiro Chiba (IBM Research - Tokyo, Japan), Manish Sethi (IBM Research, USA), Daniel Waddington (IBM Research, USA), and Swaminathan Sundararaman (IBM Research, USA)

Cost-Efficient VM Selection for Cloud-Based LLM Inference with KV Cache Offloading 175
Kihyun Kim (Sogang University, Republic of Korea), Jinwoo Kim (Sogang University, Republic of Korea), Hyunsun Chung (Sogang University, Republic of Korea), Myung-Hoon Cha (ETRI, Republic of Korea), Hong-Yeon Kim (ETRI, Republic of Korea), and Youngjae Kim (Sogang University, Republic of Korea)

ZipNN: Lossless Compression for AI Models 186
Moshik Hershcovitch (IBM Research; Tel Aviv University), Andrew Wood (Boston University), Leshem Choshen (IBM Research; MIT), Guy Girmonsky (IBM Research), Roy Leibovitz (Dartmouth College), Or Ozeri (IBM Research), Ilias Ennmouri (IBM), Michal Malka (IBM Research), Peter Chin (Dartmouth College), Swaminathan Sundararaman (IBM Research), and Danny Harnik (IBM Research)

Session VII: Programming Abstractions & Data Processing

Disk-Based Shared KV Cache Management for Fast Inference in Multi-Instance LLM RAG Systems..... 199
Hyungwoo Lee (Sogang University, Republic of Korea), Kiyhun Kim (Sogang University, Republic of Korea), Jinwoo Kim (Sogang University, Republic of Korea), Jungmin So (Sogang University, Republic of Korean), Myung-Hoon Cha (ETRI, Republic of Korea), Hong-Yeon Kim (ETRI, Republic of Korea), James J. Kim (Soteria Inc., Republic of Korea), and Youngjae Kim (Sogang University, Republic of Korea)

ClusterLink: Redefining Application Connectivity for the Multi-Cloud Era 210
Kfir Toledo (IBM Research), Pravein Govindan Kannan (IBM Research), Michal Malka (IBM Research), Etai Lev-Ran (IBM Research), Or Ozeri (IBM Research), Vita Bortnikov (IBM Research), Ziv Nevo (IBM Research), and Kathy Barabash (IBM Research)

Precomputation-Optimized Lakehouse Architecture for Online Analytical Processing Tasks 223
Haida Zhang (Shanghai Jiao Tong University), Lin Sun (Shanghai Jiao Tong University), Zhengtong Zhang (Shanghai Jiao Tong University), Jiayang Xia (Shanghai Jiao Tong University), Ziang Huang (Shanghai Jiao Tong University), Jiansi Wang (Shanghai Jiao Tong University), Haopeng Chen (Shanghai Jiao Tong University), Yan Jiao (Shapere Tech Co., Ltd), and Yongming Xu (Shapere Tech Co., Ltd)

Session VIII

Energy-Aware Resource Allocation and Container Migration in Distributed Data Centers under Variable Energy Pricing: A Genetic Programming Hyper-Heuristic Approach 233
Mathew Falloon (Victoria University of Wellington, New Zealand), Hui Ma (Victoria University of Wellington, New Zealand), and Gang Chen (Victoria University of Wellington, New Zealand)

EnergyLess: An Energy-Aware Serverless Workflow Batch Orchestration on the Computing Continuum 243
Reza Farahani (University of Klagenfurt, Austria) and Radu Prodan (University of Innsbruck, Austria)

Carbon-Aware Temporal Data Transfer Scheduling Across Cloud Datacenters 255
Elvis Rodrigues (University at Buffalo (SUNY), USA), Jacob Goldverg (University at Buffalo (SUNY), USA), and Tevfik Kosar (University at Buffalo (SUNY), USA)

Session IX: System Monitoring & Analysis

TraceWizard: End-to-End Distributed Tracing Across Host and Network Devices in Cloud 265
Kuangyuan Li (Sun Yat-sen University), Jingrun Zhang (Sun Yat-sen University), Pengfei Chen (Sun Yat-sen University; Guangdong Key Laboratory of Big Data Analysis and Processing), Hongyang Chen (Sun Yat-sen University), Ruipeng Hong (Sun Yat-sen University), Wanqi Yang (Sun Yat-sen University), and Chen Sun (Huawei Technologies Co., Ltd.)

Mind the Memory Gap: Unveiling GPU Bottlenecks in Large-Batch LLM Inference 277
Pol G. Recasens (Barcelona Supercomputing Center (BSC); Universitat Politècnica de Catalunya - BarcelonaTech (UPC)), Ferran Agullo (Barcelona Supercomputing Center (BSC); Universitat Politècnica de Catalunya - BarcelonaTech (UPC)), Yue Zhu (IBM Research), Chen Wang (IBM Research), Eun Kyung Lee (IBM Research), Olivier Tardieu (IBM Research), Jordi Torres (Barcelona Supercomputing Center (BSC); Universitat Politècnica de Catalunya - BarcelonaTech (UPC)), and Josep Ll. Berral (Universitat Politècnica de Catalunya - BarcelonaTech (UPC); Barcelona Supercomputing Center (BSC))

Efficient Microservice Monitoring via Kernel Transformation and FFT Forecasting 288
Marianna Ojanen (University of Helsinki, Finland), Maryam Sabzevari (Nokia Bell Labs, Finland), and Sandor Szedmak (Aalto University, Finland)

Session X

Efficient Versioning for Unikernels 296
Gaulthier Gain (University of Liège, Belgium), Benoît Knott (University of Liège, Belgium), and Laurent Mathy (University of Liège, Belgium)

Real-Time Interference-Aware CPU and I/O Capping Mechanism for Multi-Tenant Containers 308
MohammadReza HoseinyFarahabady (The University of Sydney, Australia) and Albert Y. Zomaya (The University of Sydney, Australia)

SLO-Aware Container Orchestration on Kubernetes Clusters 318
Angelo Marchese (University of Catania, Italy) and Orazio Tomarchio (University of Catania, Italy)

Session XI: Edge Computing

ReSACO: A Meta Reinforcement Learning Method for Fast Offloading in Mobile Edge Computing 328
Myeongjun Kim (Korea University, Republic of Korea) and Heonchang Yu (Korea University, Republic of Korea)

MSTH-Former: Optimizing Workload Prediction in Edge-Cloud Continuum with Multi-Scale Temporal and Hierarchical Knowledge Convergence and Distillation 339
Sharmen Akhter (Kyung Hee University, Korea Republic Of) and Eui-Nam Huh (Kyung Hee University, Korea Republic Of)

PROBA: Enhancing Serverless Edge Computing via Adaptive Task Scheduling and Probabilistic Resource Sharing 351
Manish Pandey (Kyungpook National University, South Korea), Byungchul Tak (Kyungpook National University, South Korea), and Young-Woo Kwon (Kyungpook National University, South Korea)

Session XII: Security, Quality, Consensus

RACS-SADL: Robust and Understandable Randomized Consensus in the Cloud 362
Pasindu Tennage (EPFL and ISTA, Switzerland), Antoine Desjardins (ISTA, Austria), and Lefteris Kokoris-Kogias (Mysten Labs and ISTA, Greece)

An Experimental Validation of Architectural Measures for Cloud-Native Quality Evaluations 374
Robin Lichtenthäler (University of Bamberg, Germany) and Guido Wirtz (University of Bamberg, Germany)

Routing Strategies for RoCE Networks in AI Clouds 385
Abdul Alim Alim (IBM Research, USA), Ali Sydney (IBM Research, USA), Liran Schour (IBM Research, Israel), Laurent Schares (IBM Research, USA), Pavlos Maniotis (IBM Research, USA), Anand Singh (IBM Research, USA), and Bengi Karacali (IBM Research, USA)

Session XIII

QPS-Fit: An Efficient and Performant Parallel Algorithm for Hybrid Optical and Packet Switching 397
Dongzhao Song (Georgia Institution of Technology), Jingfan Meng (Georgia Institution of Technology), Qianru Yu (Georgia Institution of Technology), and Jun Xu (Georgia Institution of Technology)

HEART: Heterogeneous-Aware Traffic Allocation in Multi-Replica Deployments on Kubernetes ... 409
Hokun Park (Korea University, Korea), Donggyun Kim (Korea University, Korea), HyungJun Kim (Korea University, Korea), Gyujeong Lim (Korea University, Korea), and Heonchang Yu (Korea University, Korea)

Optimizing Receive Flow Steering for Mixed Traffic in High-Performance Cloud Datacenters 420
Junseo Jang (Sungkyunkwan University, South Korea) and Jaehyun Hwang (Sungkyunkwan University, South Korea)

Avoiding Pitfalls in Networked Key-Value Store for Tiered Memory 430
Seungmin Shin (Soongsil University), Leeju Kim (Soongsil University), Wookyung Lee (Soongsil University), Eyee Hyun Nam (FADU Inc.), Seungmin Kim (FADU Inc.), Bryan S. Kim (Syracuse University), Sungjin Lee (POSTECH), and Eunji Lee (Soongsil University)

Session XIV: Short & Work-in-Progress Papers

Universal Workers: A Vision for Eliminating Cold Starts in Serverless Computing 442
Saman Akbari (Technische Universität Berlin, Germany) and Manfred Hauswirth (Technische Universität Berlin, Germany; Fraunhofer Institute for Open Communication Systems (FOKUS), Germany)

Towards Efficient Key-Value Cache Management for Prefix Prefilling in LLM Inference 445
Yue Zhu (IBM T. J. Watson Research Center, USA), Hao Yu (IBM T. J. Watson Research Center, USA), Chen Wang (IBM T. J. Watson Research Center, USA), Zhuoran Liu (IBM T. J. Watson Research Center, USA), and Eun Kyung Lee (IBM T. J. Watson Research Center, USA)

Automated LLM Deployment and Evaluation: A Cloud-Native Approach using LLM-as-a-Judge .. 448
Ansar Rafique (University of Oxford, UK) and Brian D. Marsden (University of Oxford, UK)

DNN-Adapt: Reinforcement Learning-based Hybrid Batching for Efficient DNN Serving 451
Sai Venkat Malreddy (University of California Santa Cruz, United States of America), Milind Varma Penumathsa (University of California Santa Cruz, United States of America), and Liting Hu (University of California Santa Cruz, United States of America)

Game-Theoretic Reinforcement Learning for Task Optimization under Time-Sensitive Constraints 454
Emanuele Carlini (National Research Council of Italy, Italy), Patrizio Dazzi (University of Pisa, Italy), and Matteo Mordacchini (National Research Council of Italy, Italy)

Revisiting SQL Statement Logging for SQLite on AWS S3 457
Yewon Shin (Hankuk University of Foreign Studies) and Jonghyeok Park (Korea University)

Serverless Data Analytics (Finally) Bridging the Gap: Introducing the Ortzi DataFrame 460
Germán T. Eizaguirre (Universitat Rovira i Virgili, Tarragona), Marc Hostau (Universitat Rovira i Virgili, Tarragona), and Marc Sánchez-Artigas (Universitat Rovira i Virgili, Tarragona)

Temporal Fusion Transformer Based Vertical Scaling Management for Kubernetes 468
Kemalcan Bora (Barcelona Supercomputing Center, Spain), Elli Kartsakli (Barcelona Supercomputing Center, Spain), and Eduardo Quiñones Moreno (Barcelona Supercomputing Center, Spain)

Author Index **475**