Biennial since 1998
CSNDSP
2026
Edinburgh, UK

Special Session on **Performance-centric Design of ML/DL Systems across IoT, Edge, and Cloud**

Georgios Bouloukakis University of Patras, Greece gbouloukakis@upatras.gr Georgios Bouloukakis is an Assistant Professor at the University of Patras, Greece. He previously held roles at Télécom SudParis and UC Irvine. His research focuses on IoT/Edge middleware and distributed systems, with leadership in multiple EU projects.	Marin Orlić Ericsson Research, Sweden marin.orlic@ericsson.com Marin Orlić is a researcher with Ericsson Research, Stockholm. His research is focused on applications of hybrid AI techniques in telecom networks, such as machine reasoning and cognitive architectures for the current wave of agents.	Houssam Hajj Hassan Orange Innovation, France houssam.hajjhassan@orange.com Houssam Hajj Hassan is a postdoctoral researcher at Orange Innovation. His research focuses on advancing autonomous IoT systems through hybrid AI techniques involving AI planning, RL, and causal discovery. He contributes to the EU research project PANDORA to enable trustworthy and dependable AIoT systems.

Scope of the session

The rise of *Machine Learning* (ML) has reshaped real-time perception and decision-making, powering applications ranging from smart surveillance and natural language interfaces to personalized healthcare, smart manufacturing and autonomous vehicles. Architectures like convolutional and transformer-based neural networks are increasingly deployed in latency-sensitive, resource-constrained environments. While their computational demands often limit deployment only to resource-rich settings, AI services are shifting toward the broader *Computing Continuum* (CC), spanning Cloud, Edge, and IoT devices. This migration addresses needs for low-latency responses, privacy preservation, and energy efficiency but introduces challenges in executing complex models on constrained Edge devices.

To tackle these, a combination of approaches such as specialized hardware, model compression, distributed learning (DL) and inference, etc. will be necessary across the CC. This also raises critical questions about *partitioning*—dividing workloads—and *placement*—assigning them to specific nodes—under dynamic conditions, requiring strategies that adapt to real-world variability. Hybrid cognitive architectures are one approach to adapt to the dynamic changes in the environment while keeping the balance of size and complexity, ensuring the desired inference properties, and maintaining robust execution.
An emerging paradigm is the use of small and large (foundation) models in different roles within such an architecture, enabling and orchestrating it, or participating in inference.

This special session calls for contributions on the performance modeling of ML/DL Systems and corresponding architectures to enable their efficient management and orchestration across the IoT-Edge-Cloud continuum.

Prospective authors are invited to submit original and unpublished work on the following research topics related to this Special Session:

- *Seamless IoT-Edge-Cloud orchestration*
- *Real-time AI on diverse IoT devices*
- *Low-latency approaches for time-critical ML/DL systems*
- *ML/DL performance benchmarking on IoT-Edge-Cloud*
- *Energy-aware resource management*
- *Performance modeling for ML/DL systems deployment*

Program Committee Members

- *DL partitioning strategies for split computing*
- *Distributed continual learning*
- *Lifecycle management of distributed models*
- *Techniques for DL of foundation models*
- *System infrastructure for DL of foundation models*
- *Verification and safety of distributed models*
- *Hybrid cognitive architectures*

Program Committee Members

- Alessandro Palma, Sapienza University of Rome, Italy
- Chih-Kai Huang, Télécom Paris, France
- Ajay Kattepur, Ericsson Research, India
- Zahraa El Attar, Télécom SudParis, France
- Shahin Abdoul Soukour, Télécom SudParis, France
- Ourania Manta, CyberAlytics Limited, Cyprus
- Katerina Tzompanaki, University CY Cergy Paris, France
- Swarup Kumar Mohalik, Ericsson Research, India
- Ayush Kumar Varshney, Ericsson Research, Sweden
- Xin Tao, Ericsson Research, Sweden
- Jean Martins, Ericsson Research, Brazil
- Josue Castaneda Cisneros, Ericsson Research, Sweden
- Dagnachew Azene Temesgene, Ericsson Research, Sweden
- Alex Palaios, Ericsson Research, Germany
- Jishnu Sadasivan, Ericsson Research, India
- Prayag Gowgi S K, Ericsson Research, India

Co-funded by
the European Union

This work has received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No 101135775.