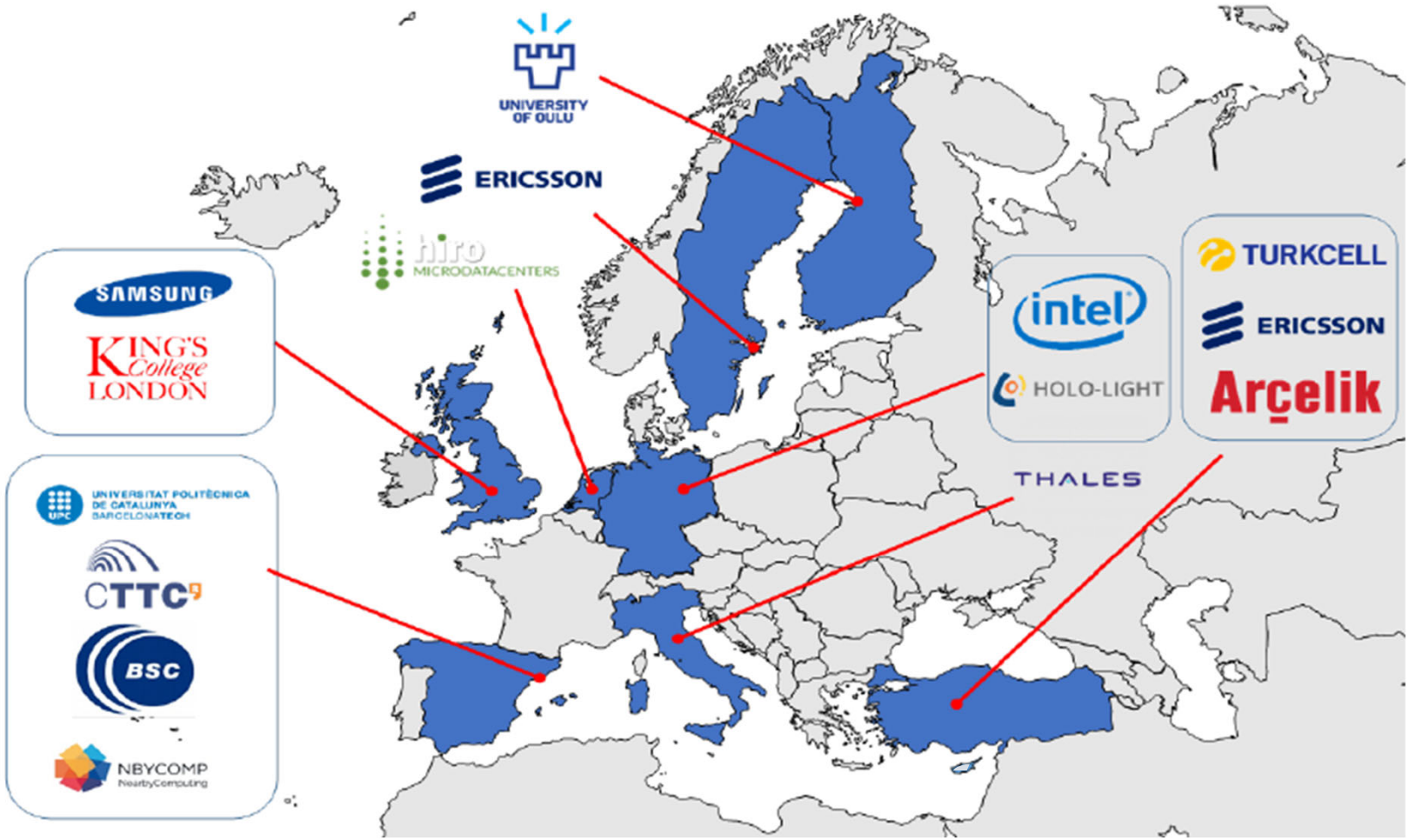
AI-powered eVolution towards opEn and secuRe edGe architEctures

O. Sallent (UPC)

SNS-JU Webinars, 20th February 2023

Consortium


UNIVERSITY OF OULU
ERICSSON
hiro MICRODATACENTERS
SAMSUNG
KING'S College LONDON
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
CTTC
BSC
NBYCOMP
intel
HOLO-LIGHT
THALES
TURKCELL
ERICSSON
Arçelik

Introduction

- Applications based on eXtended Reality (XR) expected to become key
  - Fuse digital and real world to deliver new experiences
  - Will realize the looked-for digital transformation of vertical industries, e.g., manufacturing

- Wide penetration of the Internet of Things (IoT)
  - Massive numbers of sensing devices, generating huge volumes of data
    - Leveraging Artificial Intelligence (AI) data is transformed into valuable and actionable knowledge, able to automate and optimize the decision-making process in e.g., smart city and autonomous driving

- Capabilities of communication infrastructures grow
  - Unprecedented spectral efficiencies
  - RAN disaggregation offer enhanced flexibility and scalability
  - AI-enabled solutions to achieve closed-loop automation

Introduction

- Emerging applications increasingly demanding in terms of computation
  - Necessary to **offload heavy tasks** to more powerful computing elements, typically residing at the cloud…..
  - …..however, **cloud computing** is no longer capable to meet the latency requirements of such applications, nor deal effectively with distributed and heterogeneous massive IoT deployments……
- **Edge computing** has been rapidly evolving as a novel computing paradigm that brings computational power and resources closer to where the data is generated
- **Synergy between B5G and edge computing** can provide computing and storage capabilities for applications residing at the boundary of operators' networks

Heterogeneous computing elements & architectures, and communication technologies, creating challenges in the ***development***, ***deployment*** and ***orchestration*** of innovative services

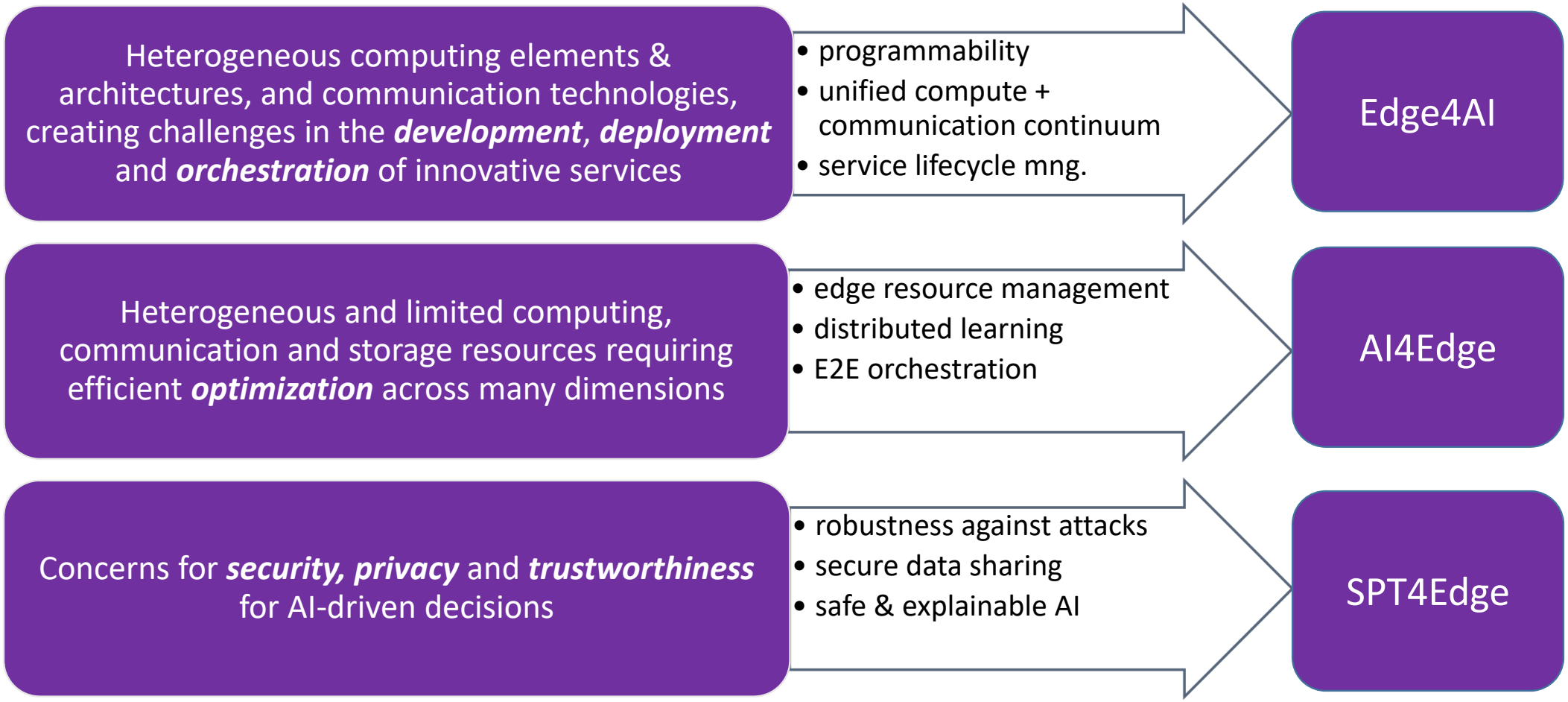
Heterogeneous and limited computing, communication and storage resources requiring efficient ***optimization*** across many dimensions

Concerns for ***security, privacy*** and ***trustworthiness*** for AI-driven decisions

VERGE's objectives and approach

- Project VERGE tackles **edge computing evolution** and envisages a solution approach sustained on three main pillars:
  - **"edge for AI"**: a flexible, modular and converged edge platform design, unifying the lifecycle management and closed-loop automation for cloud-native applications, Multi-access Edge Computing (MEC) and network services across the edge-cloud compute continuum for ultra-high computational performance
  - **"AI for edge"**: an AI-powered portfolio of solutions leveraging the multitude of collected metrics for intelligent management and orchestration
  - "**security, privacy and trustworthiness of AI-based models at the edge",** providing a suite of methods to protect AI models against adversarial attacks, increase their explainability and reliability, and ensure data privacy

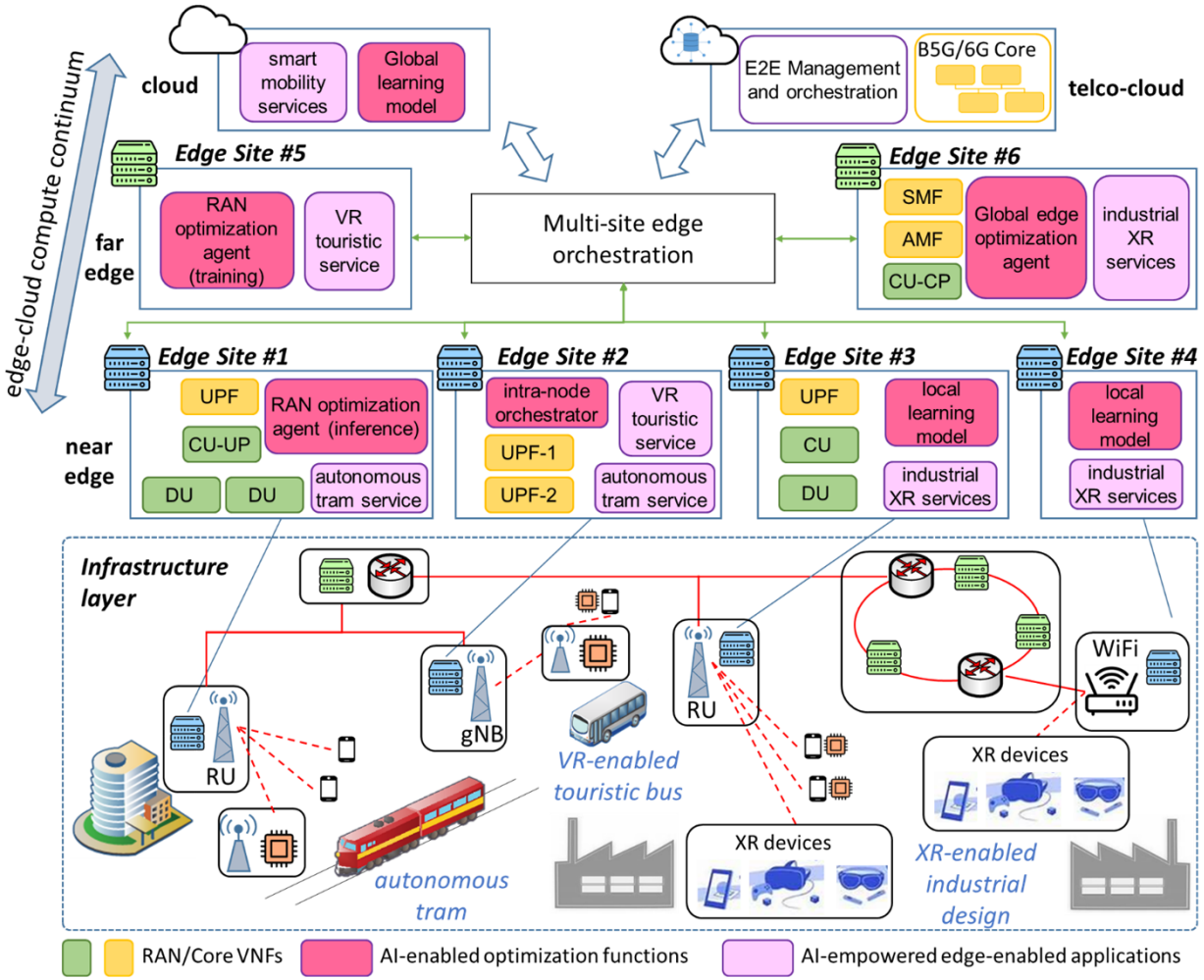
VERGE's objectives and approach



VERGE's objectives and approach - Envisioned scenario

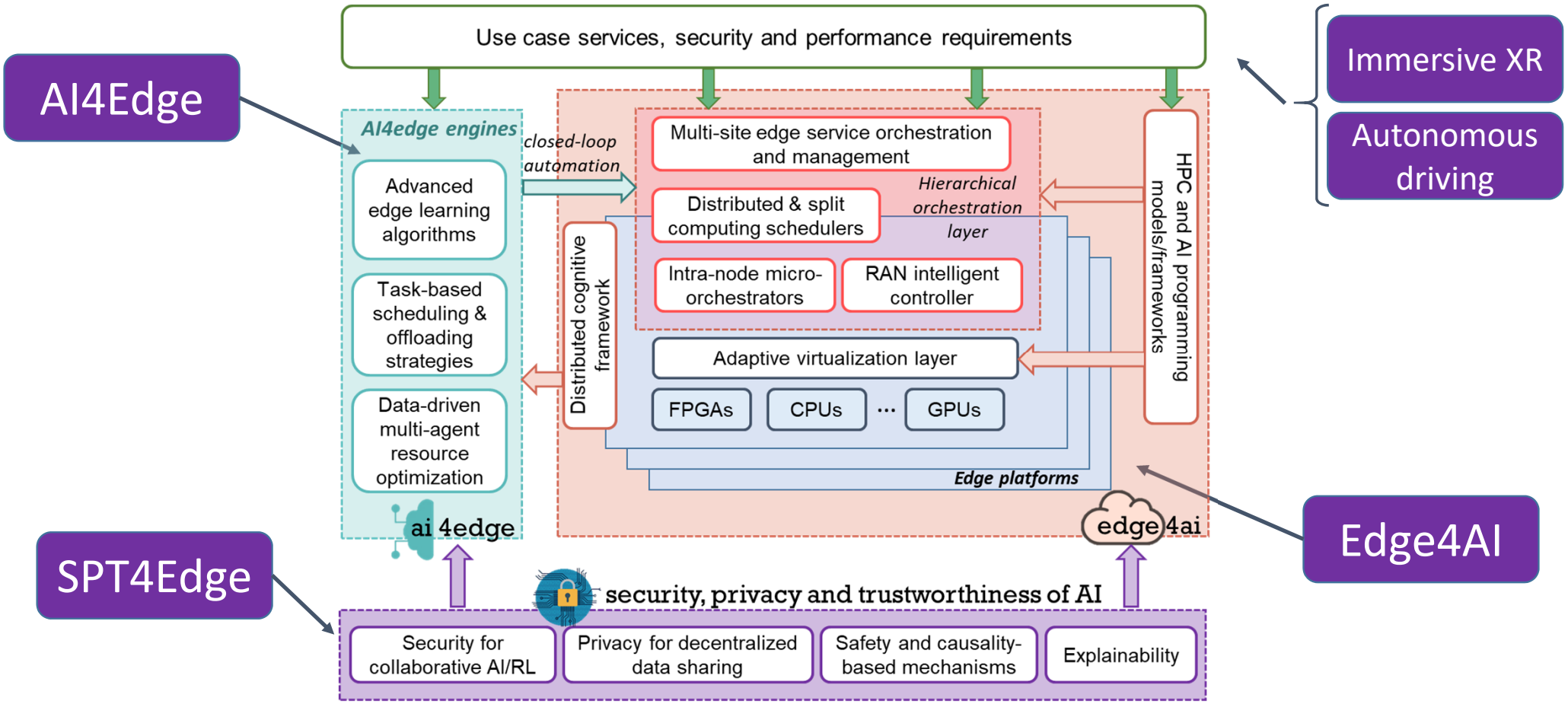
Edge-cloud compute continuum where virtualized services can be flexibly deployed and executed

- Vertical applications
- RAN and core VNFs
- AI-enabled functions for network optimization and automation.



European Commission

VERGE's objectives and approach



Edge4AI - Support for distributed and split computing over heterogeneous computation architectures

- Design of an **adaptive virtualization layer specifically targeting programmable accelerated HW platforms**, enabling the dynamic reconfiguration of functions across embedded (AI) accelerators and general-purpose computing elements

- **Programming models and practices** from the High Performance Computing (HPC) and AI domains will be employed
  - HPC parallel programming models for shared memory (e.g., OpenMP)
  - Accelerator-specific programming environments (e.g., Intel® oneAPI , NVIDIA CUDA , Vitis AI development toolkit) to exploit the inner parallelism of multi- and many-core processor architectures.
  - Tensorflow and Pytorch, as well as open-source frameworks for distributed environments such as COMP Superscalar (COMPSs) can facilitate the design of complex AI workflows, such as FL solutions.

Edge4AI - Multi-site hierarchical edge orchestration

- Design an **end-to-end orchestration framework for multi-site service orchestration and management** of heterogeneous virtualized multi-edge infrastructures
    - Extending NearbyOne solution
    - Developing the necessary interfaces and communication pipelines with network management entities of the B5G network (e.g., network slicing managers and intelligent RAN controllers) & underlying edge platform

Edge4AI - Closed-loop automation framework

- **Distributed cognitive framework**, responsible for collecting
  - RAN and core metrics, exposed by the B5G network
  - Edge platform telemetry (e.g., CPU/storage/memory utilization, etc.)
  - Application-related requirements, exposed by the employed programming models.

- Targets the execution of zero-touch orchestration and optimization operations.

- The cognitive framework will fuel the AI4Edge layer which will, then, be enforced by the multi-site hierarchical edge orchestrator, for closed-loop automation operations

AI4Edge - Advanced edge learning algorithms

- **Novel solutions for the efficient training of ML models** at the evolved edge
  - Addressing the computational distribution of FL tasks, enabling an optimal selection of UEs as learning agents
  - Addressing the trade-off between computation and communication to optimally distribute ML models across the federation

- **Dynamically splitting DNN models** into head and tail portions (deployed at UEs and edge side, respectively), optimizing the splitting point on-the-fly based on varying environmental conditions (e.g., channel quality, throughput, UE battery level, edge server load).

- **Advanced distributed learning algorithms** to achieve edge-empowered applications, such as (multi-tier) asynchronous FL, transfer learning and robust FL under limited resources.

AI4Edge - Task-based scheduling and computational offloading

- **Intelligent solutions on a variety of allocation challenges at the edge**:
  - Dynamic allocation of application tasks to the most suitable computing resources across the edge-cloud compute continuum
  - MEC selection strategies for computational offloading
  - Optimal computational splitting of AI models

- Joint consideration of metrics provided by the distributed cognitive framework of the Edge4AI layer

AI4Edge - Data-driven resource optimization

- To compose a multi-level multi-agent end-to-end framework integrating the different AI-based solutions in a **collision-free** and efficient manner
  - **Optimum splitting point of the CU and DU functions** across the edge-cloud compute continuum at the service run time based on RAN conditions and leveraging AI-based prediction methods
  - **Capacity sharing for RAN slicing in case of changes in the current deployment** (e.g., due to the activation/deactivation of RAN nodes).
    - Use of transfer learning to accelerate the training of the model under the new conditions
  - **End-to-end slicing in vehicular scenarios** (e.g., for the autonomous tram) **exploiting the predictability of the tram trajectory** to timely prepare the slice allocation.

SPT4AI - Security and privacy

- To **ensure the security and privacy** of the sensitive information carried by the data and models **against attacks** such as membership inference, model inversion, and model extraction attacks.
  - Privacy preserving technologies such as FL, differential privacy and homomorphic encryption can be leveraged

- Distributed ledger technologies (DLT), and specifically blockchains, will be employed to provide secure, distributed and decentralized data sharing, under heterogeneous communication environments.

SPT4AI - Safety and explainability

- To **increase the overall trustworthiness of the AI4Edge models**, a portfolio of algorithms for building safe, explainable and causal RL solutions will be provided, following three interconnected approaches for the design of:

  - **Safe-RL techniques based on formal verification,** which can be used to verify the correctness of AI models, both during training and during deployment

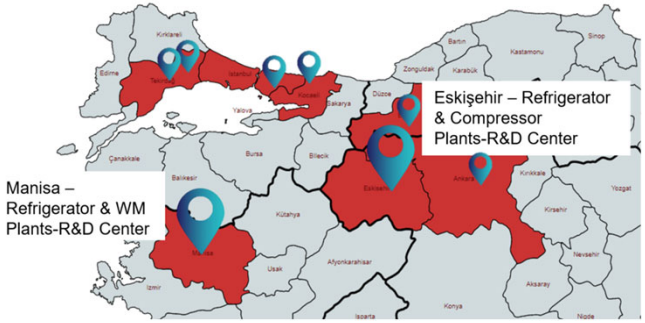
  - **Causality-based methodologies** to build causal graphs, identify causal variables and hidden confounders, and using them to build more trusted RL models.

  - **Techniques based on knowledge graphs**, their maintenance and knowledge extraction, to pose queries to the RL-based agents and obtain semantically rich explanations for the decisions.

R&D maturity

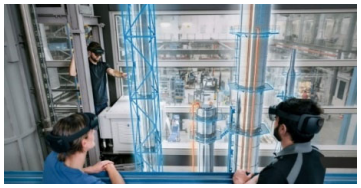
VERGE Technology	TRL M1	TRL M30
VERGE Open Edge Platform for Cloud Continuum	4	5
VERGE AI-driven service and network orchestration	3	5
VERGE programming models and split-computing practices	2	4
VERGE micro orchestration of RAN functions and computing resources in disaggregated and softwarized beyond 5G networks	2	4
Multi-level, multi-agent AI solutions jointly optimising network and computing resources, leveraging distributed learning technologies under edge constraints	2	4
VERGE framework for cloud-type edge services on a distributed edge infrastructure.	4	5
VERGE methodology for secure, trustworthy AI for Edge Computing	2	4
5G-enabled XR Design Review/Virtual Prototyping	3	5

Demonstrations overview



XR-DRIVEN EDGE-ENABLED INDUSTRIAL B5G APPLICATIONS

- Edge intelligence for XR-aided teleoperation
- B5G edge computing for XR-driven collaborative design and real-time virtual prototyping



EDGE ASSISTED AUTONOMOUS TRAM

- Dynamic and adaptive AI-based end-to-end slicing
- Edge4AI platform orchestration, distribution and cognition capabilities
- Smart micro-orchestration in disaggregated softwarized RAN elements
- Secure and trustworthy AI for smart mobility scenarios
- AI-driven split computing

THANK YOU FOR YOUR ATTENTION

Contacts

Coordinator

Oriol Sallent

Universitat Politècnica de Catalunya

sallent@tsc.upc.edu

Technical Manager

Elli Kartsakli

Barcelona Supercomputing Center

elli.kartsakli@bsc.es

Website

www.verge-project.eu

Social networks

@verge_project

www.linkedin.com/company/verge-snsproject/