Co-funded by
the European Union

6GSNS

TrialsNet: TRials supported by Smart Networks beyond 5G

Deliverable D2.2

Intermediate design of Platforms and Networks solutions

Project Details

Call	HORIZON-JU-SNS-2022
Type of Action	HORIZON-JU-IA HORIZON JU Innovation Actions
Project start date	01/01/2023
Duration	36 months
Grant Agreement No.	101095871

Deliverable Details

Deliverable WP:	WP2
Deliverable Task:	T2.1, T2.2, T2.3
Deliverable Identifier:	TrialsNet D2.2
Deliverable Title:	Intermediate design of Platforms and Networks solutions
Editor(s):	Nina Slamnik-Kriještorac (IMEC)
Author(s):	Paolo Giaccone, Maurizio Munafò (CNIT)
	Ruben Cerezo, Fernando Beltrán (ERC)
	Nina Slamnik-Kriještorac, Raul Cuervo, Hojjat Navidan, Dries Naudts (IMEC)
	Gabriele Scivoletto (NXW)
	Razvan Mihai, Marius Iordache (ORO)
	Giulio Bottari, Paola Iovanna (TEI)
	Mauro Agus, Alessandro Trogolo (TIM)
	Marco Gramaglia (UC3M)
	Andreas Georgakopoulos, Eleni (Nelly) Giannopoulou, Vera Stavroulaki, Panagiotis Demestichas (WINGS)
Reviewer(s):	Dimitra Tsakanika (DAEM)
	Vera Stavroulaki (WINGS)
Contractual Date of Delivery:	30/04/2024
Submission Date:	07/05/2024
Dissemination Level:	PU
Status:	Final
Version:	1.1
File Name:	2024-05-07_TrialsNet_D2.2_Intermediate design of Platforms and Networks solutions.docx

Disclaimer
The information and views set out in this deliverable are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Deliverable History

Version	Date	Modification
V1.0	*30/04/2024*	*Final version, submitted to EC through SyGMa*
V1.1	*07/05/2024*	*Editorial fix of the submission date.*

Table of Content

List of Acronyms and Abbreviations

Acronym	Description
3GPP	3rd Generation Partnership Project
4G	Fourth Generation of mobile communications
5G	Fifth Generation of mobile communications
5GCN	5G Core Network
A5G	Advanced 5G Technology
6G	Sixth Generation of mobile communications
AI	Artificial Intelligence
AIaaS	AI as a Service
API	Application Programming Interface
APN	Access Point Name
B5G	Beyond 5G mobile network
Bs5G	Baseline 5G
BB	Baseband Unit
BIPT	Belgisch Instituut voor postdiensten en telecommunicatie
CAPIF	Common API Framework
CGNAT	Carrier-grade Network Address Translation
CMU	Compact Mobility Unit
CN	Core Network
CNIT	Consorzio Nazionale Interuniversitario per le Telecomunicazioni
CPE	Customer-Premises Equipment
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CRC	Cyclic Redundancy Check
DBSCAN	Depth-based spatial clustering of applications with noise
DL	Deep Learning
DQN	Deep Q-network
DRL	Deep Reinforcement Learning
DT	Digital Twin
E2E	End-to-end
EAC	Edge Application Client
EAS	Edge Application Server
eBPF	extended Berkeley Package Filter
eCPRI	Evolved CPRI
EdgeApp	Edge Application
EEC	Edge Enabler Client
EES	Edge Enabler Server
EMCO	Edge Multi-Cluster Orchestrator
eNB	Evolved Node B
ERC	Ericsson España SA
ETSI	European Telecommunications Standards Institute
FEC	Forward Error Correction
GAM	Galleria d'Arte Moderna
GDPR	General Data Protection Regulation
GSMA	Global System for Mobile communications Association
gNB	gNodeB
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HARQ	Hybrid Automatic Repeat request
HW	Hardware
IEEE	Institute of Electrical and Electronics Engineers
IME	Intent Management Entity
IMEC	Interuniversity Microelectronics Centre
IP	Internet Protocol
IPSEC	Internet Protocol Security
IoT	Internet of Things
IRU	Indoor Radio Unit
ISG	Industry Specification Group
ITS	Intelligent Transport Systems and Services
KPI	Key Performance Indicator
KV	Key Value
KVI	Key Value Indicator
KVM	Kernel-based Virtual Machine
LADN	Local Area Data Network
LTE	Long Term Evolution
MAC	Medium Access Control
MANO	Management and Orchestration
MCI	Mass Casualty Incident
MCS	Modulation and Coding Scheme
MEC	Multi-Access Edge Computing
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MNOs	Mobile Network Operators
MnS	Management Service
MPTCP	Multi-Path Transport Control Protocol
MQTT	MQ Telemetry Transport
MT-M&O	Multi-Technology Management and Orchestration
NFV	Network Function Virtualization
NFVO	Network Function Virtualization Orchestrator
NPN	Non-Public Network
NR	New Radio
NSA	Non-Standalone Architecture
NWDAF	Network Data Analytics Function
OAI	Open Air Interface
OBU	Onboard Unit
ORO	Orange Romania
OS	Operating System
OSM	Open-Source MANO
PDF	Probability Density Function
PKI	Public Key Infrastructure
PNI-NPN	Public Network Integrated Non Public Network
PoliTO	Politecnico di Torino

PVGIS	Photovoltaic Geographical Information System
PRB	Physical Resource Block
PT	Physical Twin
QoS	Quality of Service
RAM	Random Access Memory
RAN	Radio Access Network
RAT	Radio Access Technology
RD	Radio Dot
RDI	Radio Dot Interface
REST	Representational State Transfer
RSU	Roadside Unit
SA	Standalone Architecture
SDN	Software Defined Networking
SEAL	Service Enabler Architecture Layer
SIM	Subscriber Identity Module
SNPN	Stand-alone Non-Public Network
SVR	Support Vector Regression
TEI	Ericsson Telecomunicazioni SpA
TDD	Time-Division Duplex
TIM	Telecom Italia SPA
UAV	Unmanned Aerial Vehicle
UC	Use Case
UE	User Equipment
UPF	User Plane Function
URI	Uniform Resource Identifier
URLLC	Ultra-reliable low latency communication
USB	Universal Serial Bus
TTI	Transmission Time Interval
V2X	Vehicle-to-Everything
VAL	Vertical Application Layer
VEC	Vehicular Edge Computing
vEPC	Virtual Evolved Packet Core
VIM	Virtual Infrastructure Manager
vLAN	Virtual Local Area Network
VM	Virtual Machine
VNF	Virtual Network Function
VPN	Virtual Private Network
vRAN	Virtualized RAN
VRU	Vulnerable Road User
VXLAN	Virtual Extensible Local Area Network
WP	World Package
Wi-Fi	Wireless Fidelity
XR	Extended Reality
ZSM	Zero-touch Service Management
ZTS	Zero-Touch Service

List of Figures

List of Tables

Executive Summary

The TrialsNet project aims at deploying full large-scale trials, combining a heterogenous and comprehensive set of innovative Sixth Generation of mobile communications (6G) applications based on various technologies such as robots, metaverse, massive twinning, Internet of Senses, and covering three relevant domains of the urban ecosystems in Europe identified by (i) Infrastructure, Transportation, Security & Safety, (ii) eHealth & Emergency, and (iii) Culture, Tourism & Entertainment. Through the results collected from the large-scale trials both in terms of quantitative and qualitative perspective, the project will support the transition from the current Fifth Generation of mobile communications (5G) to the next generation by means of definition of new requirements as well as the identification of the added values that each use case will bring to the community.

In particular, Work Package 2 (WP2) focuses on the development of the platforms and network solutions within those large-scale trials, while other technical WPs, such as WP3, WP4, and WP5, are implementing the Use Cases (UCs) and validating their performance. In collaboration with those WPs, WP2 is designing and deploying the overall network infrastructures with advanced network and platform functionalities to suit the needs (i.e., performance requirements) of each use case falling into the three abovementioned domains. While the deliverable D2.1 [1] focused on the preliminary design of both the platform and network solutions in the different trial sites, i.e., clusters located in Italy, Spain, Romania, and Greece, this deliverable provides insights into the latest progress on both Baseline 5G Technology (Bs5G) and Advanced 5G Technology (A5G) design and preliminary deployment in each of the aforementioned sites.

Within the **Italian cluster**, two different trial sites are deployed, i.e., Turin site and Pisa site. The Turin site is a multi-location trialing setup that relies on the commercial 5G network deployment as a baseline, whereas the advanced capabilities are being deployed on the service level, i.e., Virtualized Network Function (VNF) orchestration for improved service performance. The Pisa site is also a multi-location one, stretching from the CNR campus in Pisa as the primary trial site with advanced End-to-End (E2E) orchestration functionalities, over the Fondazione Monasterio hospital in Massa, and the Ericsson's 5G laboratory in Genoa that is used as setup for testing before trialing use cases in full-scale setup in Pisa, which is being deployed from scratch.

The **Spanish cluster** is centered around the 5Tonic, an open laboratory for research and innovation that focus on 5G technologies. This site is currently undergoing updates towards Release 17 to support A5G deployment and improve uplink throughput capabilities that are necessary for UC testing and implementation in the final trial sites. The current 5Tonic infrastructure has been so far used for performing preliminary tests for some of the TrialsNet use cases, such as UC1, UC6, and UC10.

The **Romanian cluster** consists of the three sites: (i) main outdoor site located in Iasi, which is based on the commercial Orange Non Standalone Architecture (NSA) deployment, (ii) indoor 5G Standalone Architecture (SA) based on Release 16, which now includes also edge computing capabilities with User Plane Function (UPF) deployed at the edge as part of the A5G, and (iii) experimental facilities based on Release 16 and A5G such as Zero-touch Service Management (ZSM) and orchestration located in Antwerp, which are used for testing advanced orchestration properties before deployment and validation take place in Iasi. The experimental facilities in Antwerp have been also provided by the project as an independent trial site for the third parties in the context of the Open Call.

In the scope of the **Greek cluster**, several locations including the Athens airport and public venues in the city of Athens are combined, all covered by public 5G network. In addition to these sites, a private network of the WINGS testbed is used for E2E 5G and Beyond 5G (B5G) testing based on Release 15 network deployment, accompanied by distributed cloud and edge computing capabilities (A5G), which are used for testing and validation of use case prior to their deployment at public locations.

Additionally, building on top of the preliminary design of the main TrialsNet innovations (horizontal and vertical) related to network and platform described in D2.1 [1], this deliverable provides the latest insights into the design updates and implementation progress, with more specific reference to their validation strategies related to specific UCs. These innovations are focused on network features that go beyond the Bs5G and create a set of advanced capabilities for enhanced performance and sustainability of all UCs (which are considered as part of A5G functionalities of network and service infrastructure). In addition to the innovations presented in D2.1 [1], a new vertical innovation is here introduced, i.e., the advanced methodology for traffic and crowd monitoring based on cellular information. This methodology uses signalling data from the network to monitor traffic and

crowds and aims to improve performance and accuracy, as an alternative to solutions based on video surveillance cameras, currently deployed in UC1 and UC4.

Finally, network and service sustainability perspectives are studied and presented in this deliverable, focusing on understanding the sustainability aspects for TrialsNet network technologies, large scale impact, and economic and societal benefits. This study is performed in alignment with the common framework of Key Value Indicators (KVIs) defined in WP6. Various sustainability solutions such as self-sustainable energy harvesting, energy utilization at edge vs. cloud, and energy-aware application design, are explored and presented in this deliverable. In addition, all innovations from WP2 are studied from a sustainability perspective, providing prospects on different sustainability angles that are aimed to be improved with their implementation.

1 Introduction

The focus of WP2 is to deploy network infrastructure with advanced functionalities (i.e., A5G) in different trial sites, which are capable to fulfill the stringent performance requirements in terms of ultra-low latency, ultra-high throughput, and enhanced reliability coming from the use cases. Such A5G deployments be used for large-scale trial activities, which are performed in the context of the other technical WPs, i.e., WP3, WP4, and WP5. The updates in the infrastructure design and deployment reported in this document are taking place with reference to requirements collected from D3.1 [2], D4.1 [3], and D5.1 [4], as well as the preliminary service performance results obtained from D3.2 [5], D4.2 [6], and D5.2 [7].

The UCs defined and implemented by WP3, WP4, and WP5, belong to the three domains:

- Infrastructure, Transportation, Security & Safety (UC1 in Madrid, UC1 and UC4 in Iasi, UC2 and UC3 in Athens, and UC5 in Turin),
- eHealth and Emergency (UC6 in Athens and Madrid, and UC7, UC8, and UC9, in Pisa), and
- Culture, Tourism, and Entertainment (UC10 in Madrid, UC11 and UC13 in Athens, and UC12 and UC13 in Turin).

To introduce the new functionalities for enhanced performance and sustainability of those use cases, WP2 is defining A5G that encompasses advanced network and service capabilities, such as E2E orchestration, VNF orchestration, Artificial Intelligence (AI) as Service (AIaaS), and Digital Twins (DTs), stretching over both experimental and commercial equipment within different sites. The impact of these advanced network functionalities on UCs will be closely monitored in both WP2 and other WPs (WP3, WP4, and WP5), and as such reported in subsequent deliverables, i.e., D2.3, D2.4, D3.3, D4.3, and D5.3.

In this deliverable, Section 2 delves into the latest advancements of the clusters' design and deployment progress, and the underlying network and service infrastructure associated with the different sites. Therefore, this document reports on the improvements between the baseline infrastructure and technology which were reported in D2.1 [1] and the current network snapshot that is used for preliminary UC testing activities.

Building on top the research-oriented activities reported as horizontal (related to transversal B5G/6G functionalities) and vertical innovations (related to particular UCs) in D2.1 [1], Section 3 of this deliverable offers insights into the latest updates in innovations' design and deployment, discussing the testing and validation scenarios focused on evaluating the impact of these innovations on specific use cases.

Given the project's strong focus on sustainability aspects, Section 4 provides details on network and service sustainability perspectives. This section lists and discusses various sustainability solutions such as self-sustainable energy harvesting, energy utilization at edge vs. cloud, and energy-aware application design.

Finally, Section 5 concludes the document, thereby providing the main outcomes of the deliverable D2.2 and shedding light on the next upcoming WP2 activities.

2 Progress on TrialsNet clusters infrastructures activities

This section provides insights into the latest advancements related to TrialsNet clusters and underlying platform and network solutions, on top of the baseline infrastructure and technology which were reported in D2.1 [1]. All advanced technologies available at different sites need to be aligned with the corresponding 5G standards, thereby ensuring successful deployment and operation of all use cases deployed in different TrialsNet clusters.

Thus, following the technological trends and standardization activities, TrialsNet defined the following two distinct sets of technologies, which are considered as reference for the platform and network solutions development D2.1 [1]:

- **TrialsNet Baseline 5G Technology (Bs5G)**: This technology represents the initial range of capabilities and functionalities accessible in the trial sites within all clusters, forming the foundation for the development phase of the Use Cases (UCs). It serves as a benchmark for measuring the initial Key Performance Indicators (KPIs) throughout the implementation of various UCs, and it is considered as the project's starting point to define the initial performance expectations.
- **TrialsNet Advanced 5G Technology (A5G)**: This technology embodies the collection of enhanced capabilities and functionalities on network and service levels. The enhancements of the TrialsNet's A5G are twofold. First, it encompasses the planned improvements of the existing technology within each cluster. Second, it takes the input from evaluations derived from the initial deployments of UCs in each cluster as feedback. Through vigilant monitoring and analysis of the outcomes from these initial UC implementations, TrialsNet aims to pinpoint areas for improvement, optimizing the technology to achieve more advanced capabilities and enhance overall performance.

Mapping the available technologies in all TrialsNet sites to the above-mentioned sets of technologies is illustrated in Figure 1, and as such will be used as a reference for reporting updates in the following subsections. In addition to the overview provided in D2.1 [1], Figure 1 also adds the aspects of VNF orchestration for improvement of service performance in the Italian cluster, Turin site, as a A5G feature that initially was not planned to be implemented.

	3GPP Releases, Edge and Orchestration		Additional functionalities	
	Per site Bs5G	Per site A5G	Per site Bs5G	Per site A5G
Italian (Turin)	Rel-15 + Edge			VNF orchestration
Italian (Pisa)	Rel-16 + Edge	E2E Orchestration	AIaaS	
Spanish	Rel-16 + Edge	Rel-17		AIaaS, DTs, TSN
Romanian	Rel-16	Edge, E2E Orchestration		AIaaS, AI Orchestration
Greek	Rel-15			AIaaS*, AI Orchestration*

Figure 1. **TrialsNet functionalities in different sites (*means at service level only) with reference to Bs5G and A5G.**

2.1 Italian cluster (Turin site)

This section describes the evolution of infrastructure components in terms of platform and network solutions for the Turin site of the Italian cluster in which UC5 "Control Room in Metaverse" [2], UC12 "City Parks in Metaverse" and UC13 "Extended Reality (XR) Museum Experience" [4] are implemented. The Turin site in the Italian cluster includes a very extensive set of technologies, that will leverage on the previously gathered experience in the 5G EVE [8] and 5G-TOURS [9] projects. Also, other available software components are the XR platform developed by TIM (Section 2.1.2) and the Symphony Internet of Things (IoT) platform provided by NXW (Section 2.1.3).

2.1.1 Site description recap

As described in D2.1 [1], the Turin site of the Italian cluster is widely based on the 5G EVE legacy infrastructure, led by TIM and supported by Italian partners. The facilities remain operational and can be expanded to support new activities, including TrialsNet UCs implementation.

The site offers various radio testing environments with dedicated Core Networks (CNs), including a commercial 5G network, based on the 3rd Generation Partner-ship Project (3GPP) release Rel-15 NSA 80 MHz at 3.7 GHz, which is accessible through commercial or private Access Point Names (APNs), such as 5geve.tim.it. The facility, equipped with Ericsson technology, is multi-location and interconnected via high-performance links (via TIM transport network or dark fiber).

According to Figure 1, Bs5G is based on the 5G commercial Radio Access Network (RAN) and CN, which eventual updates will be driven by market request and company strategy, while the A5G has been added considering the introduction of the service level VNF orchestration.

2.1.2 Design aspects

The last upgrade of the Turin cluster site is multi-location environment, and it includes TIM, COTO and CNIT facilities. As depicted in Figure 2, the overall architecture is based on widely available commercial coverage, where a private APN is active, and it can be used by authorized Subscriber Identity Module (SIM) cards (enabled by TIM). User devices can access both the Commercial and Private APNs to access the backend computational resources, deployed into dedicated hardware, PoliTO computing infrastructure, public commercial cloud. In COTO facilities, control room is implemented (UC5).

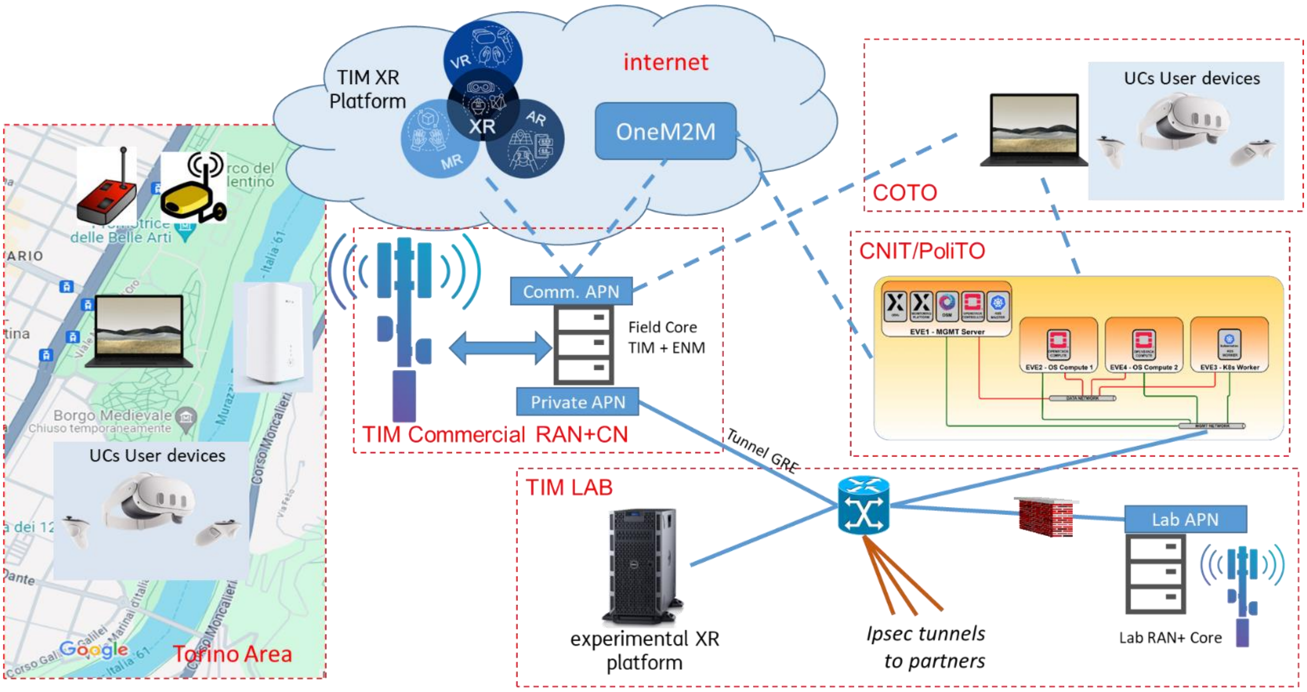

Figure 2. Overall Turin Site architecture.

In particular, to support the UCs, various XR platforms, are available on:

- Commercial cloud (TIM XR Platform)
- Dedicated server (TIM Experimental XR streaming platform)
- PoliTO computing infrastructure (managed by NXW, provided by UC owners).

2.1.2.1 The PoliTO computing infrastructure

TrialsNet utilizes the computing infrastructure (managed by NXW) hosted by CNIT at Politecnico di Torino (PoliTO) to host the VNFs used for implementing internal and third-party Use Cases. This infrastructure has been updated from the one inherited from the 5G EVE project. Figure 3 shows the high-level design of the Turin Site computing infrastructure. The infrastructure will host the backend of UC5, UC12, and UC13. for the Galleria d'Arte Moderna (GAM) and Palazzo Madama experiences.

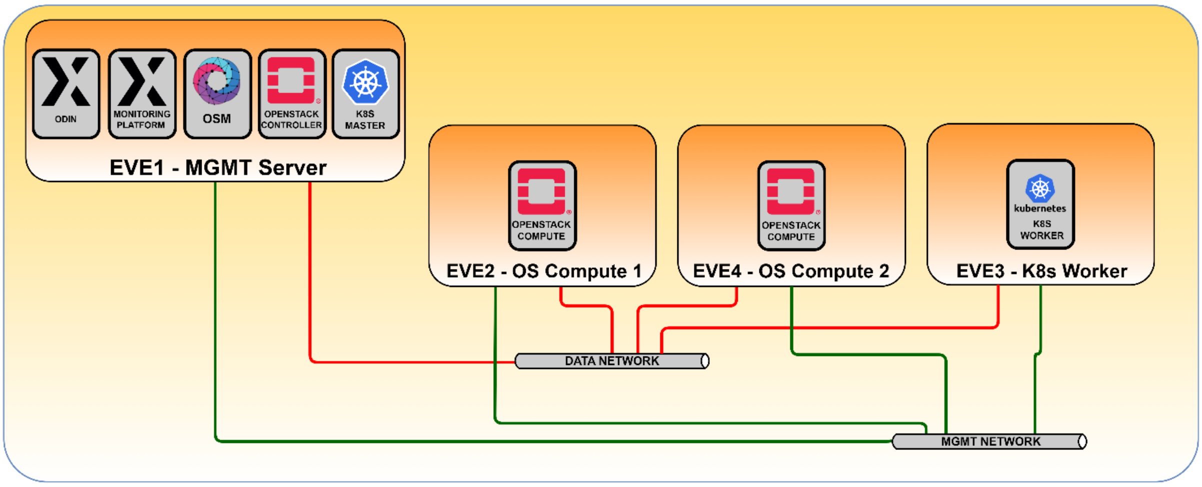

Figure 3. The Turin Site Computing platform architecture.

The system is composed of 4 physical servers, each of them implementing a specific role.

- **Management Server (EVE1):** this server acts as the control hub for the testbed and it provides a virtualized environment with Virtual Machines (VMs) hosting the various management elements of the system, i.e., the Virtualized Infrastructure Manager (VIM) controllers, the Network Function Virtualization Orchestrator (NFVO) performing VNF orchestration, the monitoring platform.
- **OpenStack Compute Nodes (EVE2 and EVE4):** two servers are used to host the VMs, providing the computing resources to instantiate network functions and application components. Using OpenStack configurations it is possible to select the target server to deploy a given VM, allowing to implement a custom resource allocation logic within the testbed.
- **Kubernetes Worker node (EVE3):** this server is responsible for hosting containerized applications and services, enabling deployment and scaling of workloads.

In details, the following software modules have been deployed:

- **OpenStack:** an open-source cloud computing platform designed to create and manage a wide range of cloud infrastructure services, including computing, storage, and networking. It is used by organizations and service providers to build private and public cloud environments. The OpenStack controller has been deployed in a dedicated Kernel-based Virtual Machine (KVM) in the Management Server, while the compute nodes have been deployed in two dedicated servers (EVE2 and EVE4)
- **Kubernetes (k8s):** an open-source solution for orchestrating and managing containerized applications and services. The k8s Master Node has been deployed in a dedicated KVM-based Virtual Machine in the Management Server, while the k8s worker node has been deployed in a dedicated server (EVE3)
- **The Open Source Mano (OSM):** an open-source community hosted by European Telecommunications Standards Institute (ETSI) developing an open-source NFV (Network Function Virtualization) Management and Orchestration (MANO) stack aligned with ETSI NFV Information Models. OSM adopts information models aligned with the ETSI standards, publicly available, suitable for generalized VNFs and VIM-independent. It has been deployed in a dedicated KVM-based Virtual Machine in the Management Server
- **The NXW Monitoring Platform:** a software component developed by NXW which implements functionalities for collection, storage and visualization of monitoring data generated from different data sources. It has been deployed in a dedicated KVM-based Virtual Machine in the Management Server

- **The NXW Continuum Multi-Technology Management and Orchestration Platform (MT-M&O):** a versatile software system designed for B5G/6G networks. It enables efficient deployment and distribution of applications across Extreme-Edge, Edge, and Cloud Continuum resources by employing Continuum resources discovery, allocation, and migration strategies. MT-M&O leverages prediction algorithms to anticipate dynamic constraints like battery levels, energy consumption, and connectivity quality. Its architecture includes a driver-based approach for platform and device agnosticism, allowing seamless integration with heterogeneous virtualization platforms (like K8s and OpenStack) and various Extreme-Edge devices, such as IoT devices and robots.

Figure 4 and Figure 5 show the management interfaces of the k8s and OpenStack instances currently running in the Turin Site.

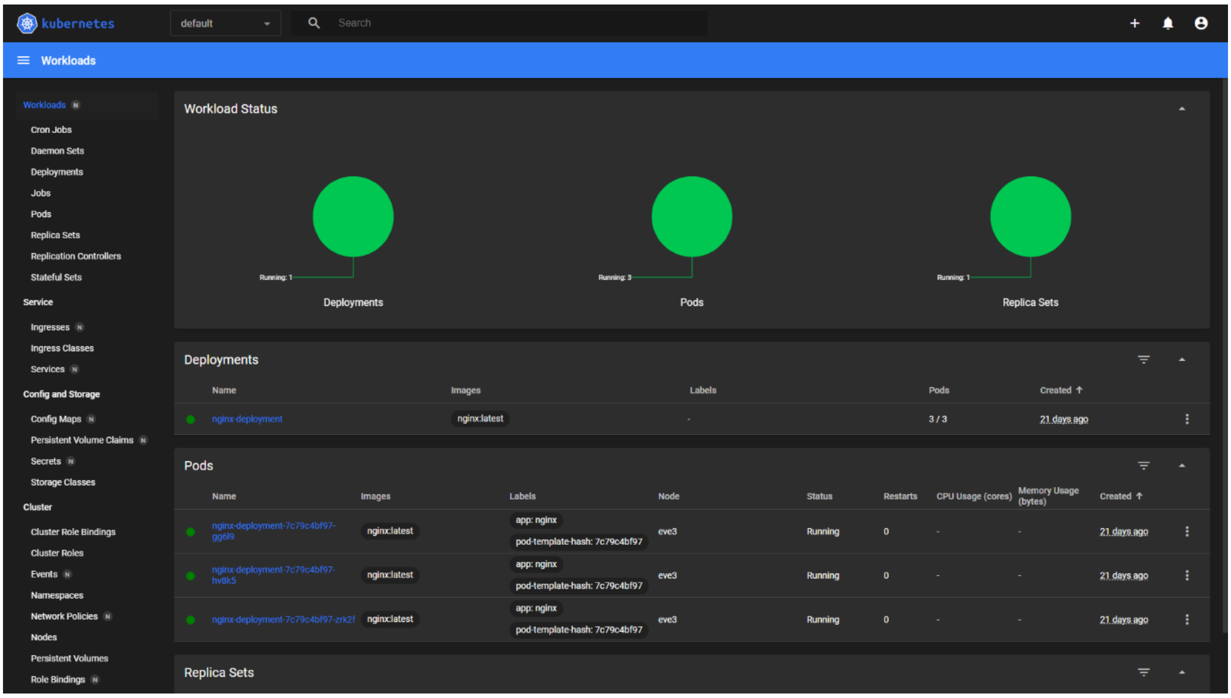

Figure 4. Turin site K8s management interface.

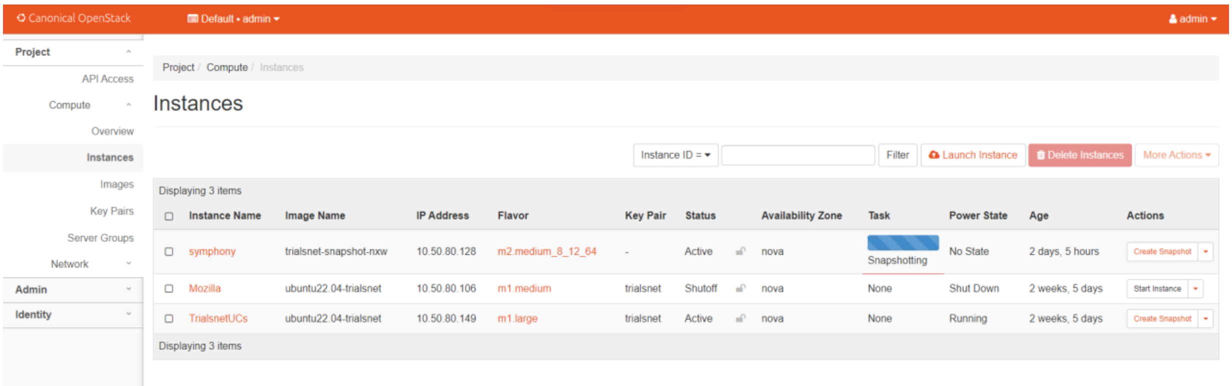

Figure 5. Turin Site OpenStack management interface.

2.1.3 Initial deployment activity

As reported in the previous section, all the use cases implemented in the Turin site will exploit the platform and network solutions that are part of the overall architecture reported in Figure 2. In this context, while UC12 and UC13 are mostly based on a client-server model in which the applications are running on the users' devices and the backend components are instantiated on the VNF infrastructure, the UC5 requires a more complex solution that includes the deployment of specific physical devices such as cameras and sensors, a dedicated mid-layer software application, for which a ad-hoc architectural solution is going to be adopted. In the following sub-section, a description of such distinctive elements.

2.1.3.1 People counting sensor

The "Area Eventi" in Valentino Park has been equipped with sensors that count the presence of people indirectly by estimating the number of Wireless Fidelity (Wi-Fi) equipped devices located nearby the sensors. Two locations have been chosen in the area, as depicted in Figure 6. The sensors denoted as "box" sensors have been installed at ground level, whereas the sensors denoted as "pole" sensors have been installed at about 4 meters above the ground level on a pole. For each location, two sensors have been installed: one off-the-shelf commercial sensor by Dropper [10] and one experimental sensor developed specifically by CNIT. Thus, a total of 4 sensors have been installed.

Figure 6. Installed sensors (left) and pole sensors (right).

2.1.3.2 Video cameras

The selected video cameras are a high-end Internet Protocol (IP) camera model (Hikvision) which renders high-quality images across a range of lighting conditions, minimizing storage and bandwidth requirements and providing data-powered situational awareness. The camera can be configured according to several parameters: video codec, bitrate, frame rate, security protocols, and Region of Interest (see Figure 7). Such flexibility is ideal to test the 5G network under different loads. The tests have been performed using Wi-Fi and a 5G connection via a router on the only model currently commercialized in Europe, which has a Fourth Generation of mobile communications (4G) connection and Wi-Fi, but the new model with 5G will be available at the time of the trials.

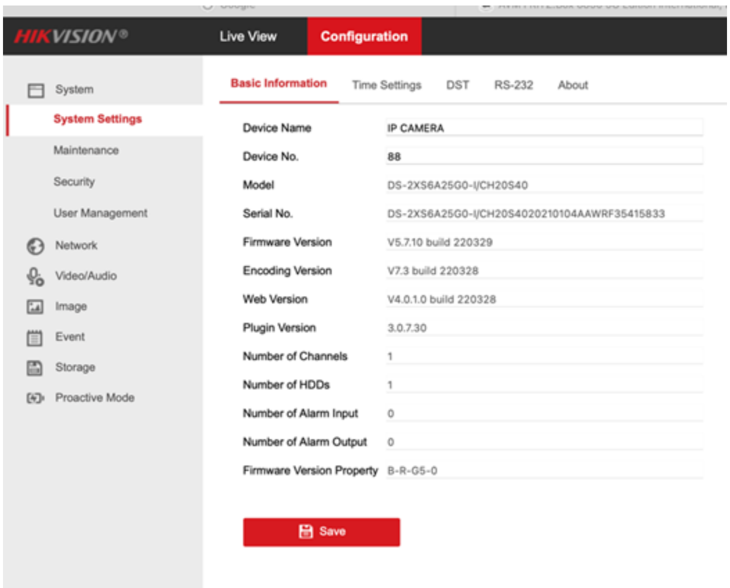

Figure 7. Configurator of the video camera.

2.1.3.3 Symphony Internet of Things (IoT) platform

The Symphony IoT platform is a central component for the Control Room and Metaverse use case. The tool is in charge of handling the entire lifecycle of the data coming from the sensors deployed in the field. A preliminary description of the high-level architecture and the main functionalities has been provided in deliverable D2.1 [1], further elaborated in D3.1 [2], particularly from the perspective of the functionalities involved in UC5. Within the Turin Site, the Symphony Platform monitors the sensors associated with UC5, while network KPIs and metrics are collected from the Monitoring Platform described in section 2.1.2.1.

Deployed as a Virtualized Network Function (VNF), the platform has been implemented within the Turin Site infrastructure as an OpenStack Virtual Machine. The connection to the sensors deployed in the field is ensured by the TIM 5G network dedicated to the Use Case, allowing the sensors to correctly push the data to the IoT platform to be processed and presented on the Mozilla Hub metaverse platform.

The southbound interface of Symphony is composed of the Hardware Abstraction Layer, which is in charge of implementing the driver to properly handle the connection with the sensors' fieldbus. For this purpose, a dedicated Message Queueing Telemetry (MQTT) driver has been developed to translate the sensors' data model into the Symphony internal one. The Symphony northbound interface for this use case is implemented through a Graphical User Interface (GUI) that has been embedded into the Mozilla Hub metaverse platform.

The data are currently being stored in the internal Symphony Data Storage (fvStorage), which incorporates a database and exposes a Representational State Transfer (REST) Application Programming Interface (API) for eventually fetching the data from external tools.

Figure 8 shows the Symphony IoT Platform management interfaces, where four sensors have been configured to receive the data streams from the devices installed in the Trial area, while the corresponding dashboard is depicted in Figure 9.

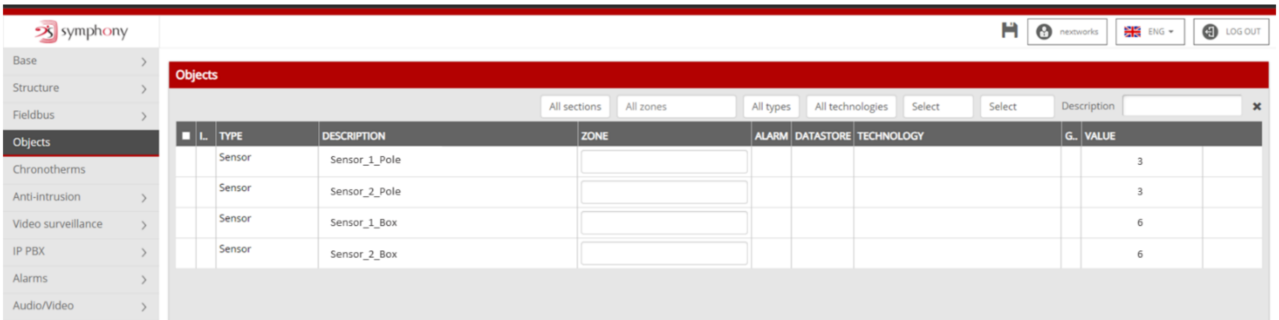

Figure 8. The Symphony IoT Platform management interface.

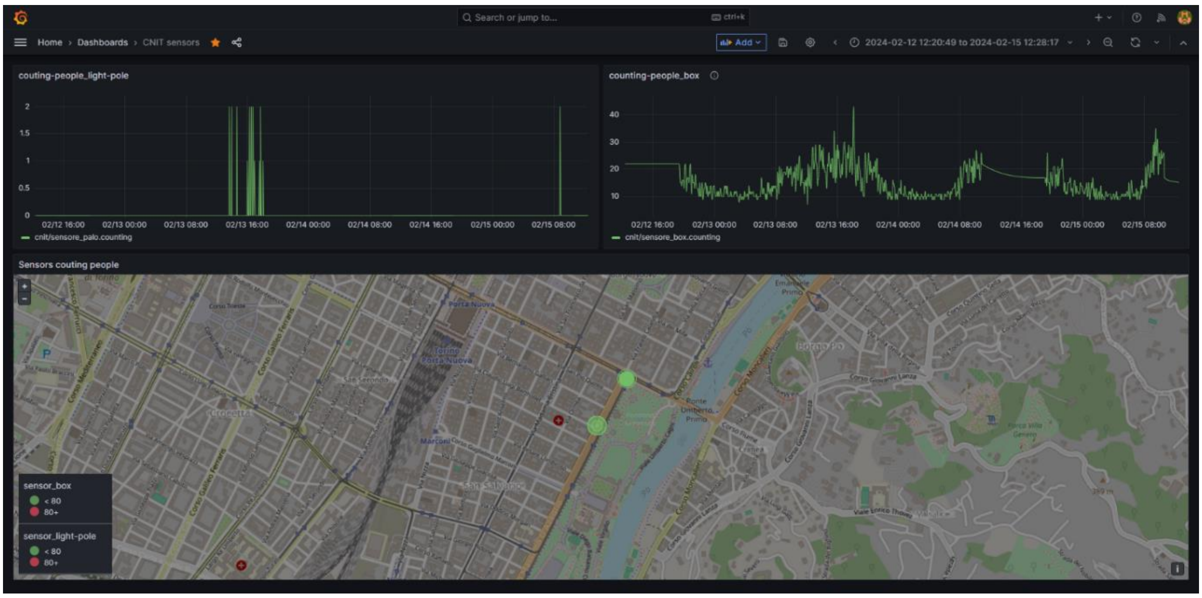

Figure 9. Dashboard exposed by the Symphony IoT Platform.

Based on the overall Turin site architecture described in section 2.12, the ad-hoc architectural solution for UC5 illustrating the connections between various elements described above is depicted in Figure 10. Sensors and cameras are connected to the 5G network provided by TIM through a priority SIM, ensuring highly reliable communication even in cases of congestion of the radio access network. The connections are directed to a public APN and the traffic is routed through the Internet to PoliTO NFV infrastructure and to COTO premises. In particular, video feeds from field cameras are transmitted through the 5G network to the 5T servers of COTO before reaching the PoliTO, while the sensors information reach directly the PoliTO. Similarly, devices used by field agents (e.g., phones, tablets, headsets) connect via 5G through priority SIM, in order to achieve high performance in terms of bandwidth and reliable communications. The connections are directed to a private APN through which the traffic is routed directly from TIM network to PoliTO, exploiting a dedicated link at high speed and low delay. On the PoliTO NFV infrastructure side, both Symphony IoT Platform (collecting and elaborating both sensors' data and video cameras feeds) and the Control Room application backend (implementing the metaverse environment in which the different fields agent will interact) run as OpenStack VMs. Design and development aspects related to the Control Room application are reported in D3.2 [5].

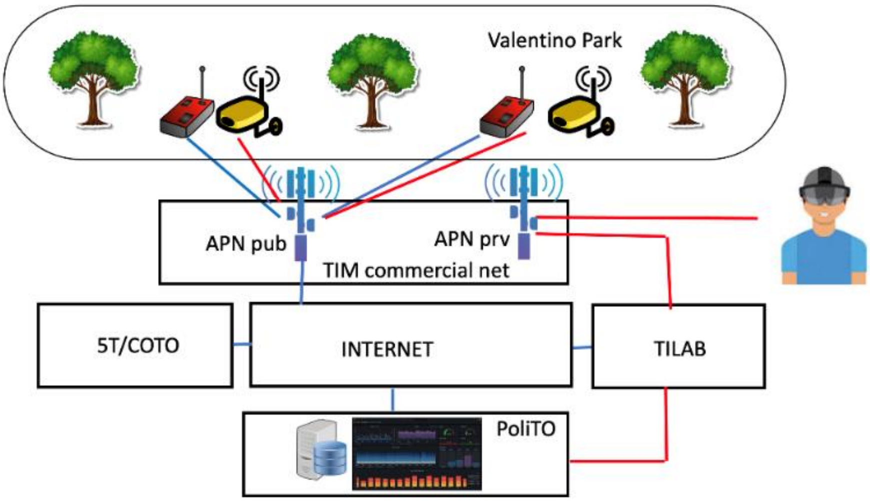

Figure 10. Ad-hoc architectural solution for UC5.

2.1.4 Preliminary network KPIs measurements

Due to the commercial network implementation, relevant network KPIs are affected by the network load imposed by commercial user distribution (density and location), so they can vary in different trial locations, different conditions (outdoor, indoor), and time (busy hours, night). Preliminary measurements mentioned in D2.1 [1] gave the feedback that use cases requirements were satisfied in most cases, accurate measurements will be performed in the actual location of trials execution in the first deployment of use cases.

2.2 Italian cluster (Pisa site)

The Italian cluster's Pisa site will be the hub for the deployment and validation of UC7 "Remote Proctoring," UC8 "Smart Ambulance," and UC9 "Hannes Prosthesis". These UCs will be supported by a 5G network that covers both indoor and outdoor experimental areas within the CNR campus in Pisa, the primary trial site. TIM site in Turin, will host part of the CN and the Ericsson Orchestration system.

Additionally, two other locations will play essential roles in the UCs' deployment such as the Fondazione Monasterio hospital in Massa, which will host UC7's training surgical room and UC8's emergency unit, and the Ericsson's 5G laboratory in Genoa, where UC9 will initially be deployed before being moved to Pisa.

The Figure 11 provides an update of the high-level architecture of the Pisa site already illustrated in D2.1 [1] in which the light-green boxes highlight the locations and systems used for UC7, the light-blue boxes refer to UC8, and UC9 is represented in light orange.

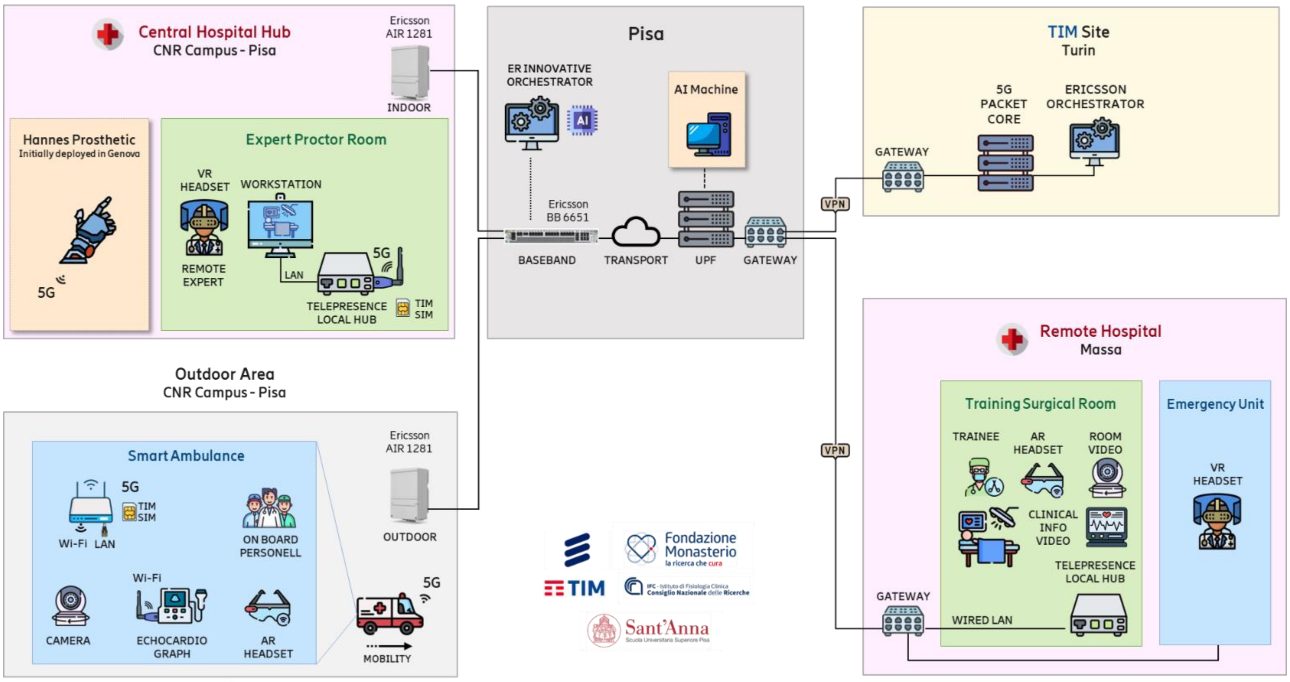

Figure 11. High level architecture for the Pisa site.

2.2.1 Site description recap

For the targeted objectives of the project's trials, Ericsson and TIM are working to deploy a dedicated private 5G network that provides comprehensive indoor and outdoor coverage as illustrated in Figure 12. Despite the fact the CNR campus in Pisa is already covered by TIM’s commercial 5G network, the use of a private solution is required to experiment the Ericsson Research (ER) Innovative Orchestrator described in Section 3.2.2 which would not have been possible to be installed otherwise. With reference to Figure 1, the ER Innovative Orchestrator represents the A5G of the Pisa site, relying on the Bs5G technology provided by the other parts of the network infrastructure described in Section 2.2.2.

With respect to the operating band, the private 5G network will use the 26 GHz spectrum band (with 200 MHz bandwidth channel) since using the 3.7 GHz band would have caused mutual interference between the private network and the commercial one. Finally, the chosen frequency spectrum and the provided network architecture, which anticipates a synergy between the radio and transport networks (coordinated by an orchestrator), is planned to ensure the required performance for all use cases. This network will boast enhanced functions based on its dedicated transport and orchestration features.

Figure 12 depicts the specific area within the CNR campus designated for experimental activities. The picture identifies an outdoor section for UC8 experiments, which will receive dedicated 5G service via an antenna mounted on the adjacent building wall. For UC7, a particular room within the campus building has been selected to host experiments, including a training session led by a proficient surgeon, remotely connected to the Massa Hospital. This same space will later serve as the testing ground for equipment pertaining to UC9, which will have been initially developed and trialed in Genoa.

Additionally, the CNR campus's primary server room will house the baseband radio system, the ER Innovative Orchestrator, and other systems for the connectivity with remote Massa and Turin sites. This space will also accommodate the AI server that will eventually support UC9. More details on the systems used in these locations are reported in the following sub-section.

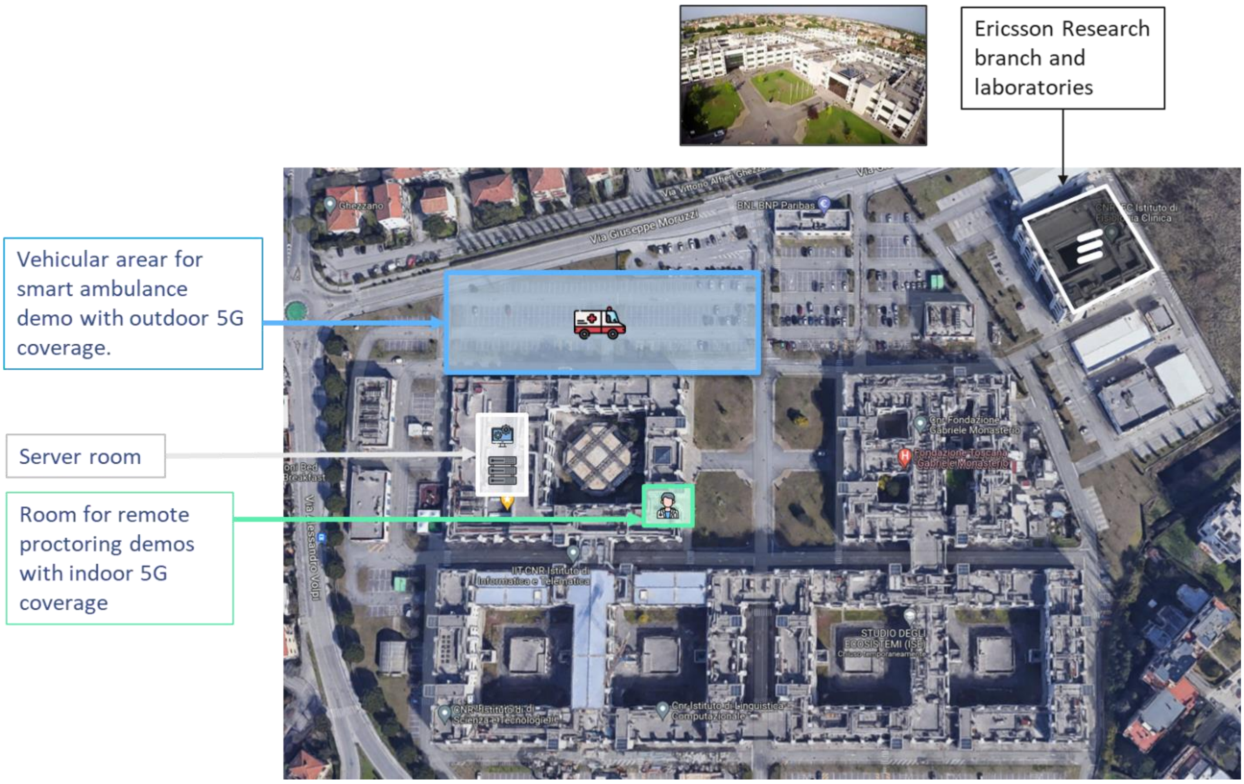

Figure 12. Overview of the experimental area with the main locations for experiments.

2.2.2 Design aspects

The deliverable D4.2 [6] provides details on UC7, UC8, and UC9, allowing the contextualization of the use of the radio network described in this section.

The antenna system selected for the trial is Ericsson AIR 1281 n258 [11]. It is part of Ericsson's Antenna Integrated Radio (AIR) product line. The product operates using Time Division Duplexing (TDD) with a downlink transmission time fraction of 75%. It is primarily designed for outdoor applications, but it can be used also in indoor scenarios. Two of these antennas will be used in the trial in support of UC7 and UC9 (indoor) and UC8 (outdoor). This radio systems natively support 3GPP Rel-16 and 3GPP Rel-17, in the trial, it will be used the mmWave band SA Rel-16, at 26 GHz on mmW spectrum (26.9-27.1 GHz, 200 MHz). The remote radio units are connected to the baseband system Ericsson Baseband Unit (BB) 6651 [12] which supports mmWave bands. This equipment will be hosted in a rack located in the server room.

The Edge 5G SA network functionalities include a local User Plane Function (UPF) breakout in Pisa and support for 5G Release 16 Slicing, specifically for enhanced Mobile Broadband (eMBB). The UPF is part of a Packet Core Gateway (PCG), which also supports non-3GPP functionalities such as Carrier-grade Network Address Translation (CGNAT), Firewall, and Multi Service Proxy. The hardware infrastructure is built around compute servers that are based on the Dell EMC™ PowerEdge™ R640 [13] rackmount server. The servers are designed to fit in a single rack unit with hot-swappable components and redundant power supplies and it will be hosted in the rack located in the server room. The virtualization platforms available support the deployment of both VMs (Virtual Machines) and Containers, relying on CaaS (Container as a Service) with SUSE K3s, and run on a SUSE enterprise Linux Host-OS.

As for terminals, TIM has selected the Customer Premises Equipment (CPE) ZTE MC889A SA (26GHz) to connect all the devices required for the UCs. This CPE provides a 2.5Gigabit Ethernet Port for wired connectivity. Two of these CPEs will be used for UC7, UC8, and UC9 respectively.

As part of the design phase, it is crucial to assess the 5G signal coverage in the outdoor area where the ambulance will operate, to pinpoint the precise location for antenna installation. Towards this goal, multiple coverage simulations have been conducted to evaluate the signal strength in the targeted area.. In Figure 13, the radio coverage footprint of the final installation is depicted.

The simulations indicate that the outdoor area of the CNR Campus, designated for the UC8, will receive optimal signal coverage. Once the 5G network becomes operational, in-field signal quality measurements will be undertaken to verify and ensure the signal's integrity. This validation process is vital to guarantee that the network performance provided to the ambulances is at its peak efficiency, given the spectrum band allocated for these experiments.

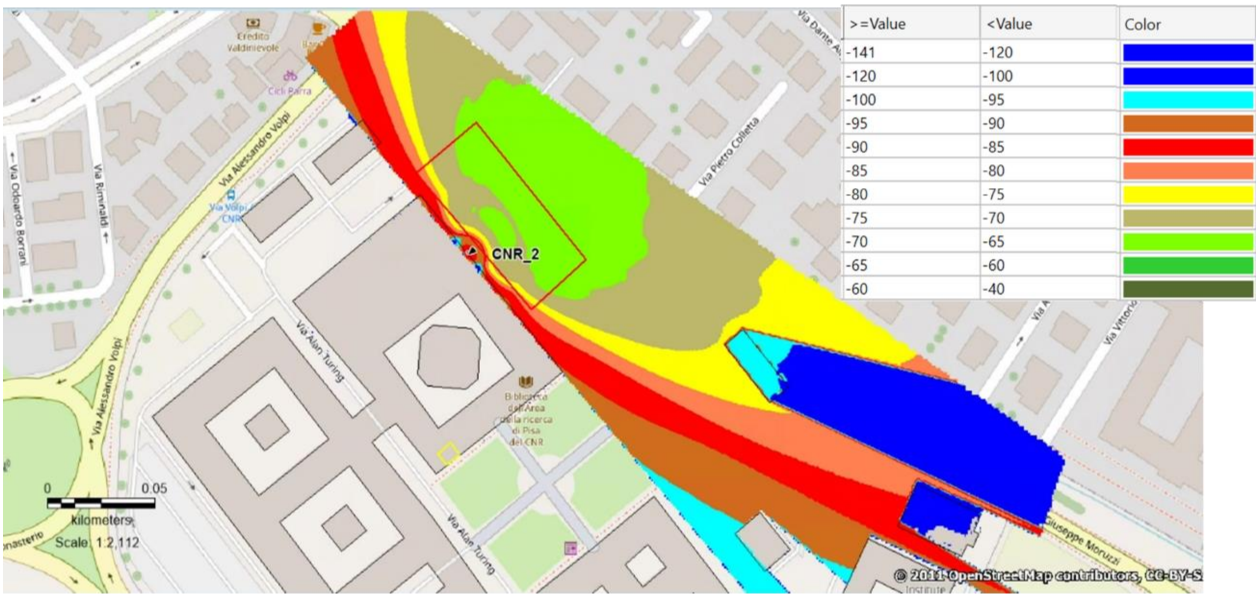

Figure 13. Coverage simulation of the CNR Campus outdoor area.

2.2.3 Initial deployment activity

The partners involved in the development of the Pisa cluster, under the coordination of TEI, have developed a detailed plan whose implementation is carried out and monitored with weekly alignment meetings. In the third week of January 2024, an inspection of the premises selected for the trial was carried out, and all practical aspects related to the cabling, antenna anchoring, signal visibility maximization (especially in relation to the outdoor area associated with UC8), and the required permits in the context of the CNR Campus hosting the trial were defined. In the months of April and May 2024, the installation and commissioning activities are planned, along with the availability of the CPEs identified for the experiments. By the end of May, the schedule envisages the possibility of conducting preliminary tests of the infrastructure and, subsequently, the start of the integration of eHealth devices and related applications within the network context.

2.2.4 Preliminary network KPIs measurements

The network planning and development is proceeding according to plan. However, the 5G network is still not active so testing activities have been limited to application level in the context of WP4. These activities are reported in D4.2 [6] also in relation to the KPIs.

The deployment and set-up of network infrastructure for Pisa and Massa site, including the activation of the 5G network in Pisa, is planned within Q3 2024, according to the reported time-plan. This has restricted testing activities to the application level, within the scope of WP4. Detailed accounts of these testing activities, alongside their correlation with the KPIs, are documented in D4.2 [6].

2.3 Spanish cluster

The UCs implemented in the Spanish cluster are UC1 “Smart Crowd monitoring” [2], UC6 “Mass Casualty Incident (MCI) and Emergency Rescue in Populated Area” [3] and UC10 “Immersive Fan Engagement” [4] and were carried out using the 5Tonic infrastructure with a 5G Standalone Rel-16 con-figuration operating in the mid-band of 3.7 GHz. During the tests, a lack of UL throughput and the need for latency/jitter improvement were evidenced, even using the full spectrum available in this band. Due to this, the infrastructure will be evolved to Rel-17, using new Hardware (HW)/Software (SW) to use 5G high-band, i.e., mmWave, as part of the A5G. This new deployment requires UE/CPE capable to be connected to a gNodeB in the FR2 frequencies without the support of FR1 frequencies (i.e., NSA mode).

2.3.1 Site description recap

Preliminary test have been performed in the 5Tonic infrastructure located at IMDEA Networks Institute (see Figure 14).

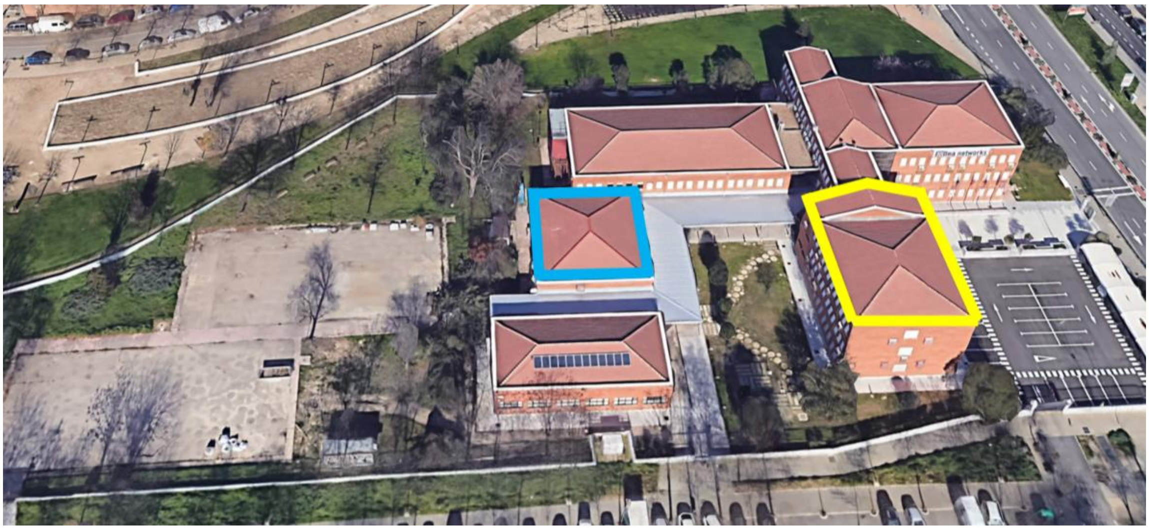

Figure 14. 5Tonic experimentation area.

The rooms where the experiments were conducted were:

- **X3 room** (blue): As described in D2.1 [1], the goal of this area is not only to have access to 5G technology but also to have access to the experimentation tools like for example the KPI framework or the graphical dashboard of the experiments and also an indoor area where devices can move throughout the room. This room will contain the new portable flight rack, which will be evolved to A5G integrating the new 5G band. Once tested, this flight rack will be installed in the Wizink Center Arena.
- **5Tonic lab room, 1S4** (yellow): This room is the main office and its coverage is provided using DOTs antennas (small cells) that operate in mid-range bands (same n78 band) for 5G New Radio (NR) and has indoor coverage purpose. This gNodeB will be kept as it is for Midband.

2.3.2 Design aspects

The 5G network solution remains unchanged since deliverable D2.1 [1]. The results of the tests indicate that the current mid-band configuration does not meet the requirements for the intended use, therefore an upgrade of the infrastructure to Rel-17 is necessary. To support the remote tests on the premises, a flight rack with high-band antennas and UPF on the Edge will be used. This will require the new antennas AIR5322 for mmW and a new Baseband6648.

The current solution includes 5G RAN equipment in mid-band (n78 - 3.5 GHz) and Stand-alone technology, 5G SA. Telefónica, provides this Midband spectrum to be used during the trials. The new Highband spectrum will be also provided by Telefónica. Software installed in RAN and Core components are 3GPP Rel-16 compliance. The hardware elements deployed in the rooms are:

- **X3 room**: The gNB consists of a small antenna Radio4408 n78L. Although the unit is designed for outdoor use, it can be used indoors. A baseband 6630 which provides switching, traffic management, timing, baseband processing and radio interfacing. It has the capability to be configured in mixed mode (more than one Radio Access Technology (RAT). It also includes a Router 6675 designed to provide high capacity pre-aggregation and aggregation routing in a compact and hardened one radio unit form factor. For time and phase synchronization required for 5G with TDD, a Global Positioning System

(GPS) receiver is used. This component consists of a Global Navigation Satellite System (GNSS) Active antenna.
- **5Tonic lab room, 1S4:** This gNB is built with the same baseband 6630, Router 6675 and GPS receiver as X3 room. In this case, the antenna system uses the smallest Ericsson's radio to cover indoor areas for NR 5G purpose, radio DOT (RD). It is a 4x4 Multiple Input Multiple Output (MIMO) antenna that offers an innovative and high performing solution that effectively connects indoor users to the whole mobile eco-system. This antenna requires equipment to provide the Radio Dot Interface (RDI) called Indoor Radio Unit (IRU). It connects to the baseband via Common Public Radio Interface (CPRI), and aggregates signals and provides power to Dots.

The new 5G RAN equipment in high-band (n258 – 26GHz) will be Rel-17 compliant:

- **Flight rack:** For the 5G SA high-band implementation (see Figure 15), a new portable flight rack gNode will be integrated. This gNB keeps the same Router 6675 and GPS as X3 room. The antenna will be an AIR5322 in 26GHz band (B258), an active antenna system with beamforming. It is designed for outdoor use, intended for pole, wall, or rooftop mounting, but it can be used indoor too.
- **Baseband**: The new Baseband 6648 is in 19-inch format and it has 9 CPRI ports, such as Evolved CPRI (eCPRI), enabling increased connectivity for radio units.

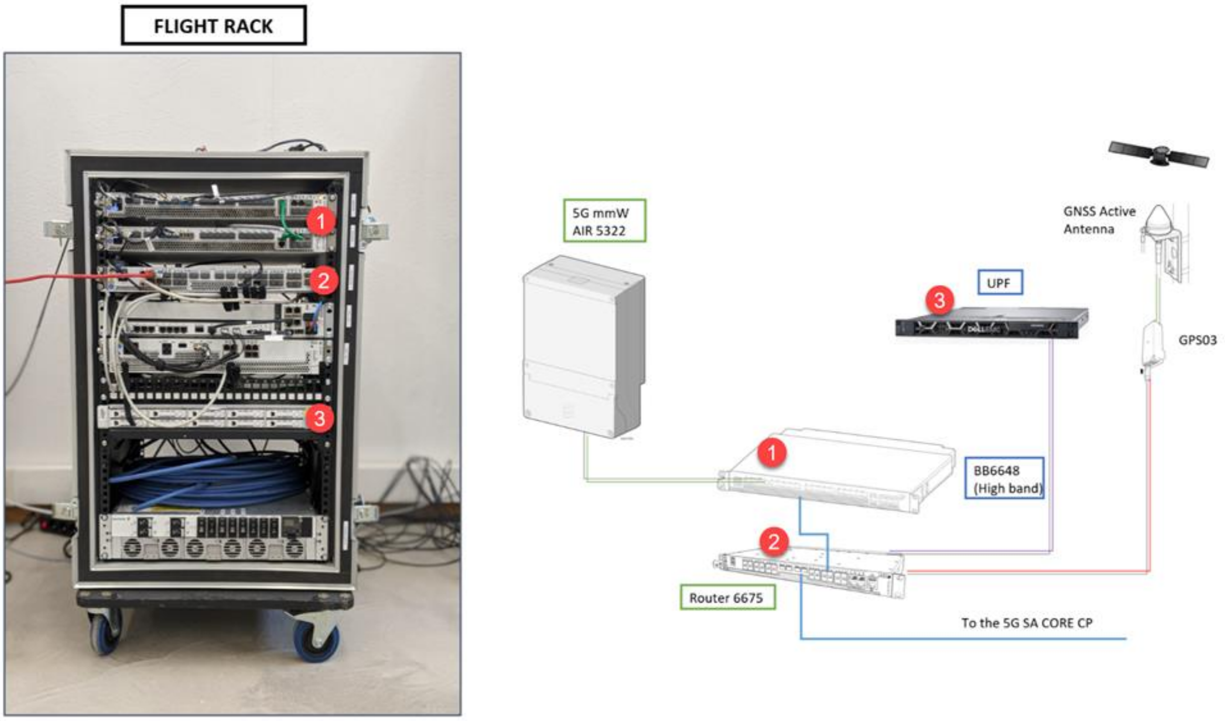

Figure 15. **A5G Flight rack configuration.**

To provide the connection to the various devices, two CPEs have been selected such as Askey NUQ3000M for UC1 and UC6, and Askey NUQ3000M and Teltonika RUTX50 5G router for UC10 respectively. The new A5G implementation will require new CPE capable to support 5G SA high-band.

2.3.3 Initial deployment activity

Development work in the lab was focused on configuring the 5G mid-band network to achieve the best possible performance for the UC1, UC6 and UC10.

Bandwidth tests were performed based on the best throughput performance. It was obtained using the highest available bandwidth in mid-band, 100 MHz.

Another important parameter is the TDD Pattern. To improve downlink throughput a TDD pattern DDDSDSU (10:2:2) will be used and to improve uplink throughput will be used DDDSUDDSUU (10:2:2:2:2). Although in both UC1 and UC10 devices generate more uplink traffic than downlink traffic, due to a regulatory issue all operators must use the same TDD pattern which is DDDSU (10:2:2:2).

The network tuning work was mainly focused on determining the optimal transmit power. In indoor cells this parameter has a big impact on the performance of the network due to saturation. The shape of the room, walls, materials, etc. have an important impact in the indoor propagation.

The X3 room has an RRU4408 antenna that can radiate a maximum power of 20 W, so tests were performed starting at 1W and going up to 10 W in 1W steps. With each power change, traffic was generated in the 5G network, and the maximum uplink and downlink throughput and latency were measured. The results obtained at power levels between 1 W and 5 W were practically the same, but from 5W onwards the network performance started to deteriorate due to saturation. So, for the X3 tests a power of 1W was set.

2.3.4 Preliminary network KPIs measurements

The 5Tonic 5G network has been operational for two years and during this period a lot of fine-tuning work has been done. The limits are now known and therefore, the KPIs have been measured directly for UC1, UC6, and UC10 and the results are detailed in deliverable D3.2 [5] (Section 3.1), deliverable 4.2 [6] (Section 3.1) and D5.2 [7] (Section 3.1) respectively.

2.4 Romanian cluster

2.4.1 Site description recap

The Romanian cluster hosts the deployment and validation of the UC1 “Smart Crowd Monitoring” and UC4 “Smart Traffic Management” and is based in Iasi, leveraging both the indoor 5G SA Rel-16 infrastructure from Bucharest and Iasi Orange 5G Labs as well as the outdoor 5G NSA commercial network provided by Orange Romania.

In relation to what was already described in WP2’s D2.1 [1] about B5G and A5G features of the Romanian testbed, ORO progressed with the implementation of Edge Computing and Orchestration features. More exactly, on the Edge Computing domain, ORO started working on the integration of the Iasi local UPF into the 5G SA Rel-16 Core Network existing within the Bucharest 5G Lab venue; this UPF will be configured to route all the traffic coming from the Iasi gNBs to the computing servers that will also be deployed within the TUIASI datacenter alongside the UPF.

In addition, ORO, together with IMEC and NXW are also working on bringing orchestration capabilities to the testbed based on the work already performed by the three partners within two other research initiatives [14]. The target platform will be tailored to the specific needs of the TrialsNet project, allowing for facilitated onboarding of third party applications and further integration with the ZSM and security modules that are to be developed by IMEC and NXW. Furthermore, based on the TUIASI applications that are currently under development, AIaaS capabilities could be offered to third party experimenters based on commonly agreed access and security policies by the end of 2024.

2.4.2 Design aspects

The final design of the 5G and Edge-Compute UC1 architecture was finalized and is indicated in Figure 16. The architecture will be implemented starting from the end of 2024 and will be based on the current in-field deployed hardware described in the following section. In this approach, the RAN equipment that covers the Stefan cel Mare pedestrian area will be upgraded to support 5G SA & NSA hybrid mode, allowing for the interconnection with ORO’s Bucharest 5G Lab 5G SA Nokia Compact Mobility Unit (CMU) Core Network for the control-plane traffic and access to the available compute resources.

The Iasi 5G Lab datacenter is the central point of this updated architecture, hosting the 5G SA local UPF unit and the edge-computing servers on which the UC1 video analytics application will be installed. The user-plane traffic coming from the Stefan cel Mare Blvd. 5G SA RAN units will be routed to the local UPF, and the video surveillance feeds will be transmitted directly to the Iasi 5G Lab datacenter edge servers over a dedicated Ultra-reliable Low Latency Communication (URLLC) slice that allows for a prioritized latency of about 8 ms E2E.

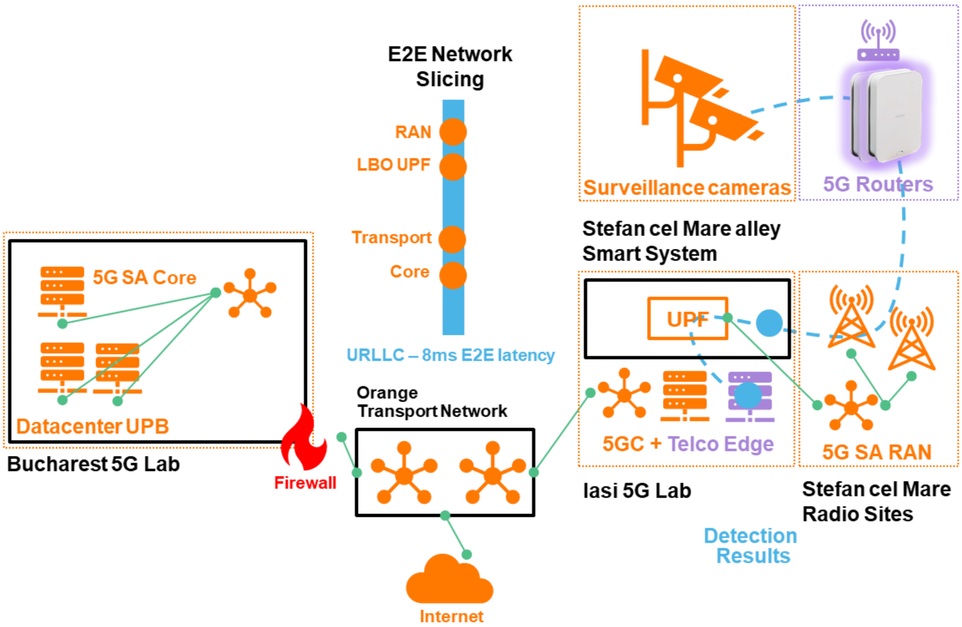

Figure 16. UC1 5G and Edge-Compute final architecture.

In the same manner, the final 5G and Edge-Compute designed architecture for UC4 is reported in Figure 17. As per UC1, the RAN units that are covering the Podu Ros intersection area will be also upgraded to support 5G NSA & SA hybrid mode and will be connected to ORO's Bucharest 5G Lab 5G SA Nokia CMU Core Network for control plane traffic and access to the available compute resources. Also resembling the UC1 setup, the user plane traffic that flows through the UC4 RAN units will be routed directly to the Iasi 5G Lab local UPF unit over an URLLC slice capable of an 8ms E2E latency, allowing for a rapid interconnection between the video surveillance feeds coming from the cameras and the edge-computing servers that host the video analytics application developed by TUIASI.

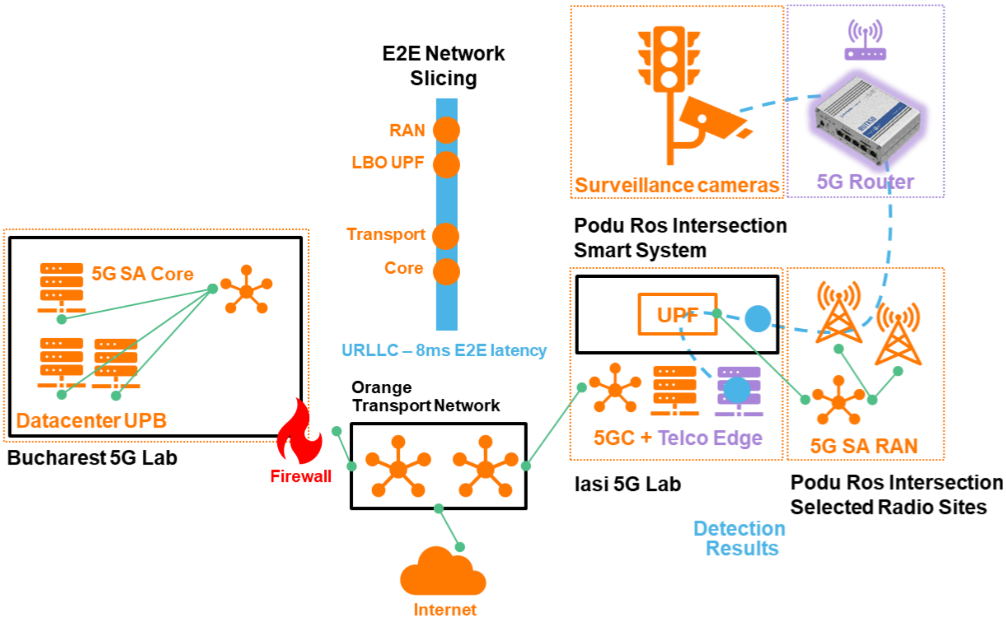

Figure 17. UC4 5G and Edge-Compute final architecture.

2.4.3 Initial deployment activity

In terms of the current infrastructure that is supporting the Romanian cluster use cases and is based on surveillance cameras, several updates have been performed since what reported in the previous deliverable D2.1 [1]. With reference to Figure 18, in the context of the surveillance cameras installations, Nokia Fastmile 5G14-B [15] 5G routers were installed on the supporting lighting poles and connected to the cameras via dedicated

CISCO switches as showcased in Figure 19. The 5G routers were fitted priorly with specifically provisioned SIM cards that support the TrialsNet project private APN and configured to host the connected cameras. Through the private APN, the video surveillance feeds can be reached directly from ORO's Bucharest 5G Lab infrastructure. To streamline the access of TUIASI to the cameras and feeds, an Internet Protocol Security (IP-SEC) Virtual Private Network (VPN) tunnel has been configured between ORO's 5G Lab and TUIASI's network infrastructure. Based on this architecture, the UC1 video analytics application runs now on TUIASI's servers, performing inference and data visualization tasks.

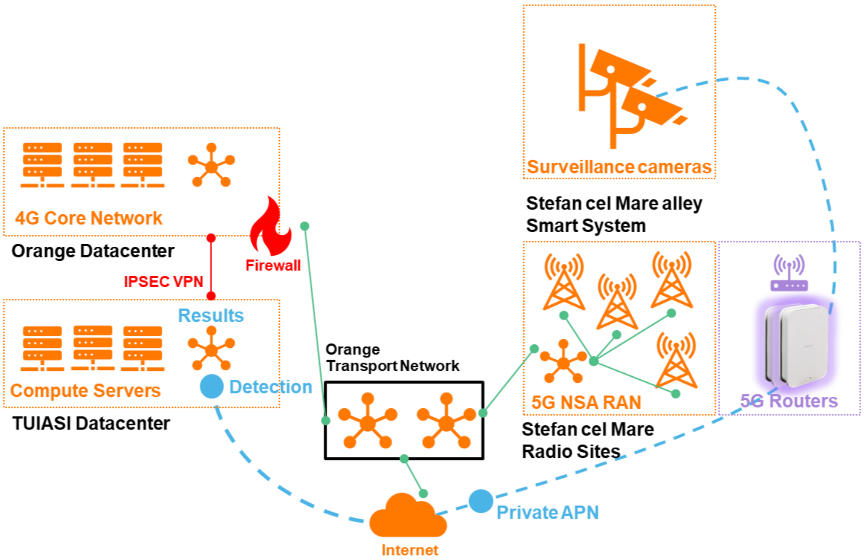

Figure 18. UC1 5G and Edge-Compute current architecture.

Figure 19. UC1 Cameras and 5G CPE installations.

Regarding the UC4 setup, the outdoor hardware infrastructure showcased in Figure 19 was finalized with the installation of a Teltonika RUTX50 [16] 5G router, that aggregates the traffic from all the previously deployed surveillance cameras from the Podu Ros intersection area (see Figure 21), in the metaled cabinet near the intersection. The 5G router is fitted with a SIM card provisioned with the TrialsNet project private APN, through which the video feeds are directly accessed by TUIASI computing servers, through the IPSEC VPN tunnel between them and ORO. Based on this architecture, the UC4 video analytics application runs now on TUIASI's servers, performing inference and data visualization tasks, as showcased in Figure 20.

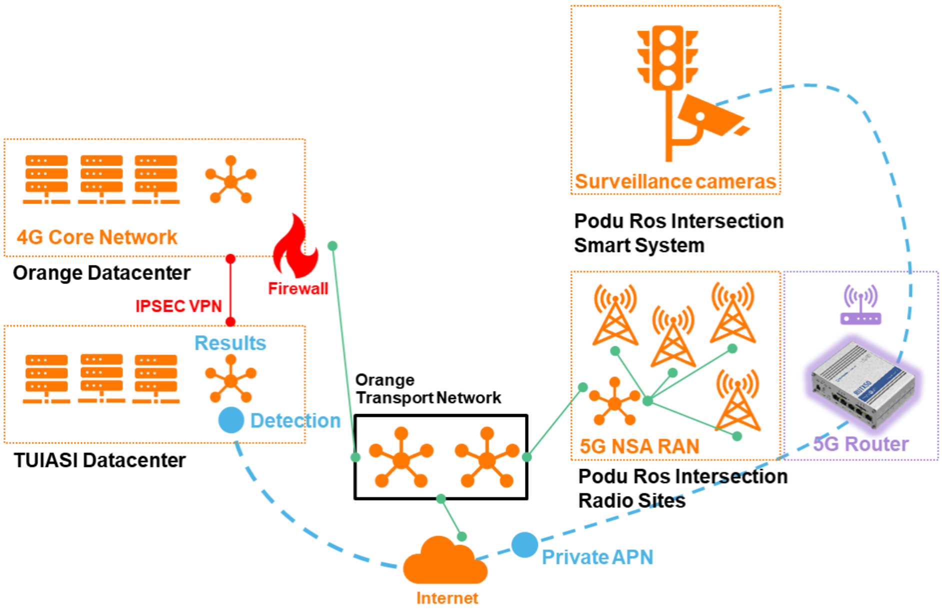

Figure 20. UC4 5G and Edge-Compute current architecture.

Figure 21. UC4 Cameras installations.

2.4.4 Preliminary network KPIs measurements

Several field tests were performed in order to assess the performance of the UC1 and UC4 current technical solutions. The throughput and latency performance of the network was assessed using the CPE provided testing capabilities, using Speed Test by Ookla and ping probes between the 5G CPEs and a virtual machine from the 5G Lab Edge Computer infrastructure, as already showcased in D3.2 [5].

During the live network tests performed in October and December 2023, ORO also measured network performance KPIs at the NR gNodeB level using proprietary monitoring solutions. The autumn results are depicted in Figure 22, and show that the peak number of users was reached during 14th of October on the UC1 radio site (BA0133), while on the UC4 radio site (BA0250) the average number of users was constant during the selected period. In regard to the uplink traffic, there was a peak starting from 11th of October that continued through 14th reaching 500 GBs per day/per site for UC1.

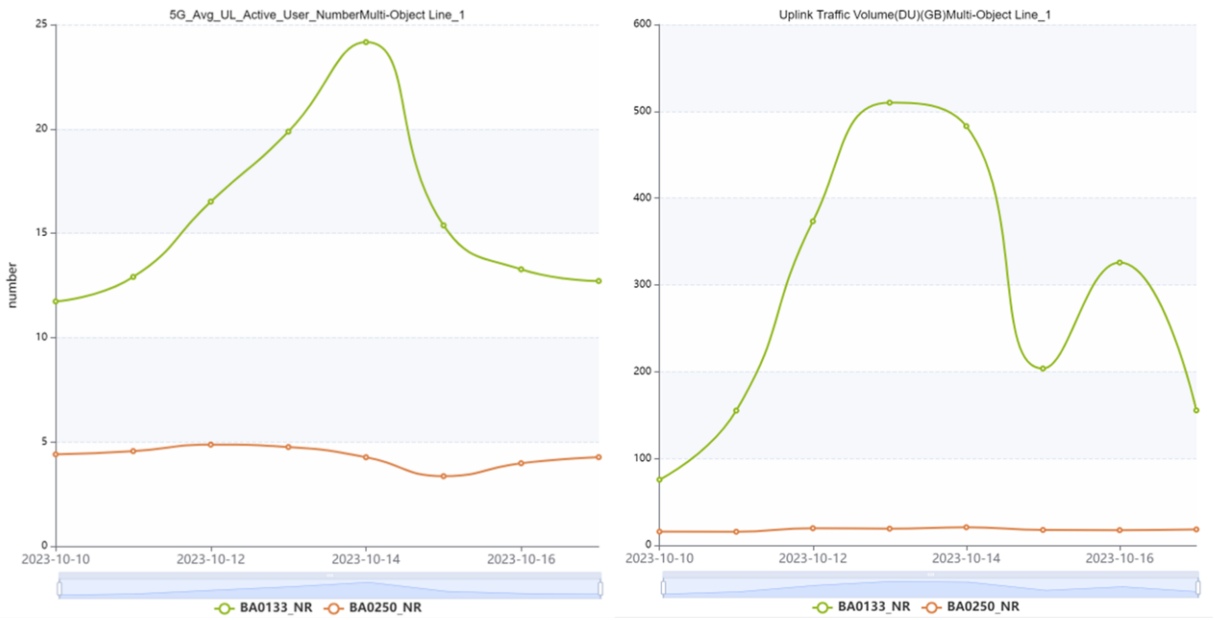

Figure 22. Measured number of users and uplink traffic at the NR gNodeB level during October tests.

The live network tests were repeated during the winter holidays, between 10th and 16th of December 2023, the corresponding results being depicted in Figure 23. As showcased, the numbers of 5G users increased for both the UC1 and UC4 radio sites, while the uplink traffic for the UC1 site peaked between 10th and 13th of December with an average of 200 GBs per day/per site.

Figure 23. Measured number of users and uplink traffic at the NR gNodeB level during December tests.

2.5 Romanian cluster (experimental facilities in Belgium)

As reported in D2.1 [1], the experimental facilities in Belgium refer to the testbed ecosystem managed by IMEC, comprising a set of small-scale test sites located in Antwerp and Ghent (Belgium). This section briefly describes the testbeds that are used in the project activities (development of horizontal innovations and UC4), highlighting the updates from the state reported in D2.1 [1]. In addition, it has to be highlighted that this experimental site has been proposed as a completely standalone (i.e., independent from the Romanian cluster) testing setup with network and virtualized infrastructure resources to support the implementation and trials of third parties use cases in the context of the project's Open Call.

2.5.1 Site description recap

The **Smart Highway** experimentation site is designed for wireless networking research in the automotive context, offering Vehicle-to-Everything (V2X) technologies, as well as distributed and edge computing capabilities. It is deployed on the Antwerp highway E313, as a set of seven distributed Road Side Units (RSUs), which, in combination with Onboard Units (OBUs) deployed in test vehicle(s), create a unique real-life environment for experimentation in a V2X realm. The computing-communication continuum consists of far and extreme edge computing units placed in the RSUs and OBUs, respectively, as depicted in Figure 24. This testbed is used for testing and validation of advanced network capabilities such as edge computing and ZSM and orchestration as part of A5G, which will be further transferred to Iasi trial site for the purpose of trialing UC4.

This communication-computing continuum is orchestrated using end-to-end orchestration techniques, both in terms of resources and services that are deployed to serve the end users, i.e., vehicles, leveraging the technologies such as Kubernetes [17] as a baseline. The application services running within the continuum aim to deliver the required levels of Quality of Service (QoS) to the users. This experimental facility is used for experimentation with zero-touch service management tools, therefore, it provides means for using Artificial Intelligence (AI)-based zero-touch services that help native Kubernetes-based orchestrators to make more efficient orchestration decisions that in turn result in improved service performance. On top of that orchestration framework, IMEC is deploying an intent-based network and service management layer, which bridges the gap between user/service requirements in terms of expected performance (end-to-end latency, uplink/downlink throughput, reliability). Following the progress on the KVI definition and measurement methodology in WP6, reported in D6.1 [18], the intent-based management will subsequently include specific KVI requirements related to sustainability into the decision-making process, and thus, the deployment and orchestration implications on the services.

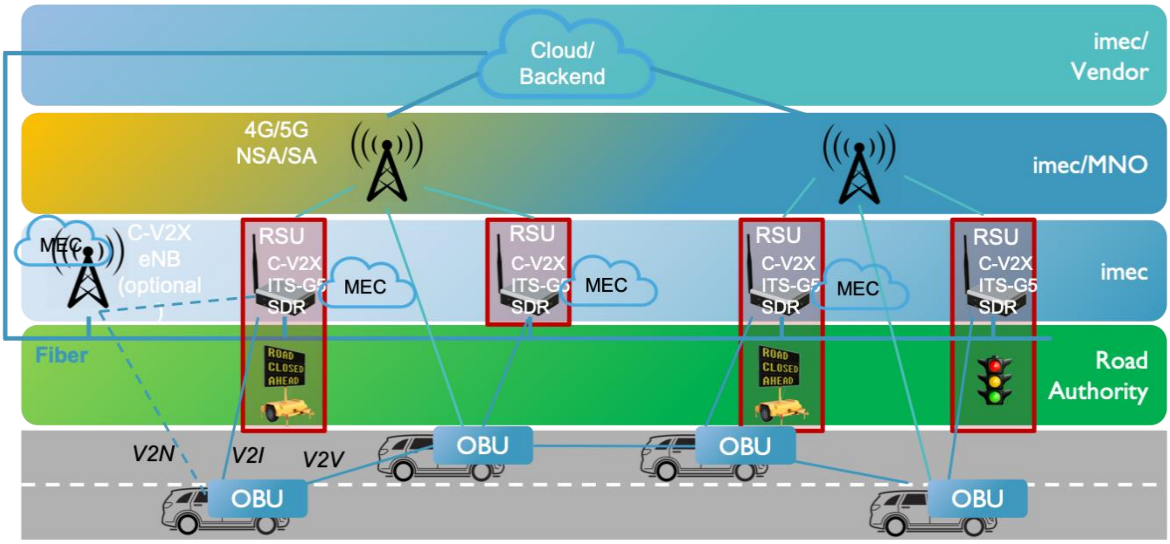

Figure 24. Overview of the Smart Highway testbed (technologies, stakeholders).

Finally, Figure 25 shows the coverage area of the Smart Highway testbed, i.e., the site where the experimental license has been issued to IMEC by the Belgian regulatory body Belgisch Instituut voor postdiensten en telecommunicatie (BIPT), for the frequency range 4035-4085 MHz (n77 band).

Figure 25. Coverage area of the Smart Highway testbed for which an experimental license is issued.

The **5GOpen testbed** is currently deployed as an indoor testing site (Figure 26), located at IMEC offices in Antwerp, thereby offering indoor and temporarily outdoor 5G Standalone (SA) connectivity. The testbed is fully compliant with OpenRAN initiative, thereby being built with the open source solutions on top of the off-the shelf equipment. The testbed consists of multiple gNodeBs deployed throughout the seventh floor of the building, and a 5G Core Network (5GCN) deployment, leveraging Open Air Interface (OAI) for both radio and core, Open5GS [19], and Free5GC [20] for core, and srsRAN [21] for radio. This testbed is being upgraded with new functionalities (more edge computing nodes, base stations, sensors) to support new releases of 5G and to enable testing of novel AI-based network optimization techniques towards 6G. As such, it is used in combination with the Smart Highway testbed for the purpose of evaluating performance of zero-touch service management techniques, but with more focus on the network performance evaluation. The setup is also installed on one instance of RSU located at the University Campus, offering 5G SA capabilities outdoors in the granted frequency range.

Figure 26. Open5G testbed with coverage maps.

The **IMEC-Ghent 5G testbed** is currently deployed within IMEC's premises in Ghent, facilitating initial development and testing. This testbed employs a heterogeneous network consisting of a 5G NR and a Wi-Fi network. The 5G NR network is implemented by Open5GS for the core and srsRAN for the radio and the Wi-Fi network is implemented by commercial equipment. The NR radio link is established using Universal Software Radio Peripheral (USRP) [22] in the gNB and Quectel RM500Q-GL in the UE.

2.5.2 Design aspects

The RSUs within the **Smart Highway** testbed are installed on top of the gantries along the E313 highway in Antwerp, and they are connected to the fiber network of the road operator, which provides them with a stable backhaul connection for edge processing. It is important to note that each of those RSUs is fully accessible and managed remotely, which is particularly important for external experimenters.

The software and hardware capabilities of these RSUs, as well as the OBU, are already reported in D2.1 [1]. Nevertheless, to facilitate the connectivity of Smart Highway BMW test vehicle to 5G SA network, which could be used interchangeably with both Smart Highway and Open5G testbed (to test service performance outdoors but in close proximity from the indoor setup, Figure 26), IMEC created a mobile User Equipment (UE) box with 5G modem (Peplink), GPS device, and camera. This box is shown in Figure 27, and as such could be easily transported to any test vehicle (not limited to cars only) to provide computing and communication capabilities with a basic camera installation that could be used for obstacle detection purposes. The computing unit installed in the box is NUC-based [23], and as such it provides extreme edge capabilities, with capabilities to offload tasks to far edge (Smart Highway RSUs).

Figure 27. Box with the 5G Modem (Peplink), GPS, and camera.

Concerning **Open5G testbed** specifications, all capabilities are reported in D2.1 [1], whereas the visualization of all testbed components is illustrated in Figure 28. In particular, Figure 29 shows the planned upgrades and deployment of new base stations and UE testing units in the indoor environment. For example, UEs are equipped with both Intel NUC [23] computing units and Quectel 5G modem [24]. This setup is planned to be used together with the Smart Highway testbed to test 5G SA network performance and performance of automotive services over 5G, as a result of applying AI-based orchestration techniques in the context of zero-touch service management.

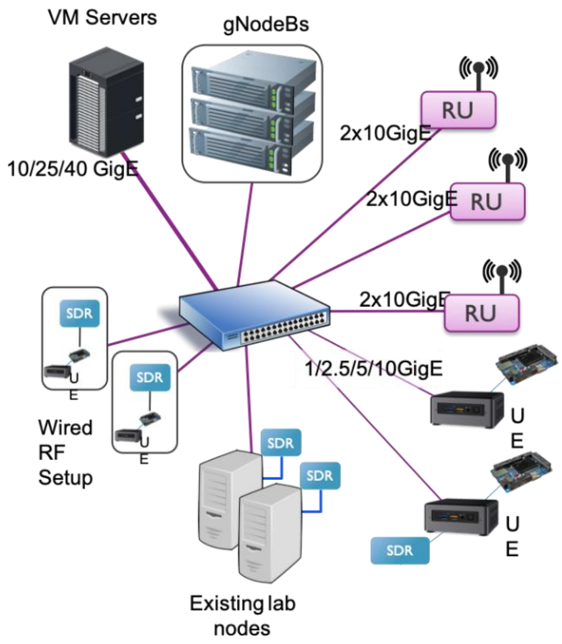

Figure 28. Components of the Open5G testbed.

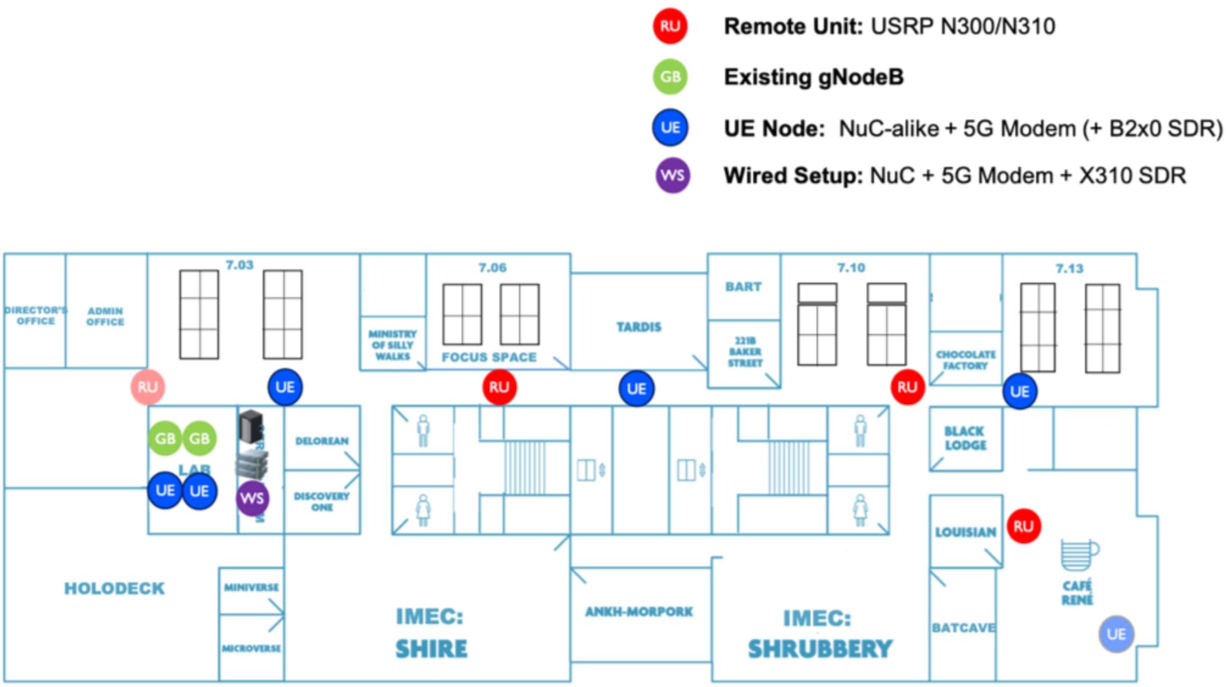

Figure 29. Upgrades of Open5G testbed.

In the **IMEC-Ghent 5G testbed**, various hardware and software, illustrated in Figure 30, are utilized. The setup employs a server and client setup, both utilizing Intel Next Unit Computing (NUC) as the endpoints for the connection. Hardware components such as 5G radio units, network switches, and Wi-Fi access points have been deployed. This setup is planned to be used for the initial development and testing of the Multi-Path Transport Control Protocol (MPTCP) solution for enabling the concurrent transmission of data over multi-RAT networks. The main design focus is not merely limited to utilizing MPTCP, but to delve network optimization for the efficient transmission of video data.

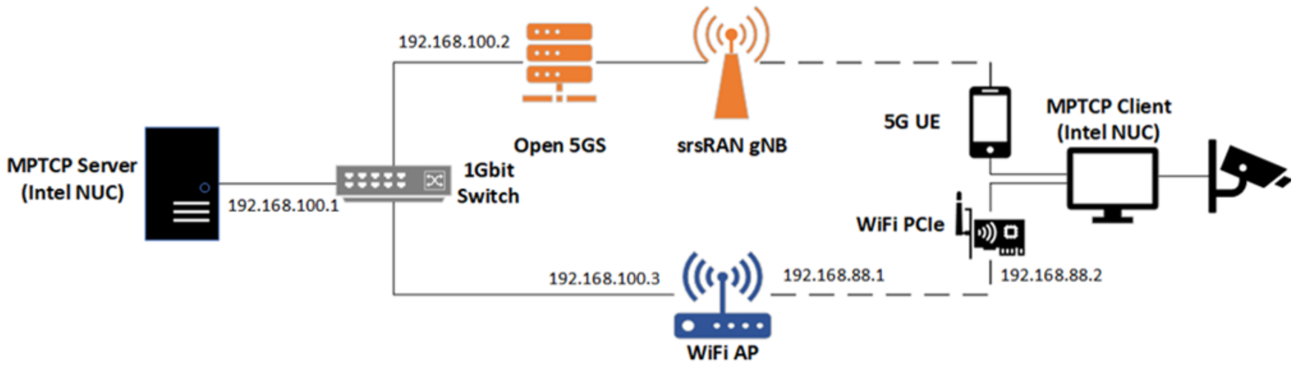

Figure 30. Schematic of IMEC-Ghent 5G testbed.

The development and implementation of an Intelligent network selection algorithm based on MPTCP are critical advancements in enhancing connectivity and network performance. In scenarios where reliable and real-time data transmission is paramount, the ability to intelligently select the optimal path can significantly improve the efficiency and reliability of communication networks. This becomes especially crucial in use cases demanding high reliability and minimal latency (such as UC1 and UC4), as it ensures that data is transmitted in the most efficient manner possible.

By integrating MPTCP, which allows for the simultaneous use of multiple network paths for a single TCP stream, the algorithm will intelligently distribute traffic across available network paths, thereby optimizing the use of network resources and improving the quality of video streaming. This selection is done based on a set of metrics, including throughput, latency, network reliability, and congestion levels. This scenario is illustrated in Figure 31, where an algorithm running at an endpoint (client) is tasked with selecting the optimal network based on the KPIs it measures from the different RANs, i.e., Wi-Fi and 5G NR.

The development of this algorithm is a key future direction, promising to significantly enhance the capabilities of the network solutions by ensuring high performance, reliability, and optimal resource utilization across all network conditions. The system can also benefit from predictive analytics for network conditions, enabling it to anticipate network congestion and adjust the traffic distribution accordingly. This can be done through advanced AI/Machine Learning (ML) algorithms. For instance, neural networks can identify complex patterns in network metrics and learn over time which network paths tend to become congested at certain times of the day.

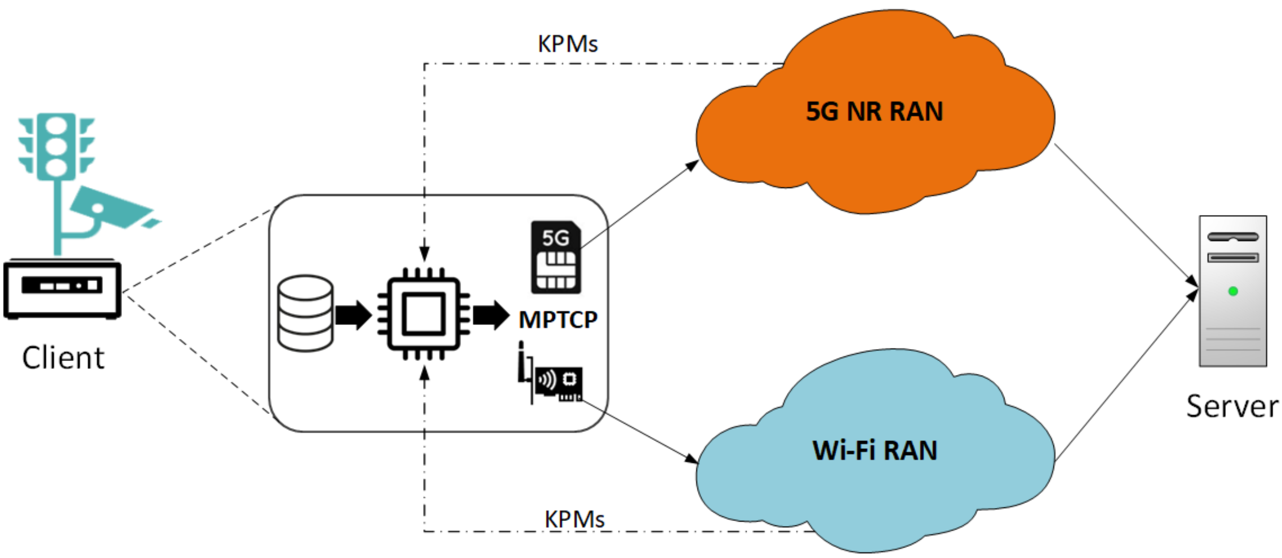

Figure 31. Intelligent network selection based on MPTCP.

2.5.3 Initial deployment activity

The on-field deployment of Zero-touch Service Management (ZSM) framework, presented in more details in Section 3.1.1, is created as a Proof-of-Concept on top of the **Smart Highway testbed**, as illustrated in Figure 32. This testbed enables collecting data based on realistic environments thanks to its seven RSUs deployed on a section spanning four km long of the E313 highway in Antwerp (Figure 32). The RSUs are computing-capable units, and they support hosting containerized applications such as vehicular services, media content publishing, resource monitoring, such as RSUs availability based on their Central Processing Unit (CPU) load, memory workloads and latency, decision-making processes and orchestration. Each RSU used in this on-field deployment features the following:

- Radio and antenna units to support wireless communications with Cohda MK5 and MK6c [25] over IT-G5 and Long Term Evolution (LTE) V2X [26] in the bands 5.9 GHz through a PC5 interface, and radio with Peplink 5G2 [27],
- Computing processing units provided by General Purpose Computing Units (GPCUs) that enhance the edge computing capabilities with Intel Xeon 8 Cores and 32 GB Random Access Memory (RAM),
- Power control unit to provide and manage energy throughout the RSU components,
- Backup and recovery to deal with events of failures,
- Remote access to manage the deployed setups,
- An On-board Unit (OBU), which is a mobile unit BMW X5 xDrive25d LO equipped with similar appliances as the RSUs, to evaluate the end-to-end latency of the RSUs which is fundamental for the performance of the vehicular services.

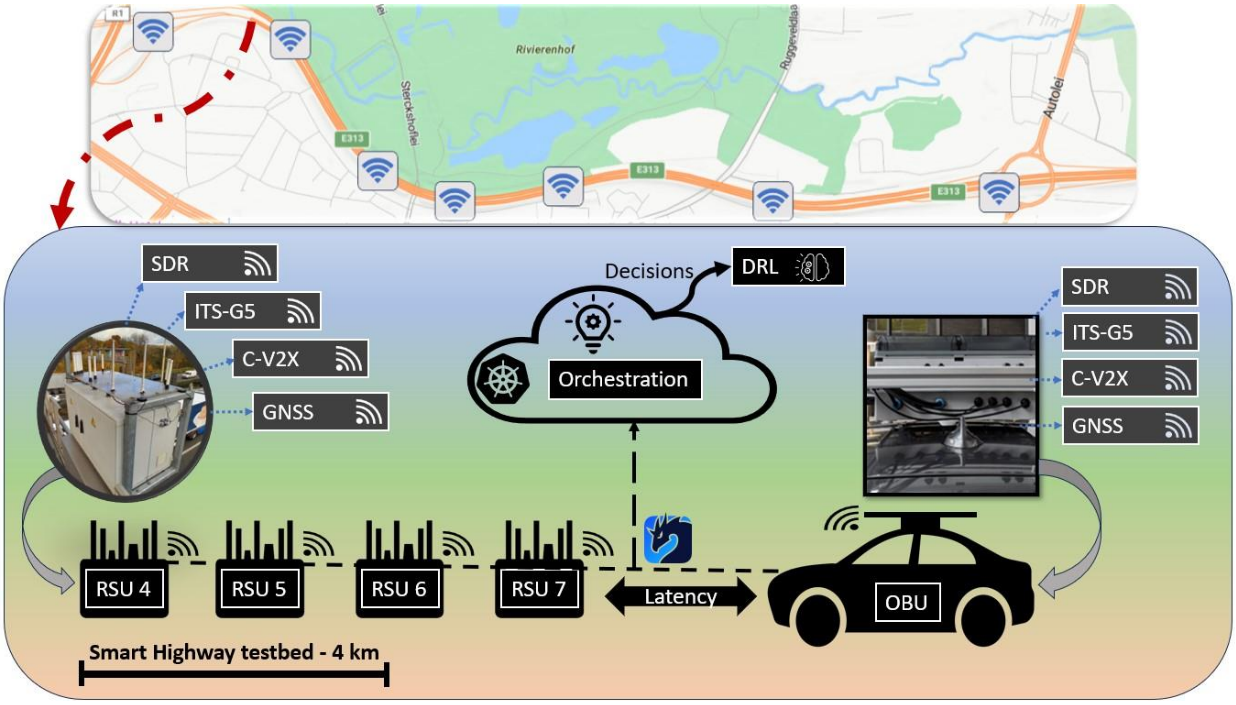

Figure 32. Experiment setup on the Smart Highway testbed located at the E313 highway in Antwerp, Belgium.

This experimental setup has been used in various testing scenarios for ZSM framework and vehicular services as container-based application deployed across the communication-computing continuum based on Kubernetes. More details about this setup are provided in Section 3.1.1, which focuses on advancements and updates related to the TrialsNet horizontal innovation, i.e., ZSM for enhancing performance and sustainability of vertical services.

To make a decision on the placement or life-cycle management of such services, the ZSM framework relies on the real-time monitoring of data (computing metrics collected from each RSU such as CPU, memory, storage, and network end-to-end latency measured at the client side), which is fed into simpler rule-based zero-touch services and the more advanced ones based on Deep Reinforcement Learning (DRL). The zero-touch services learn the relationship between computing load and end-to-end network and service performance, and while predicting the future load, they pre-empt improvements on the service placement or resource allocation, trigger migration and scaling operations before the service performance deteriorates and client experiences drops in QoS.

In addition to ZSM framework deployment on the Smart Highway testbed, Open5G is concurrently used for 5G network performance testing and dynamic network slice configuration. These capabilities of Open5G will be leveraged in the Q2 and Q3 of 2024 to extend the ZSM framework with automatic and intelligent network slice configuration. As additional tests are ongoing and are planned for Q3, the implementation progress and results will be reported in D2.3.

The deployed **IMEC-Ghent 5G** setup is instrumental in validating the theoretical constructs of the hybrid multi-Radio Access Technology (RAT) communication system. The controlled environment deployment enables the assessment of various KPIs concerning MPTCP and routing protocols. In alignment with the projected roadmap, a preliminary on-field deployment is set to commence in collaboration with ORO at the UC4 location. This next phase is designed to replicate the controlled environment setup at IMEC, but on a larger scale and in a real-world setting, to further validate the system's efficacy and performance.

2.5.4 Preliminary network KPIs measurements

Preliminary results obtained using the Smart Highway testbed are related to end-to-end latency measured at the client side and are reported in Section 3.1.1. As preliminary testing campaigns are ongoing, more results will be reported in the next deliverable.

2.6 Greek cluster

This section provides an updated description of the Greek cluster where the UC2 “Proactive Public Infrastructure Assets Management”, UC3 “Autonomous APRON”, UC6 “Mass Casualty Incident (MCI) and Emergency Rescue in Populated Area”, UC11 “Service Robots for Enhanced Passenger’s Experience”, and UC13 “Extended Reality (XR) Museum Experience” are going to be implemented.

In the next sub-sections, there is a recap of the site description, main design aspects about hardware and software components of the infrastructure (in terms of platforms and network solutions), initial deployment activity and preliminary KPI measurements.

2.6.1 Site description recap

As mentioned in TrialsNet D2.1 [1], UCs of the Greek cluster are trialed at the Athens airport (i.e., UC3 and UC11) and at public venues in the city of Athens (i.e., UC2, UC6, and UC13). In the Greek cluster a public 5G network is used, for leveraging its high-speed connectivity, low latency, and wide coverage. Specifically, the network uses NSA architecture and operates at a frequency of 3.5 GHz. The allocated band for this network is 80-100 MHz, which provides high-speed connectivity and low latency to support the data-intensive applications required by the UC. Future versions of the public network will also be used. Also, the integration of AI as a Service (AIaaS) continues in order to empower the relevant UCs with advanced AI capabilities, enabling them to leverage AI algorithms and models to enhance their operations and applicability. Such algorithms include AI for image recognition as well as efficient resource management and diagnostics.

In addition, a WINGS owned, private network infrastructure is utilized in order to conduct testing activities, validation and demonstration, prior to the deployment in the field. All Greek site use cases have been tested in the WINGS testbed.

The WINGS testbed provides E2E 5G/B5G functionality, along with extensive cloud and edge computing capabilities, leveraging the 3GPP (Release-15 and beyond) Public Network Integrated Non Public Network (PNI-NPN) with shared CP (at a first phase) and isolated, Stand-alone Non-Public Network (SNPN), with all NFs (UP and CP) inside WINGS premises, isolated from the public network in the final phase. The site offers a range of 5G/B5G services and will be gradually evolved to 6G. It supports various vertical domains, with WINGS providing the necessary HW, software, and configurations to enable the testbed to handle these UCs. WINGS testbed serves as a testing ground for services, equipment, and new features before they are commercially released.

WINGS has progressively extended the existing software, HW, and network functionality to support Cloud, Mobile Edge Computing (MEC), Extreme Edge and IoT functionalities. WINGS has demonstrated advanced

UCs on DTs, Collaborative Robots with native-AI B5G/6G capabilities of the system. Also, the required frameworks to build, test, and validate innovative 6G applications are part of the overall infrastructure.

The WINGS testbed utilizes AI mechanisms to support diagnostics, intelligent management, and orchestration. The management of the facility is done using a combination of existing and new software, covering DevOps, AI/MLOps, monitoring, profiling, diagnosis, and service-aware resource allocation and orchestration. An AI-enhanced MANO component inherited from 5G-TOURS project [28] is used to enable advanced automation and optimization. Monitoring, profiling, and diagnostic components provide information on available resources and network capabilities to help find the optimal deployment of a vertical service.

The WINGS testbed utilizes open-source software such as OpenStack, Kubernetes and Open-Source Management and Orchestration (OSM MANO) to provide a flexible and scalable infrastructure for verticals, as well as other open-source tools like Kafka, MQ Telemetry Transport (MQTT), and Robot Operating System (ROS). These components can support Virtual Machines (VMs), containers, and serverless execution of code from cloud to extreme edge devices such as raspberry pi.

The WINGS testbed prioritizes protection and privacy-preserving mechanisms to ensure reliability, security, privacy, confidentiality, and integrity of data. Open-source tools like Open-Source Network Access Control (OpenNAC) for network access control is utilized. Strong authentication, user management, and secure services will be provided with minimum effort from the verticals. The solutions selected will comply with General Data Protection Regulation (GDPR) article 5 to ensure appropriate security and protection against unauthorized or unlawful processing, accidental loss, destruction, or damage. They will employ consolidated Public Key Infrastructure (PKI) certificate-based cryptographic systems for critical communications to encrypt personal data during transmission.

2.6.2 Design aspects

Figure 33 shows the WINGS private testbed architecture. In this figure the Campus network is deployed on 3 floors of the building that hosts WINGS in Athens. The Campus network is connected via secure interfaces with the Core network of an Operator and through it to the Internet. The installed small cells work in 3.5GHz 5G band and the operator's Core network currently is NSA and will be upgraded to SA in due time. WINGS campus network users (SIM cards) have access with networking only in the Campus network coverage area, while access to the public network is not available. Public network users have no access to WINGS Campus network. The operator is responsible for delivering the traffic data from Campus network to specific servers in WINGS premises.

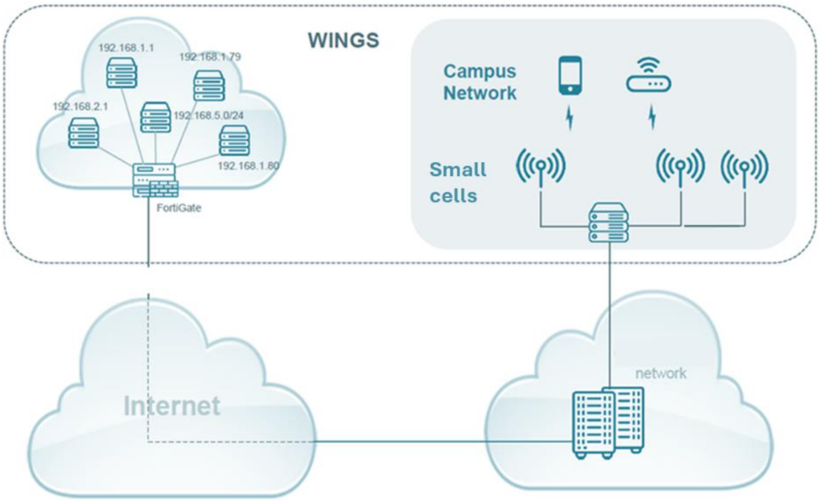

Figure 33. WINGS private testbed architecture.

2.6.3 Initial deployment activity

The purpose of this work is the development of a private high access network specifications (Private Campus Network) for the environment of the company WINGS which is based on the latest 5G mobile technologies and will provide the required coverage, capacity and connectivity to the end user for access in a set of various applications of innovative solutions developed by company in the context of TrialsNet project. In the first year, new equipment has been installed as mentioned in the table below. The applications (use cases) that WINGS intends to utilize in the Campus network concern Internet of Things technologies (IoT) and Robotics with 5G connectivity for low response time communication and high bandwidth. These communications include (indicative) control data, measurements, audio and video. Some of these applications are hosted on the premises and WINGS equipment (closed loop control) and others will be routed by the network of WINGS in cloud. The current list of equipment and devices used in the Greek site for the supported use cases is presented in Table 1, providing three new private small cells. It should be noted that the list may be augmented as the implementation of infrastructure progresses.

Table 1. Networking equipment and devices deployed in the Greek cluster (WINGS private 5G network).

Equipment	Item	Description
	Small cell	Three small cells have been deployed in WINGS testbed in order to create a private cellular network as described in the Design aspects.
	Indoor radio units	The Indoor Radio Unit connects to the baseband via CPRI, and aggregates signals and provides power to small cells.
	Baseband unit (BBU)	The BBU is responsible for communication through the physical interface.

2.6.4 Preliminary network KPIs measurements

Some preliminary results have been measured with the usage of iperf. For the results the new 5G private network of WINGS was utilized. Special SIM cards were used in a mobile phone in order to connect to the private network. Application-related results are analyzed in the deliverables D3.2 [5], D4.2 [6], and D5.2 [7].

Figure 34 provides the measurement of end-to-end latency which is measured in the WINGS 5G private network. The average value is 24ms.

Figure 34. Measurement of end-to-end latency.

3 Preliminary development of TrialsNet Innovations

This section reviews the latest updates on the research and development activities related to horizontal and vertical innovations as defined in D2.1 [1]. In particular, horizontal innovations refer to transversal or horizontal 6G technological/network enablers that are applicable to various vertical contexts, aiming to advance their performance and sustainability of the overall ecosystem (5G/6G network and vertical infrastructure). On the other hand, vertical innovations are more specific to particular use cases and trial deployments, thereby focusing on use case-specific challenges characteristic for one or subset of verticals. Both types of innovations are studied and reported in WP2, but as such are applicable to various use cases from WP3, WP4, and WP5.

With respect the vertical innovations reported in D2.1 [1], this deliverable introduces a new innovation related to the use of the cellular network information for crowd and traffic monitoring.

3.1 Horizontal Innovations

This section provides insights on the horizontal innovations tackled in the TrialsNet project. The content described in this section is in reference to the D2.1 [1] and the respective input provided on state-of-the art and specific context in which these horizontal innovations and technologies are being used in the TrialsNet framework. Detailed information about the intent-based network management as part of the ZSM framework are reported in the Annex A, while the background for the different application frameworks presented in Section 3.1.2 is provided in Annex B.

3.1.1 Zero-touch service management

Zero-touch Service Management (ZSM) techniques are becoming an indispensable part of A5G, and 6G systems, as they are critical for ensuring the proper functioning of vertical services, by granting the stability and performance of the network [29]. For example, in the context of Smart Traffic Management (UC4), mission-critical services that aim to ensure safety of Vulnerable Road Users (VRUs) at the busy intersections, have stringent performance requirements such as ultra-low-latency and extremely high reliability that need to be met regardless of the network load. Therefore, the goal of the ZSM framework, i.e., the intelligent and automated network and service management operations, is to make proactive service placement and runtime service reconfigurations that will in turn result in reliable vertical service operation (for UC4 services or any other vertical service with strict performance requirements).

One of the key goals of ZSM is to enable automated network management across different vendors and domains. To achieve this, a common interface is needed for data analytics and closed-loop control components. This interface facilitates the creation, execution, and governance of single or multiple closed-loops within end-to-end networks and supports the adoption of ZSM by the Mobile Network Operators (MNOs). The ETSI introduced the ZSM Industry Specification Group (ISG) in 2017. This group is dedicated to the design of a framework for network automation based on ZSM principles, such as the specification of solutions for the orchestration and automation of network resources allocation [30]. The design of the ZSM architecture reference shown in Figure 35, supports open interfaces, model-driven services and resource abstraction to provide modularity, flexibility, scalability and extensibility. In this figure, the general structure proposed in the principles of ZSM [31] is illustrated, where the domains are defined and managed by an End-to-End (E2E) Service Management Domain. This figure displays the composition of one single Management Domain as a templated pattern to be followed in each Management Domain within the architecture of ZSM.

One of the core ideas from the ZSM principles is the separation of concerns. This means that each management domain should be defined by clear boundaries around administrative, geographical, or technological factors [32]. Each of these management domains is structured into the following blocks as seen in Figure 35:

- **Management Functions**: to provide capabilities through the exposition/consumption of endpoints,
- **Data Services**: to enable the integration of authorized data management across domains for persistence and sharing data,
- **Domain Integration Fabric**: to expose and control the management services beyond their domain boundaries,

- **Cross-domain Integration Fabric**: it plays the roles of both service consumer and service producer. Its function is to facilitate the communication between management functions within and across management domains.

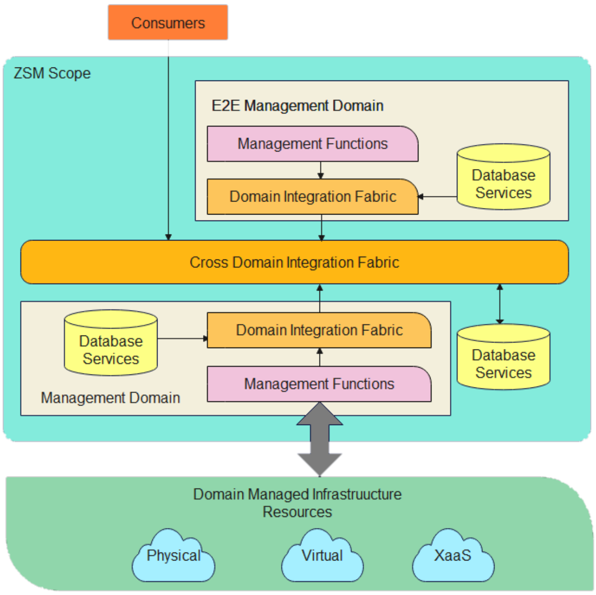

Figure 35. ZSM architecture high-level overview reference.

Concerning the consumers, vertical industries pose specific network and service requirements that need to be met. In the automotive case, the main challenges for urban traffic management are to ensure the smooth and safe flow of traffic, especially with the growing number of vehicles and new modes of transportation such as self-driving and micro-mobility vehicles. To achieve this, B5G technology can provide faster and more reliable data transmission at higher rates, which is essential for real-time traffic surveillance for safety functions. The evolution of the AI/ML techniques has reached a mature point to provide solutions for tackling the limitations of traditional MANO operations of managing complex and heterogeneous networks, thanks to their capacity for dealing with optimization and decision-making processes.

The increasing demand for new business-oriented services in the B5G towards 6G networks is growing along with the constant evolution of vertical applications [33]. The requirements of these applications are scaled up by the need for more agile use and configuration of programmable network functions using Software Defined Networking (SDN) and NFV, and creating Network Slicing (NS). In the case of automotive scenarios, the vehicular systems are currently counting on transportation communication standards like Intelligent Transportation System (ITS-G5), and Vehicle-to-Everything (V2X), and the possibility to enhance the services by getting closer to the users from the cloud and the edge computing.

To test the performance of decision-making processes that allocate network resources for vehicular services based on Roadside Unit (RSU) workload, IMEC runs series of experiments on the Smart Highway, thanks to the possibilities it brings for deploying a PoC within realistic conditions. Such an implementation is illustrated in Figure 36. The resulting datasets are valuable for the training of the DRL that would support the decision-making processes of the zero-touch services.

For monitoring network and vehicular services, execution of stress tests, orchestration and data collection are run as services in a containerized environment supported by Kubernetes [17], as already illustrated in Section 2.5, Figure 32. The logical setup for the experiment with the ZSM framework and vertical services is organized as follows and illustrated in Figure 37:

- RSUs 4, 6 and 7 replicate multi-site hosting deployment while publishing status stats for monitoring purposes.
- RSU 5 contains algorithms that mimic the demands from users by performing requests to specific RSUs selected randomly in a specific time-lapse.

- RSU 7 also contains algorithms for orchestration, decision-making and data collection about performance and available resources, e.g.: Central Processing Unit (CPU), memory and latency.
- OBU 2 that works as a mobile unit that measures the response time from the RSUs deployed in the Smart Highway and collects data for latency records.

For communication between the running applications across the RSUs and the OBU, Zenoh [34] is used as a publish/subscribe technology that provides high compatibility within distributed systems, which makes it ideal for a typically diverse environment at the edge. It is also noticeable how Zenoh provides a low burden of extra overhead information in the messages.

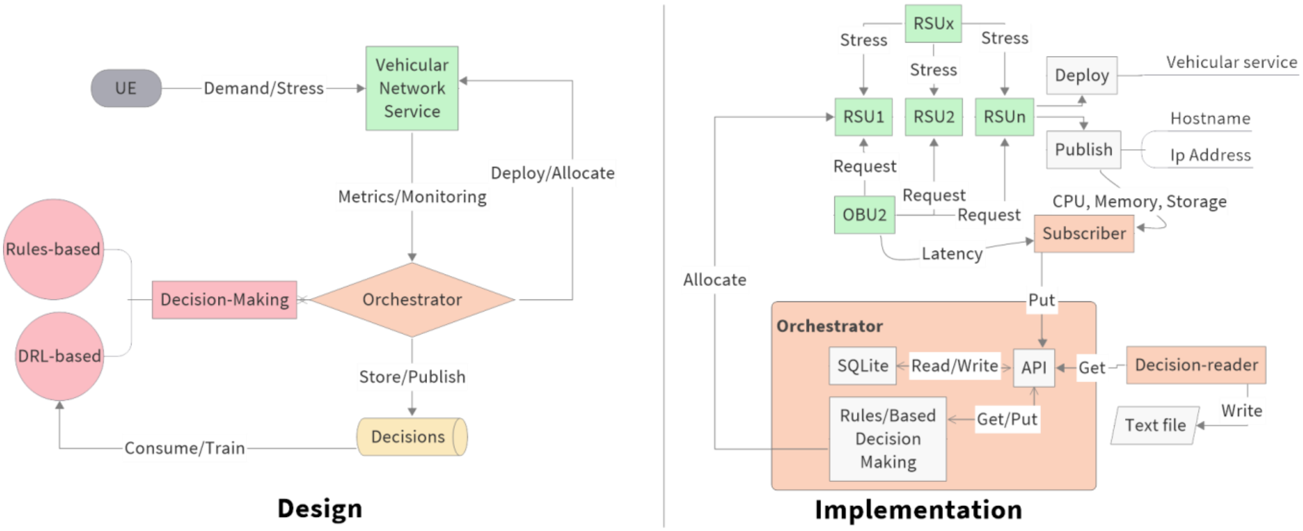

Figure 36. An example of ZSM implementation.

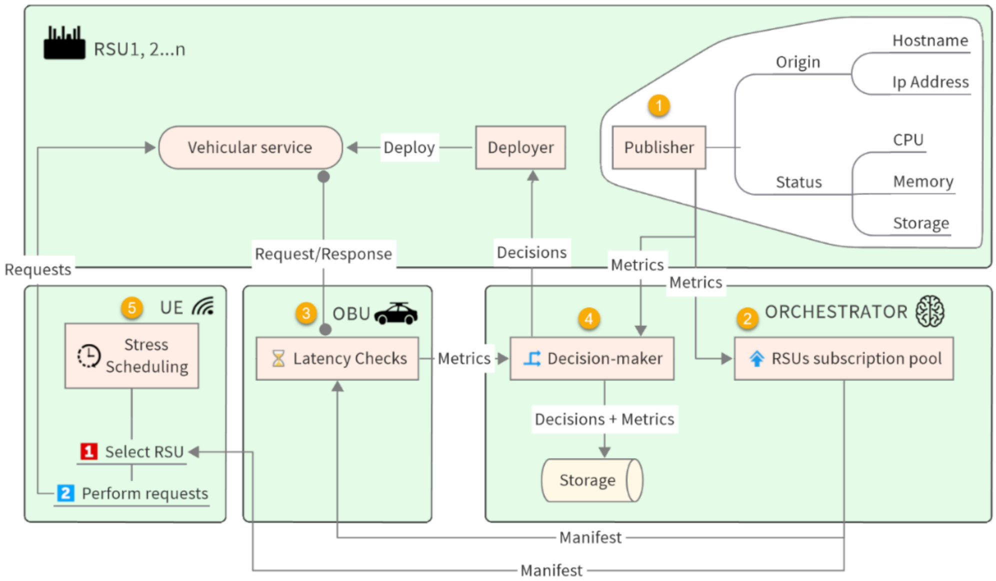

Figure 37. Logic flow of PoC at the Smart Highway testbed.

3.1.1.1 Experiment setup in realistic environment at the Smart Highway testbed

The main goal of the testbed setup is to create an environment to deploy PoC for testing and validation of zero-touch services within the realistic environment conditions offered by the Smart Highway. These services are AI-based functions and/or algorithms that support the orchestration service by improving the decision-making process. The goal of this PoC is to run algorithms that make decisions based on available resources.

To take advantage of the discoverable features enabled by Zenoh, the subscription of the RSUs to the orchestration is made automatically as each dataset of metrics that is published, includes information about the origin entity like hostname and IP address. The origin data is sent and managed by the subscription pool which make it available to the rest of the applications. This way, no manual configuration is needed prior to the execution of the PoC, providing flexibility to the setup and real-time feedback from the running services.

The execution of the PoC proceeds as follows (and as displayed in Figure 32):

- The publisher module of the hosting RSUs sends status values about performance status such as CPU load and Memory use. The dataset also hosts identification properties.
- The RSUs subscription pool includes the newly arrived RSU identity and shares it with the rest of the applications involved in the PoC.
- The OBU performs E2E latency checks among the RSUs.
- The decision-making algorithm processes the datasets sent by the publisher module from the RSUs and the OBU, making choices on the most suitable RSU based on the availability of its resources. The resulting decisions are stored and used to train a Deep Reinforcement Learning (DRL)-based decision-making algorithm.
- The stress-inducing scheduler application sends requests to a random RSU that is periodically selected from the manifest shared by the subscription tool. This task is performed to emulate the behavior of the dynamic changes in demand of vehicular network resources, which trigger actions in the orchestration operations for balancing and adjusting the running services.

The experimentation consists of deploying virtual machines on RSUs 4, 5, 6 and 7 and OBU 2 of a mobile unit. Within each RSU, there is a deployment of Kubernetes-based nodes that provide vehicular services to be consumed through the network such as assisted navigation, speed optimization, obstacle detection, collision avoidance, and infotainment. To connect the nodes, Zenoh is used, which enables dispensing with hypostatic addresses and lightweight transfer of information with a low burden of data overhead, to monitor the state of resource consumption on RSU 4-6 and 7 taking into account the load level of the processors and the E2E latency, which is vehicular communication services such as notifications sent to the vehicles and vice versa. To generate behavior in the RSU that enables data collection, an additional algorithm was created and installed in the RSU 5. This algorithm is responsible for emulating the demand for RSU resources. To do this, it randomly selects an RSU from 4, 6 or 7 every 180 seconds, which is the optimal interval that has proven to be adequate to the setup to make enough impact for the experiment in the available resources of the RSUs.

There is a correlation between the increase of CPU load and E2E latency. The relationship between them is learned by using the Support Vector Regression (SVR) method. As concurring requests are made to push the availability of the RSUs capacities to their limits, the generated stress increases CPU consumption and eventually affects the E2E latency of vehicular services depending on how busy the RSU computing units are while processing the requests. Under these conditions, the stressed RSU behaves as if it is under a high level of demand.

The decision-making algorithm reviews the data obtained by the rest of the RSUs and selects the RSU that indicates the lowest latency and CPU consumption. The main decision algorithm is rules-based designed while the supporting decision algorithm is DRL based and uses the outputs from the rules-based implementation to be trained. The overall decisions are expected to be reinforced by using DRL, whose performance is later compared along with the rules-based algorithm in the next section of this paper.

3.1.1.2 DRL support for decision-making processes

The available distributed edge computing in the testbed, in the form of accessible RSUs, creates a complex and dynamic environment. In such environments, DRL has emerged as a suitable option for decision-making processes in recent years, as it has been shown to effectively learn optimal strategies in complex network settings while adapting to changes [35]. Additionally, DRL offers a solution capable of making near real time decisions regarding which RSU should be employed for the computational task of the user. In this experiment, the Deep Q-Network (DQN) algorithm is adapted (see Figure 38), a well-known DRL approach, in order to address the problem of RSU selection [36]. This is relevant for the users as it can affect the availability and performance of the vehicular services when saturation is caused due to the demand not being effectively managed, and that implies an efficient RSU selection.

Parameter	Value	Parameter	Value
n	3	k	8
episode length	5295	episodes	10
batch size	32	activation function	ReLU
γ (reward discount)	0.9	dense layers	2
dense layer size	32	memory capacity	2000

Figure 38. DQN Hyperparameters.

Figure 39 illustrates how the DQN algorithm learns from the data collected from the smart highway testbed on December 13, 2023. The collected data is split into training and test data and then used different percentages of the training data to test the DQN performance. When only a fraction of the training dataset is used, about five percent, the performance of the DQN is only slightly better than that of a random selection approach, i.e., selecting RSUs at random. However, the performance improvement is fast and results in better performance than that of the conventional decision algorithm. This indicates that the DRL can quickly adapt to the dynamic environment of a testbed, as it only requires a relatively small number of data samples to outperform conventional rule-based approaches.

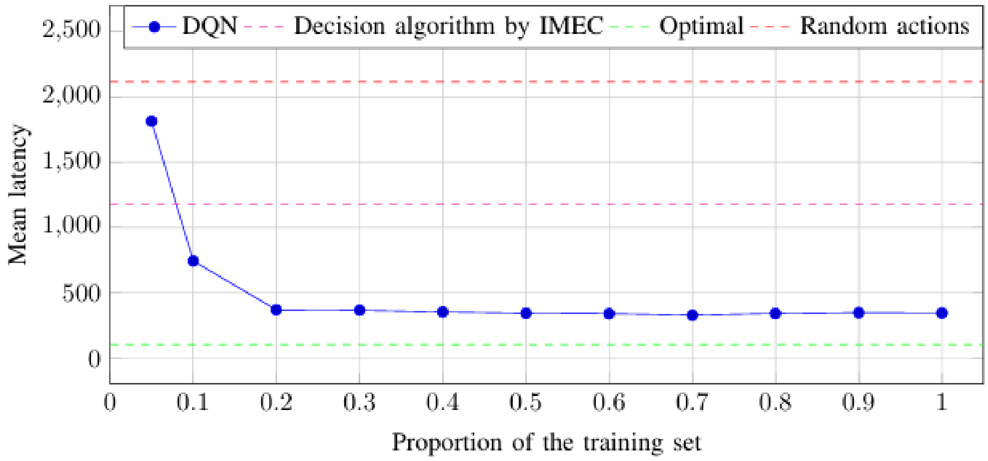

Figure 39. Change in mean latency over the proportion of used measurements gathered from the Smart Highway testbed.

The performance of DRL using the datasets from the decision-making process during the orchestration of the vehicular services, indicated promising results based on the rapid adaptability of this algorithm in the dynamic environment of the Smart Highway testbed. This characteristic made DRL outperform the conventional rule-based approach for decision making, presenting it as a feasible candidate for the support and improvement of the orchestration of vehicular services.

3.1.1.3 Management of continuum resources for E2E service orchestration

The programmability of computing resources throughout the Extreme-Edge, Edge, and Cloud Continuum, along with the increasing prevalence of Extreme-Edge devices, has become critical in the development of 6G. These assets can be used to efficiently deploy and distribute applications by implementing strategies such as Continuum resource discovery, allocation, and migration. Such strategies are based on the constraints and characteristics of Continuum nodes and are augmented by prediction algorithms that anticipate dynamic constraints over time (e.g., battery levels of Extreme-Edge devices, energy consumption, connectivity quality, computing loads from concurrent applications).

The NXW Continuum MT-M&O Platform software component plays a central role in enhancing management and orchestration functions across the Extreme-Edge, Edge, and Cloud Continuum. Designed to be agnostic and adaptable, this platform facilitates tasks such as accommodating diverse virtualization platforms and Extreme-Edge devices, optimizing orchestration mechanisms for proximity-sensitive applications, and automating the discovery of virtualization platforms and Extreme-Edge devices.

The architectural design of the Continuum MT-M&O Platform has been preceded by a requirements analysis to determine how to model the resources offered by different types of nodes in the Extreme-Edge, Edge and Cloud Continuum in order to build the required information models needed by the platform to carry on its discovery, orchestration and migration operations. The Extreme-Edge, Edge and Cloud resources have been modelled into the Continuum-MT-M&O Platform generic and specific information models so they can (I) support multiple and heterogeneous virtualization platforms and their computing capabilities (e.g., Kubernetes, K3s, Microk8s, OpenStack, etc.), (II) seamlessly support various Continuum segments (e.g., Extreme-Edge, Edge or Cloud), (III) support different types of Extreme-Edge devices (e.g., IoT devices, Sensors and Actuators, Robots and Cobots, etc.) including also the usage of ontologies to better characterize specific Extreme-Edge device categories (e.g., Dronetology for Drones, OCRA for Cobots).

To ensure platform and device agnosticism, a Driver-based approach is employed. This approach enables the orchestrator to access platform-specific computing capabilities and device-specific information through dedicated drivers, managed by an agnostic manager. This manager oversees dynamic discovery, continuous monitoring, and inventorying of resources across the Extreme-Edge, Edge, and Cloud Continuum.

Similar requirements analyses are conducted to model platform-agnostic virtualized service deployment requests. These requests, handled by the Continuum MT-M&O Platform, include high-level application requirements, reference specific orchestration templates, and specify target virtualization platforms. The Driver-based approach is again utilized to support multiple virtualization platforms and orchestration mechanisms, while presenting a unified interface.

Figure 40 depicts the architecture of the Continuum MT-M&O Platform highlighting the Extreme-Edge, Edge and Cloud resources discovery, monitoring and inventory workflows as well as the virtualized services orchestration workflow.

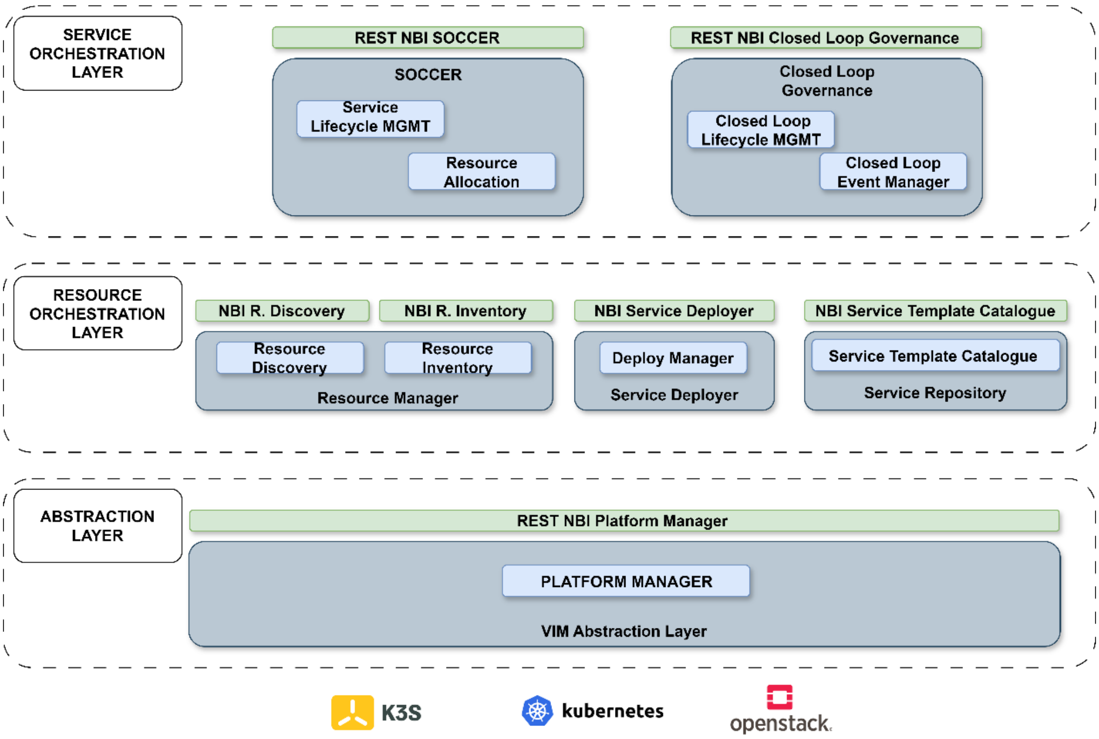

Figure 40. High Level Architecture of the Continuum-MT-M&O Platform.

The architecture of the Continuum MT-M&O Platform is based on microservices, with each submodule contributing to orchestration functionalities. The Platform Manager facilitates dynamic resource discovery and monitoring, while the Resource Manager enables a unified interface for accessing Continuum resources. The Service Deployer manages platform-agnostic virtualized service orchestration requests, translating them into platform-specific operations.

The introduction of the Continuum MT-M&O Platform aims to surpass existing architectural standards, such as NFV-MANO and ETSI-MEC, by embracing cloud-native approaches. This shift addresses the distinct requirements of vertical applications, decoupling network service orchestration from application layer orchestration.

3.1.1.4 Planned deployment and integration

The NXW Continuum MT-M&O Platform has been deployed in the Italian (Turin) infrastructure for supporting the deployment and metrics collection for UC5, UC12 and UC13, as well as providing the management and orchestration mechanisms to the computing resources and the applications running on top of it. There are ongoing discussions for evaluating the implementation of closed-loop functions for the automatic migration of services across devices in the Continuum. Moreover, a subset of the MT-M&O functionalities (mainly the ones related to the resource management) will be deployed in the Romanian infrastructure for supporting the management of the network and computing resources, providing the mechanisms for the implementation of the actions taken by the Decision Making module at the Smart Highway testbed.

In addition, Figure 41 and Figure 42 illustrate the planned deployment and integration of the zero-touch service management components in Iasi network facilities, which will be tested and validated in the scope of UC4. Both figures combine ZSM framework with the network setup in Iasi. In such a setup, UC4 services and data collection probes (needed for real-time data retrieval and decision-making at the ZSM level) will be deployed on the Telco Edge, while ZSM framework with Zero-Touch services will be running on the 5G Core, as it requires more computing power to produce enhanced service orchestration decisions. The real-time metrics that are currently being collected from the RSUs on the Smart Highway will be associated with the Telco Edge computing units in Iasi, using the same publish/subscribe mechanisms, while those measured from the test vehicle will evaluate the end-to-end performance measured at the client side (UEs in Iasi).

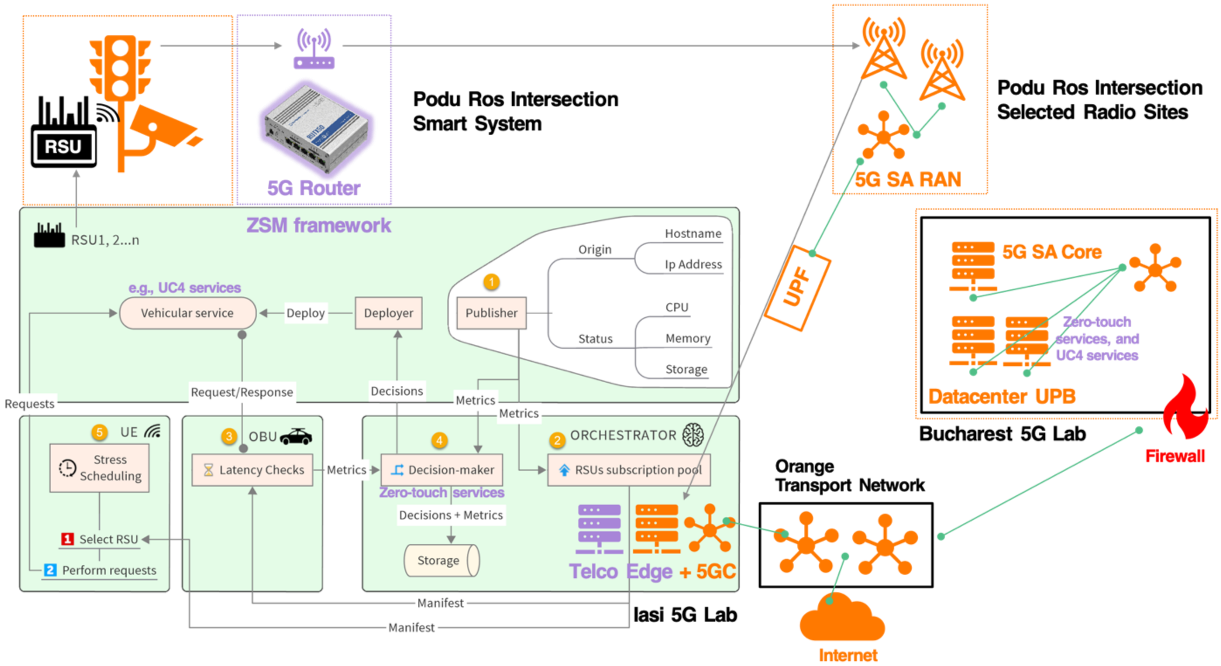

Figure 41. Planned integration of ZSM framework in Iasi network facilities.

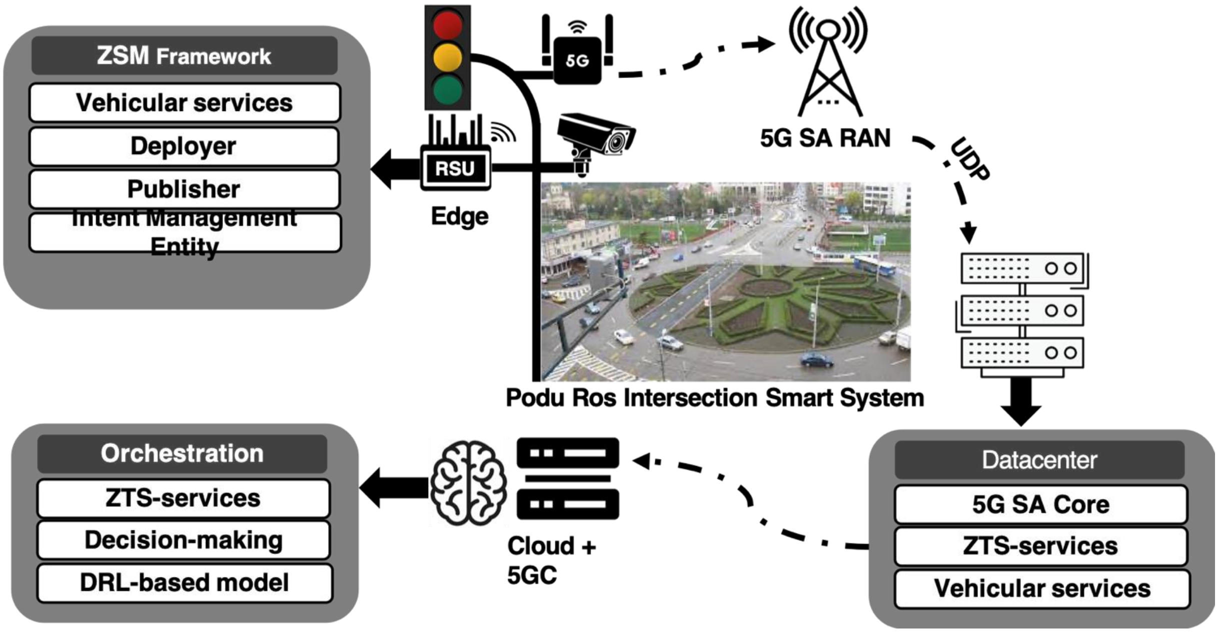

Figure 42. Planned deployment of ZSM in the scope of UC4 trials.

3.1.2 B5G applications framework

The technological advancements in the scope of 5G/6G will continue to enable a plethora of new and resource-intensive applications, by connecting over 125 billion devices worldwide by the year 2030 [37]. Given the ever-growing popularity of edge computing, due to its capabilities to offer resources at the network edge, i.e., closer to the end users, thereby preserving privacy and security unlike the centralized cloud environments, Edge Applications are gaining momentum in B5G/6G context. To enable proper management and orchestration of any vertical application, it is important that these applications are designed and developed in line with the standardized frameworks proposed by standardization bodies, such as 3GPP or ETSI. Thus, to seize the benefits from advanced ZSM functionalities presented in Section 3.1.1, which are aimed at improving vertical service performance, services and applications need to be modular and programmable, as described in this section. Therefore, in the context of the TrialsNet project, this EdgeApp framework will be mainly utilized for vertical applications developed in UC4, as these will be tested in the context of ZSM. Nevertheless, the B5G applications framework is generic and can be applied to any vertical application, especially for the purpose of creating sustainable vertical applications, as described further in Section 4.2.3.

Edge Applications (EdgeApps) or Edge Native Applications are a natural evolution of Cloud native applications, being built from virtualized network and computing resources, and usually consisting of multiple microservices that are dynamically interacting via programmable interfaces to perform specific tasks (cloud-native architectural design). According to ETSI [38], the inherent difference between EdgeApps and cloud-native applications are the edge environment characteristics that EdgeApps need to consider to best utilize the capabilities of any edge environment. To deploy an EdgeApp, which is either to linked to a specific vertical/use case or acting as vertical-agnostic, it is important to identify key requirements for network and service performance that can be translated into functional specifications, i.e., a set of instructions that will be applied by service orchestrators and controllers. This is in line with the intent-based network and service management reported in Section 3.1.1, where the identified key requirements are indicated in the form of an intent, which is further processed by intent interpreters and as such further communicated with zero-touch managers to make final decisions on necessary service life-cycle management operations that need to take place to ensure the required levels of performance. One of such initiatives is also presented by Linux Foundation open-source project called Edge Multi-Cluster Orchestrator (EMCO) [38], which delivers required service quality based on the high-level declarative intents (onboarding applications via Kubernetes Helm charts).

Given the resource constraints in edge networks, the design of edge applications is highly important because of resource consumption [39]. If an EdgeApp is designed to perform all separate processes or groups of processes, in separate containers, the orchestration entities can scale containers independently, and potentially save more computing resources than is the case of scaling all processes inside one container at once. Such a design, which decouples the main application logic into several independent and loosely coupled microservices, allows ZSM orchestrators to rapidly and flexibly deploy services and make sure that application performance matches the required level of quality of service (specified in the intents). Therefore, ZSM systems that manage and orchestrate edge resources and services need to dynamically create external interfaces for EdgeApp deployments and make them accessible for the users (e.g., vehicles, robots, vulnerable road users), dissemination services, orchestrators, and peering application instances in other edge platforms. This interface towards orchestration entities can be further used for informing orchestration layers about some internal application procedures (e.g., energy consumption increasing due to heavy AI/ML inference) so that the life-cycle management of applications can be improved by deploying additional instances in other relevant domains.

However, resource constraints are not the only characteristic of edge environments that need to be taken into account when designing vertical applications. In their recent whitepaper [38], ETSI summarizes the list of characteristics of EdgeApps, highlighting their impact on EdgeApp design:

- **Low latency**: With reference to vertical industries, numerous mission-critical and latency-sensitive applications require real-time response and fast data processing. EdgeApps can be optimally placed at the closest or fast-performing edge computing nodes, as orchestration layer (defined by ZSM framework and orchestrators as described in Section 3.1.1) takes into account various metrics (proximity from user, real-time network and service performance, edge computing load) to make well-informed decisions that comply with key requirements identified and specified in the intents.
- **Enhanced security**: EdgeApps keep data processing locally on the edge computing units, therefore, reducing the amount of potentially sensitive information to be sent to the centralized cloud environments across the network. Nevertheless, it is still important to properly design the overarching ZSM framework to also take into account security and privacy aspects, ensuring better resource isolation, data encryption and efficient authentication methods.
- **Resource constraints**: As described above and in [39], edge computing environments can suffer from lack of resources to perform compute-intensive tasks, unlike the traditional centralized cloud environments that are not constrained by computing and storage resources. This is particularly relevant for applications that are expected to run some heavy data processing, AI/ML operations, which could be heavily affected. Therefore, the intents mentioned earlier need to include the computing requirements of applications that need to be deployed or relocated to ensure the proper translation of requests to corresponding orchestration/ZSM operations (e.g., relocating application to another edge, or moving it to the cloud for retraining purposes).
- **Mobility/portability**: In case of mobility scenarios, such as those involving smart vehicles and other traffic participants (e.g., vulnerable road users), it is important that optimal application placements are selected for serving end users. This often means that applications need to be relocated/migrated from one edge computing unit to another to maintain required levels of Quality of Service (QoS). Usually, mobility/portability of EdgeApps also means that some metadata needs to be transferred from one application instance to another in case of service relocations. This scenario needs to be supported by application framework, therefore, enabling application instances to dynamically established east/west-bound interfaces with peering application instances in various domains to timely transfer required data that is necessary for service continuation in the target domain. This approach will at the same time enable resiliency of EdgeApp deployments but needs to be carefully considered as all overhead communication between application instances could increase computational load, i.e., carbon footprint and lead to energy-inefficient operation.
- **Network Awareness**: Interaction between EdgeApps and network is gaining more momentum with the Open Gateway initiative from Global System for Mobile Communications Association (GSMA) [40]. This initiative allows application developers to leverage the open Application Programming Interfaces (APIs) for exposing network capabilities (QoS on demand, location) that could enhance the operations of applications. More specifically, ETSI enables Radio Network Information Service (RNIS) to retrieve complex mobile network parameters. However, as such complexity might not be comprehended by simple application logic, they either require a ‘translation layer’, i.e., middleware that will understand the parameters and impose certain actions to improve application logic. Therefore, it is important to distinguish which type of information to allow applications to consume in order to not burden them with

unnecessary requirements for understanding complex network operations, which could be rather handled by the ZSM layer.

Therefore, all of these characteristics highlight the importance of the tight interaction between ZSM layer and application instances, as illustrated in Figure 43, to make ZSM layer and applications mutually aware in order to enhance their respective operations.

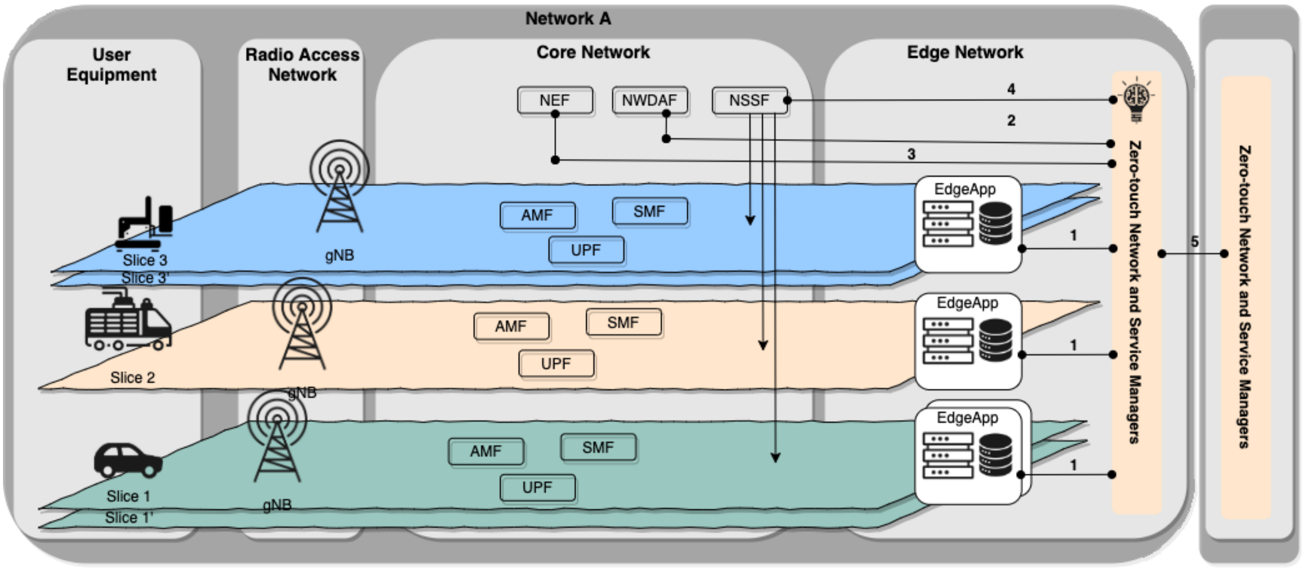

Figure 43. EdgeApps in the End-to-End 5G/6G ecosystem.

While ETSI is focused on the Multi-Access Edge Computing (MEC) standardization, including aspects of native edge applications, 3GPP is actively advancing towards incorporating those applications into the 3GPP, i.e., 5G/6G context. In particular, 3GPP is standardizing application enablement initiatives through creating service frameworks, which should be used as a baseline for designing and developing EdgeApps in the context of 6G deployments.

The following three service frameworks have been defined by 3GPP SA6 group [41]:

- **Common API Framework (CAPIF)**: a unified Northbound API framework across network/application functions to facilitate a harmonized approach for API development within 3GPP. This ensures that there is a single entry point for vertical applications (a.k.a. API invokers) towards the common API aspects (also called CAPIF APIs) such as onboarding, discovery, authentication and authorization (Refer to 3GPP TS 23.222).
- **Service Enabler Architecture Layer (SEAL)**: specifies application plane and signalling plane entities for application-enabling services (e.g. group management, configuration management, location management, identity/key management, network resource management); that can be reused across vertical applications (Refer to 3GPP TS 23.434).
- **Edge Application Enablement**: provides an edge enabling layer and application architecture for enabling Edge Applications on the Edge Data Network, including the exposure of northbound APIs towards Edge Applications, integration with the 3GPP Network, and to facilitate communication between the Application Clients running on the UE and the Edge Application Servers deployed on the Edge Data Network with capabilities such as service provisioning, rich application discovery, and service continuity (Refer to 3GPP TS 23.558).

The SEAL framework was already introduced and described in D2.1 [1], therefore, this section provides more insights into the other two frameworks in order to create a solid knowledge base for creating an overarching application framework.

A crisp overview of service-based interactions based on the CAPIF framework are presented in Figure 44 and Figure 45. In order to attempt to standardize some of the common capabilities that are exposed by the Northbound APIs, the 3GPP have introduced CAPIF (Common API Framework for 3GPP northbound APIs). Within the standardization of CAPIF, the 3GPP addressed a variety of different processes, including onboarding/offboarding of Application Functions, service discovery and management, event subscription and notification, security and charging.

As one of the efficient ways to support service continuity, application context relocation is being studied in 3GPP standardization track. In the framework of Release 17, 3GPP standardized the architecture for enabling edge applications (Edge Application Enablement), which creates mutual awareness between edge client applications, and edge application servers running in the edge data network in 5G system. This 3GPP standardization track created: i) the application layer architecture, shown in Figure 46, ii) procedures, and iii) information flows that are necessary for enabling edge applications in 3GPP networks. Such an architecture enables:

- Application portability, which makes it specifically important for distributed edge deployments in vertical scenarios where application-context needs to be frequently transferred from one edge-host to another to follow the movement of the users consuming the service,
- Service differentiation,
- Flexible deployment, and
- Inter-working with 3GPP networks, where different capabilities such as location services, QoS management, etc., are provided to edge application servers to perform their work more efficiently, e.g., by subscribing to events such as user plane path management events, i.e., user plane path change, and application context relocation

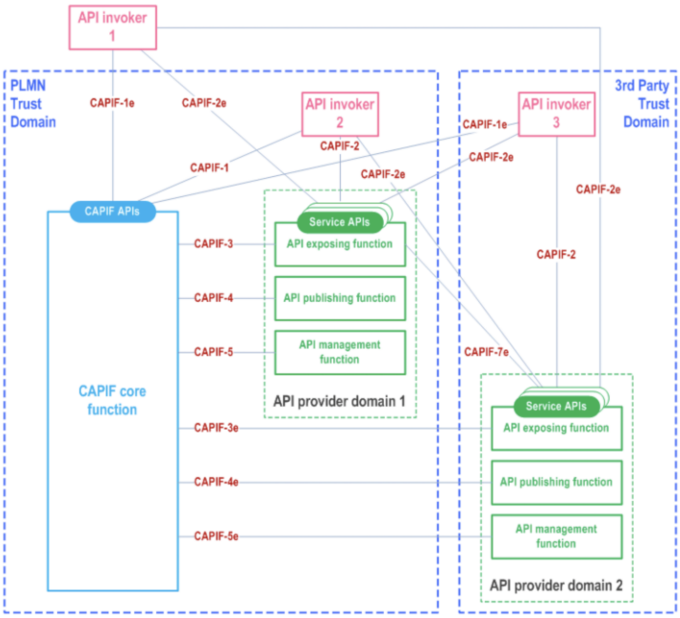

Figure 44. Functional model of CAPIF to support 3rd party API providers.

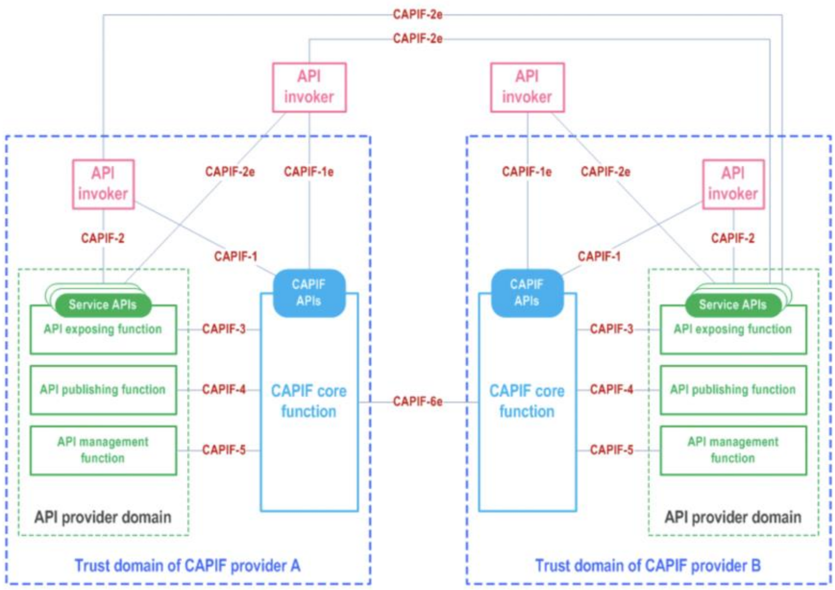

Figure 45. High level functional architecture for CAPIF interconnection with multiple CAPIF provider domains [42].

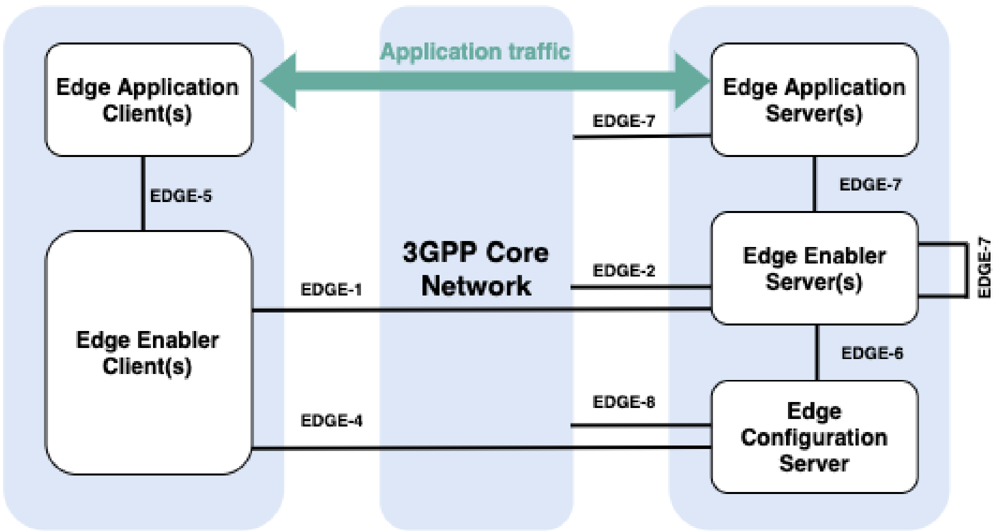

Figure 46. 3GPP Architecture for enabling edge applications.

In particular, the architecture shown in Figure 46 consists of the edge network (right), 3GPP Core Network (middle), and client side (left). The edge data network is built by:

- **Edge Configuration Server (ECS)**: it provides configuration data, i.e., Local Area Data Network (LADN) URI, to the Edge Enabler Client (EEC) to connect to the Edge Enabler Server,
- **Edge Enabler Server (EES)**: it interacts with 3GPP core to collect network and service capabilities (e.g., location services, QoS management, etc.) that will improve the performance of edge application server, thereby enabling Edge Application Client (EAC) to connect to the server, and
- **Edge Application Server (EAS)**: it performs server functions and exchanges application data traffic with the client. Furthermore, on the client side, EEC discovers the edge network, retrieves the necessary information for connecting to the edge (e.g., coverage area/service area, types of application servers or MEC applications, etc.), and connects to it via IP address provided by EES.

As Figure 46 introduces only high-level interfaces between EdgeApps and 5G/6G Core, it is important to explore the capabilities offered by specific Core network functions, such as Network Data Analytics Function (NWDAF) that stands for Network Data Analytics Function (see Figure 47). In particular, NWDAF is a component of the 5GCN architecture that provides network data analytics capabilities to support network operations, optimization, and service assurance. NWDAF is in charge of providing real-time and historical data analysis of various network elements, such as network performance, resource utilization, and service quality. The data collected by NWDAF is used to support the optimization of network resources, ensure the proper functioning of network elements, and detect and resolve potential issues before they impact the user experience. In this way, NWDAF plays a crucial role in enabling 5G/6G networks to deliver high-quality services and meet the requirements of various vertical industries.

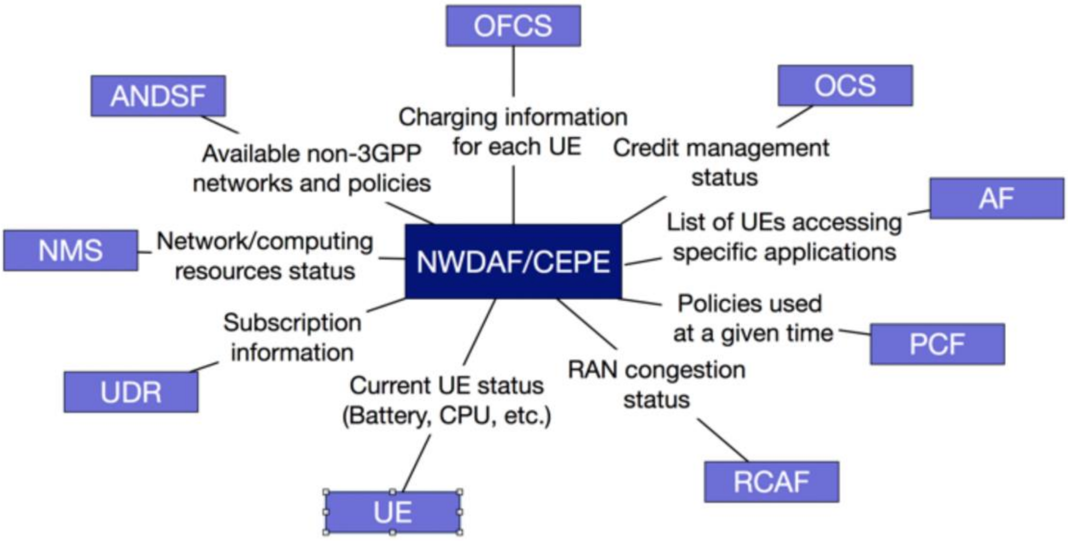

Figure 47. Provision of data to NWDAF [43].

Concerning current deployments of NWDAF in 5G Standalone networks, it depends on several technical factors, such as the specific requirements of the network operator, the maturity of the technology, the availability of suitable network equipment, but also non-technical ones, such as regulatory and economic factors. Depending on the country and operator, NWDAF may already be deployed in 5G SA networks, while in others it might be still in the planning or development stage.

3.1.2.1 Planned deployment and integration

In alignment with the development and integration of the ZSM framework outlined in Section 3.1.1, the vertical applications will follow the modular and programmable design presented in this section. In particular, the applications associated with UC4 will be embedded into containers with network-aware and energy-aware (see Section 4.2.3) features. Given the early stages of the deployment of NWDAF, there is still a vast room for exploring opportunities that it might bring to B5G networks, and as such, this specific activity in TrialsNet will further study how NWDAF can be loosely coupled with applications deployed on the e.g., network edge, and make them network-aware. The network-awareness is being particularly studied and enabled in the context of Open Gateway initiative, as well as the actual implementation in another Linux Foundation project called CAMARA. The goal of CAMARA is to create “a federated platform solution for exposing operator network capabilities to external application”. These capabilities will be further explored to ensure proper operation of services and as such will be tested and deployed specifically in the Romanian cluster, both at IMEC facilities and ORO 5G testbed.

3.1.3 Digital Twins applied to next generation mobile network

Motivated by the scenarios discussed in D2.1 [1], (Section 4.1.3), this section presents a case study of Digital Twin (DT) application to Network Appliances, with a specific focus on virtualized Radio Access Networks (vRAN) environments. This innovation will not be showcased directly in one of the UCs but it is propaedeutic for the faster development of UCs involving the radio access, as algorithm that require a finer interaction with the underlying radio environment that use ML as a basis. This is because the interaction with AI/ML algorithms can be time-consuming and resource-expensive in such systems, especially due to the difficulty of common software models in exposing data at a very high pace. Thus, learning algorithm may require lengthy iteration before converging to an optimal state. This case study particularly focuses on the lower layers of the protocol stack.

3.1.3.1 Scheduling task in Virtualized RANs

Processing is identified as the most computing-expensive operation in a mobile network stack [44]. In such systems, the amount of computing resources plays a significant role in the overall performance. Shortages in computing resources can lead to detrimental effects on the perceived throughput of users.

This aspect becomes critical in scenarios with synchronization constraints, such as in the case of Hybrid Automatic Repeat reQuest (HARQ) processing [45]. HARQ imposes stringent deadlines on the decoding of uplink (UL) wireless frames [46]. This task is not only computationally expensive but also lacks deterministic execution time. It is influenced by factors including (i) the allocation of the Physical Resource Blocks (PRB), (ii) the selection of a Modulation and Coding Scheme (MCS), and (iii) the user's perceived signal-to-noise ratio (SNR). The task's complexity also depends on the available computing capacity at the base station.

In vRAN systems, where resources are pooled among different processes [46], RAN procedures like Medium Access Control (MAC) scheduling need to be computationally-aware [47] to prevent disruptions caused by computing capacity shortages. Allowing users to send more data increases computation times, potentially leading to decoding deadline violations. Increased data transmission can be achieved by using more complex modulation or transmitting over wider bands. Although the deadline for decoding is configurable in 5G systems, with the default for 5G user traffic being 3 ms, maintaining reliable decoding times may result in a decrease in overall throughput. The impact of computing capacity on scheduling decisions, due to its non-deterministic nature, is challenging to predict, highlighting the need for data-driven approaches, including AI/ML methodologies, which are examined in the following discussion.

3.1.3.2 Digital Twin for Virtualized RANs

The design of an AI/ML scheduling algorithm involves electing, training, and deploying Deep Learning (DL) models in a production environment (within the pipeline, workflow, etc.). This can be challenging due to the complexity of real systems. In the presented case study, grant assignment decisions are made at the MAC layer, with CPU resource input coming from the MANO, and decoder performance recorded at the Physical layer. A significant challenge is the high complexity involved in training. To develop an effective model, an AI/ML algorithm must explore all possible input combinations multiple times. This can be a lengthy process in a real system, as (i) real scheduling decisions are limited to once per every scheduling interval (1 ms by default), and (ii) reaching certain combinations of inputs may be difficult due to channel conditions. Performing this process on the fly in a production system is not feasible, as systems like Open RAN necessitate the deployment of offline pre-trained models [48].

These factors make it impractical to learn directly from a real system, considering numerous aspects such as model architectures, hyperparameter configurations, analysis, and model interpretability. To address these challenges, a DT of a vRAN network appliance is introduced, aimed at enhancing the operation of computationally aware MAC schedulers.

3.1.3.3 Digital Twin design

In the described scenario, the physical system requiring control is the Forward Error Correction (FEC) decoder, as it provides feedback on both decoding time and the Cyclic Redundancy Check (CRC) result. Consequently, the DT of such a system must replicate the physical system's distribution of decoding times and the success and failure rates of decoding. To achieve this, a dataset with real-trace performance measurements is created, followed by training a supervised model to capture these measurements in the DT.

The process of deciding what aspects of the system to twin and how to approach this task involves querying a real decoder under various combinations of the input space. These inputs include CPU Capacity (in % of maximum CPU resources), the user's SNR (in dB), the MCS index, and the number of PRBs. The decoder's response is observed, capturing the decoding result with a binary variable, i.e., CRC = 1 if the frame is successfully decoded, or CRC = 0 otherwise, and the decoding time. The decoding time is modeled as a continuous normally distributed variable, truncated to positive values. The DT then outputs (i) the probability of successful frame decoding, and (ii) the mean and standard deviation of the normal distribution modeling the decoding time.

3.1.3.4 Dataset collection

Software

With 6G still under specification by standardization bodies and lacking an open-source reference architecture, the srsRAN [49], an open-source software implementing the 4G/5G functionality of the whole mobile networking stack of the evolved Node B (eNB)/gNodeB (gNB), was utilized. The version 22.04 of srsRAN, which implements the Rel-15 of the 3GPP standard, was employed.

Data

The decoding time of the UL frame was measured by recording the wall time of the decoding process. To mitigate noise from other processes, decoder threads were pinned to a specific CPU set, and the default Linux scheduler was prevented from preempting these threads in favor of other jobs. The CRC result was directly retrieved by the decoder.

Method

To thoroughly explore the entire input space, the following process was implemented. The srsRAN's UE and gNB processes were set up on different host machines. A bandwidth of 10 MHz was used (up to 45 PRB for data transmission in the UL) with a Transmission Time Interval (TTI) of 1 ms and one user per TTI. To span the PRB and MCS sets, the default srsRAN scheduler was replaced with a custom one that randomly selects the number of PRBs and the MCS index, directing the user to transmit a new frame. The UL frame, transmitted over an Additive White Gaussian Channel [50], was decoded by the gNB's decoder threads, which output the decoding time and CRC. To cover different SNR levels, the wireless channel was controlled using an automated process that selects the target SNR. This process also managed the CPU capacity of the decoding threads by adjusting the allocated CPU cycles. This procedure, representing the data collection phase, is illustrated in the left block of Figure 48. Approximately 14 million samples were collected, encompassing extensive combinations of the input parameters.

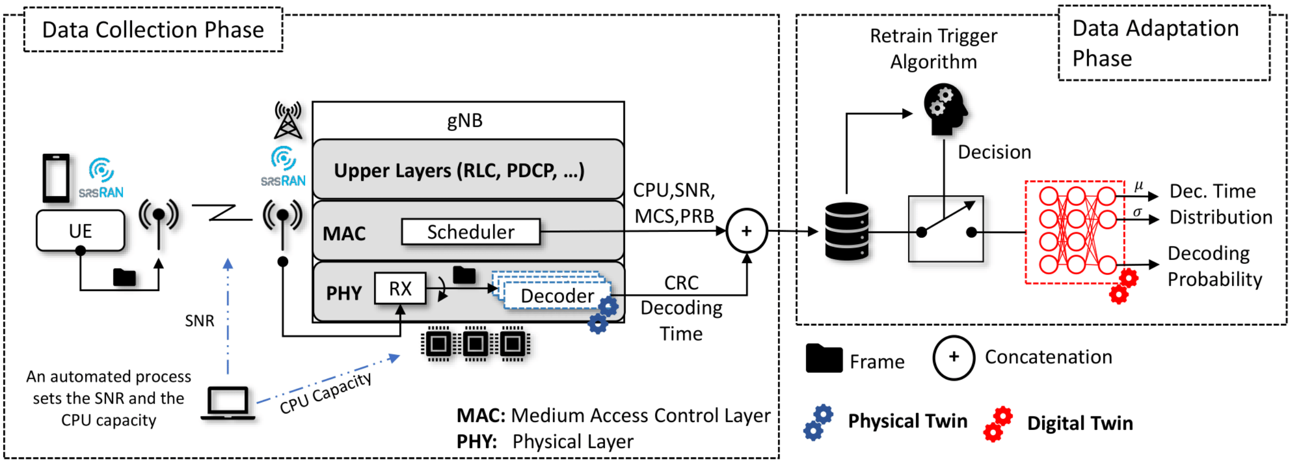

Figure 48. The building blocks of the proposed DT for vRAN systems.

3.1.3.5 DT performance

The evaluation of the DT performance is divided into the two tasks described in the following:

- **Decoding time prediction task:** The Kernel Density Estimation method is used to approximate the normalized Probability Density Function (PDF) of the real distribution. This real distribution is plotted in Figure 49, alongside the predicted distribution for conditions of 15 dB SNR, 40 PRBs, 100% CPU Capacity, and MCS index ∈ {2, 12, 20}. Observations from the real PDF indicate that lower MCS indices, which imply lower decoding complexity, result in smaller mean decoding times. Conversely, higher MCS indices are associated with greater variability in decoding times, attributed to the increased complexity of the task the decoder software implementation must solve, potentially leading to irregularities such as cache misses that unevenly reduce performance, thus increasing uncertainty. The modeling assumption that the decoding time can be captured with a normal distribution (truncated to positive values) is corroborated by performing the Kolmogorov-Smirnov test for various scenarios, where the null hypothesis was accepted with a 99% confidence interval for all tested combinations of selected MCS.

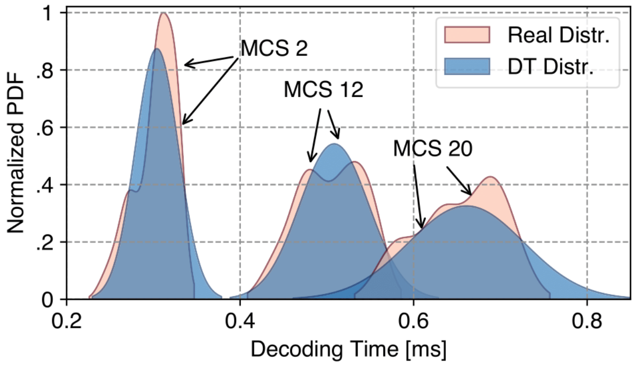

Figure 49. The PDF of the real and the predicted decoding time distributions, truncated to positive values.

- **Decoding success probability task:** The performance of the predictions for the decoding success probability task is then evaluated. It is observed that the decoding probability is high with a low MCS index and number of PRBs, as the data rate is lower. However, as either the MCS or the number of PRBs increases, leading to an increase in carried data, the Shannon capacity of the channel for this SNR level is reached, and the probability gradually drops to 0. In Figure 50, the BCE loss is plotted for 10 dB of SNR, various MCS indices, and numbers of PRBs. It is noted that the DT successfully learns the trend when the probability is either very high or low, while exhibiting a small prediction error in the transition region.

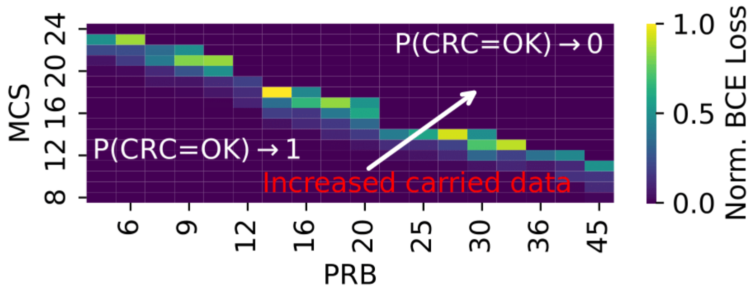

Figure 50. The real and DT's prediction of decoding probability, and the BCE loss of the prediction task.

3.1.3.6 Distribution shift

In this experiment, the interaction between the Physical Twin (PT) and the Digital Twin (DT) is studied, particularly focusing on the DT's ability to adapt to previously unseen inputs. Such situations may arise when the available historical data does not cover certain regions of the input space, for example, due to sudden failures in CPU capacity or new implementations or computing infrastructures not used in creating the DT.

To simulate the first scenario, the available CPU capacity of the PT is drawn in the interval of 90-100% of the maximum achievable, considering only very high CPU capacity (distribution A). The collected dataset is divided into a training set, for training the DT, and a validation set, for evaluating its performance across the training epochs. The normalized validation loss is plotted in Figure 51, where the DT was trained for 30 epochs.

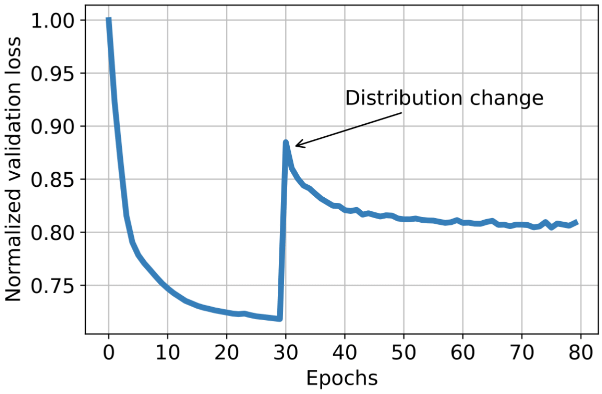

Figure 51. The DT validation loss (sum of BCE and NLL losses) when distribution changes happen.

Subsequently, the available CPU capacity is drawn in the interval of 10-90%, considering a wider range for the CPU capacity, yet lower than the previously used range. The retraining of the DT can be triggered at a set frequency or when the similarity between the DT predictions and recent observations falls below a certain threshold. In this case, a change from distribution occurred, leading to a drop in similarity and triggering a retraining of the DT. A spike in the validation loss is observed as previously unseen observations cause high prediction errors, which then decrease as the model adjusts.

3.1.3.7 Complexity and time of inference gain

The gain of using Neural Networks (NN) for predicting the decoding probability, as opposed to a pure simulation approach, is quantified by evaluating the complexity and inference time required to generate a new data sample. In a 5G testbed, generating a new data sample takes at least every 1 ms, considering a TTI of 1 ms. Moreover, this approach introduces additional engineering complexity in building a reliable end-to-end system for taking measurements. In contrast, with a NN-backed DT, the time required to generate a single sample is equivalent to the inference time of the forward pass of the NN.

When the DT was queried on a computing platform with 1 million combinations of input features, the average inference time was measured at 2.8 µs per sample. This represents a 350× speed increase compared to training that does not utilize a DT-based approach. The complexity of setting up this DT is primarily associated with initiating a new process and loading the weights of the NN, which incurs a negligible cost relative to the training of the algorithm.

3.1.3.8 Outcome of the study

This section and the preliminary work reported in Section 4.1.3 of D2.1 [1] discussed the advantages and benefits of using a DT in the context of next-generation mobile networks. Three possible application scenarios have been proposed, followed by detailed design and implementation for one of these scenarios, specifically a DT for a virtualized RAN system. The results demonstrate that the DT accurately reflects the characteristics of the physical object, successfully and scalably twinning it, and adapts effectively to changing contextual conditions.

3.2 Vertical Innovations

This section briefly tackles a set of specific vertical innovations in the TrialsNet project, which are advancing specific use cases. Same as in the case of horizontal innovations, this section is also organized in reference with D2.1 [1] and it provides respective updates on the analysis and deployment of these specific vertical innovations in the context of the TrialsNet use cases.

3.2.1 AI mechanisms for diagnostics and resources efficiency

As reported in the previous deliverable D2.1 [1], 6G will support the delivery of solutions for verticals. Solutions rely on various AI components and are distributed in the infrastructure. TrialsNet conducts work on and enhances a diagnostics framework to assure that the proper performance is delivered by different segments of the 5G/B5G infrastructure (often assuming that some segments are a "black box"). Moreover, the diagnostics framework will monitor the vertical components hosted in the infrastructure, with respect to their non-functional behaviour to ensure compliance with the anticipated behaviour. AI mechanisms will be comparing (infrastructure and vertical components) to propose potential improvement actions.

The overall aim is to support the delivery of solutions for verticals in a manner that will be characterized by sustainability, trustworthiness, performance and efficiency. In this direction, the WINGS testbed has progressed with the development of diagnostics tools which have been a critical aspect of ensuring the resources efficiency and optimal performance. Diagnostics and monitoring tools have evolved to address the challenges associated with managing and optimizing network resources. These tools are designed to continuously collect and analyze various parameters, such as throughput, latency, availability, signal strength, in real-time. By monitoring these parameters, network administrators can identify potential issues, diagnose problems, and implement proactive measures to maintain efficiency and prevent disruptions. AI/ML algorithms are utilised for predicting the behaviour and performance of monitored services and nodes across the underlying infrastructure, based on time series forecasting. In this case, the goal is to create a view of the future state of the infrastructure and deployed services. This future state is used to determine if there is a need for pre-emptive actions to prevent upcoming critical events. Moreover, AI and ML technologies enable advanced pattern recognition and predictive analytics, allowing the tools to identify deviations from normal network behavior that may indicate problematic issues. By leveraging AI and ML, monitoring tools can not only detect problems but also learn from historical data to make predictions about potential future issues. This proactive approach enhances resource efficiency by enabling preemptive actions, reducing downtime, and optimizing the allocation of network resources. Framed in this context, B5G/6G infrastructures are deployed to serve diverse verticals. These verticals will be requesting a set of services, which can generate diverse traffic profiles. For instance, a utility can generate traffic from sensors, from video streams, or even audio; these streams can vary in requirements, for instance in terms delay and reliability. Systems should be prepared to handle situations that exhibit a behavior, beyond the normal operations, agreed service provisioning, etc. This is important for preserving the services of all verticals served by a network segment. The reason for the unordinary behavior can be due to malevolent reasons or to something else, e.g., some device malfunctioning, extraordinary requirements from the vertical, underestimation (or overestimation) of resources, etc. This problem falls in the class of problems denoted as anomaly detection. Anomaly detection can also be detected in the future state of infrastructure and deployed services, in which case appropriate resource optimisation mechanisms may be proactively triggered to decide on actions to prevent the predicted anomaly.

One of the most widely investigated approaches in the problem of anomaly detection problems is clustering. Density-based spatial clustering of applications with noise (DBSCAN) is a classical approach that has been applied in several outlier detection problems. DBSCAN performs unsupervised learning and classifies the analysed data points into core points (normal behavior), border points (potentially anomalous behavior) and outliers (anomalous behavior). Moreover, a modification of the DBSCAN clustering algorithm is proposed, which takes into account the temporal evolution of the data. Given that new points are generated continuously, these new data points could affect the overall clustering result and might cause true outliers to blend with regular data points. This phenomenon poses a significant challenge in maintaining the integrity of the clustering process. The proposed solution involves removing older points from the dataset periodically and then rerunning the DBSCAN clustering algorithm.

The process of removing older data points and rerunning the clustering algorithm can be computationally expensive. Therefore, a method that balances the need for accurate clustering and computational resources is proposed. This method is guided by a metric that is called average anomaly score that serves as a measure of "how anomalous" the outliers are; a higher average anomaly score means that the outliers exhibit a very different behavior from the normal points, whereas a lower average anomaly score suggests that the outliers are beginning to blend with the normal points. The use of the average anomaly score serves two purposes in the proposed method. Firstly, it allows us to monitor the performance of the clustering algorithm. A decreasing average anomaly score can be a warning sign that the clustering algorithm is starting to struggle to separate normal points from outliers, which could suggest that it is time to remove old points and rerun the clustering algorithm. Secondly, the average anomaly score can guide the decision of when to rerun the clustering algorithm. Running the clustering algorithm is computationally expensive, so the running frequency must not be too high. On the other hand, very low running frequency is not a preferable option either, because the phenomenon of mixing outliers as inliers could start becoming more prominent. Therefore, the clustering algorithm should rerun whenever the average anomaly score decreases by a certain threshold. This ensures quick reactions to changes in the data, while also minimizing the computational cost.

The mechanisms described in this sub-section as part of the WINGS testbed mainly in the application level are used in the scope of the UCs associated with the Greek cluster, mostly focusing on UC2, UC3, UC6 and UC11. For instance, AI mechanisms for diagnostics and resources efficiency are used for efficient object detection in UC2 as well as efficient and automated robot control in UC3 as reported in D3.2 [5], anomaly detection in data of wearables in UC6 (see D4.2 [6]), and efficient human detection in UC11 described in D5.2 [7].

3.2.2 Automatic orchestration of network slices to ensure QoS in mobility

This connectivity infrastructure supporting the use cases of the Pisa site, as illustrated in Section 2.2, hinges on the seamless interaction of three primary domains: radio, transport, and cloud. These domains are coordinated under the shared framework of an End-to-End (E2E) Orchestrator.

As illustrated in Figure 52, the network supporting the use cases of the Pisa Cluster (UC7, UC8, and UC9) exploits the commercial Ericsson Orchestrator (EO) operating in the remote TIM site in Turin. The EO is a comprehensive suite of modules which can be combined in various ways to provide life cycle management of Virtual Network Functions, control the allocation and use of virtualized infrastructure resources. In TrialsNet additional functionalities will be provided by the ER Innovative Orchestrator, located in the Pisa site. This software module is introduced to enhance the connectivity with additional optimization features, particularly in relation to the awareness of the transport domain. In fact, the E2E QoS is determined by the combination of QoS in the radio layer (Radio Access Network/Core Network) and the QoS in the transport layer. This optimal combination can be achieved using techniques to create intelligent mapping between RAN/Core Network (CN) QoS and transport QoS, which considers the specific technology of the infrastructure.

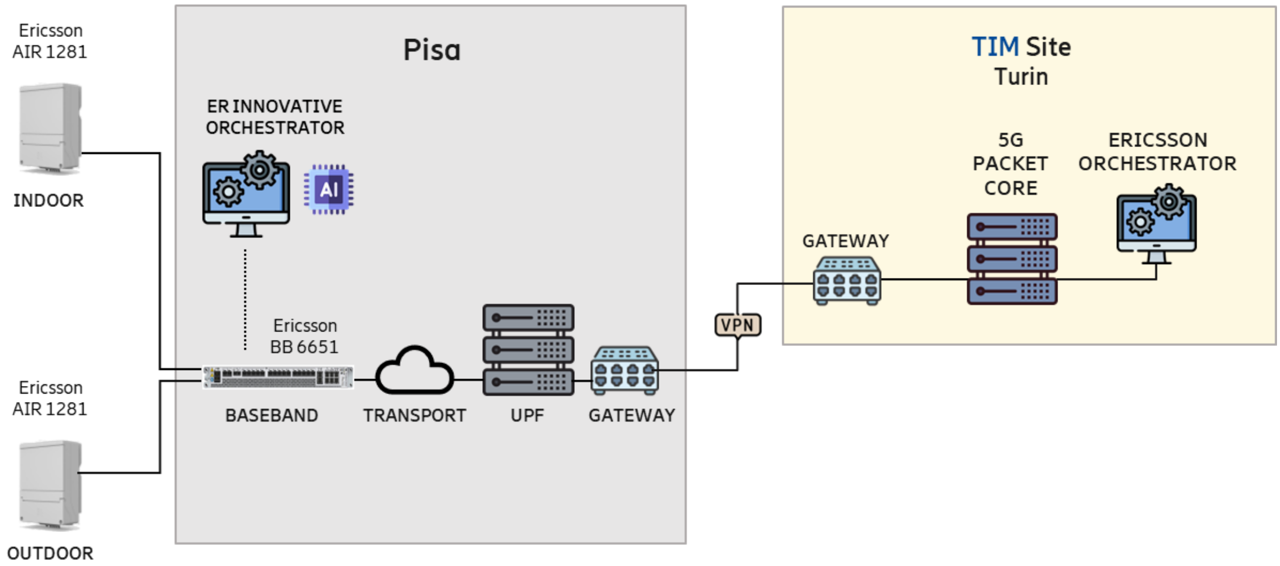

Figure 52. Orchestrators in the network scenario (Pisa site).

The ER Innovative Orchestrator utilizes transport awareness to ensure a broad spectrum of QoS levels, including very high performance needed to meet the most stringent requirements on a large-scale without the need of over-provisioning of transport resources. Specific techniques in the RAN/CN, like Dynamic Radio Resource Partitioning (RRP), will be used for the real-time allocation of spectrum resources to guarantee the required service quality. Following the allocation of radio resources, the ER Innovative Orchestrator activates the transport domain via the transport controller to configure the transport connectivity as needed.

However, implementing this process "on the fly" (when connectivity with certain requirements is needed) might not be compatible with emergency situations, such as in the UC8 case of the Smart Ambulance. These situations can pose significant challenges due to the dynamic nature of mobility, such as the high-speed velocity of an ambulance during an emergency. Changing network conditions, user mobility, and handovers between different network cells can introduce complexities that could impact QoS and disrupt communication of critical clinical data between the ambulance and the hospital. Additionally, it could also happen that the attempt to activate a high-performance slice in a specific area fails due to a lack of network resources. Therefore, it could be appropriate that certain portions of the transport bandwidth are "reserved" for emergency use (and utilized for best-effort traffic when there are no emergencies). This connectivity is seen as a sort of "emergency lane" in the network. In case of planned special events, associated with high radio traffic density in a particular area, portions of bandwidth can be reserved beforehand to establish a sort of temporary "emergency channels" for ambulance communication in case of emergency. In ordinary conditions, the ER innovative Orchestrator can use additional techniques to mitigate congestion situations and facilitate the availability of resources for emergency case. This aspect will be evaluated in relation with the UCs requirements defined in WP4.

3.2.3 Cellular network information for crowd and traffic monitoring

As an alternative to crowd or traffic monitoring solutions based on video surveillance cameras, ORO is internally developing a proprietary solution based on the signaling data coming from the commercial cellular network. Even if not related directly to the TrialsNet project proposal, the inclusion of this application in the D2.2 deliverable is important due to the flexibility and efficiency of the solution compared to other alternatives.

3.2.3.1 Application architecture

The high-level architecture of ORO's cellular data analytics solution is showcased in Figure 53. As can be seen, for the analysis of urban mobility based on data provided by mobile devices in the mobile telecommunications network, ORO uses location information obtained from the signaling data generated by mobile devices. Each procedure initiated by a mobile device as a result of an event (e.g., call setup, handover, etc.) generates signaling data in the network that always includes information about the cell from which the procedure was initiated. In the mobile network of Orange Romania, a system of probes is implemented that captures all the signaling information circulating on the interfaces of the subsystems that make up the mobile network. The collected data are used for legitimate purposes to determine various events that cause a malfunction of the network, as well as to determine the impact on the services offered. To determine the impact on the services offered and the number of customers affected in the event of a network malfunction in a specific geographical area, information regarding the total number of signaling events of each mobile device in the network every 15 minutes is stored. This information is also used by the mobile data statistical analysis application. Thus, based on the International Mobile Subscriber Identity (IMSI), the radio cell from which the mobile signaled, the number of signaling events of the mobile device in a certain radio cell, and the period of time in which the mobile signaled in the radio cell, different urban mobility statistical indicators can be calculated.

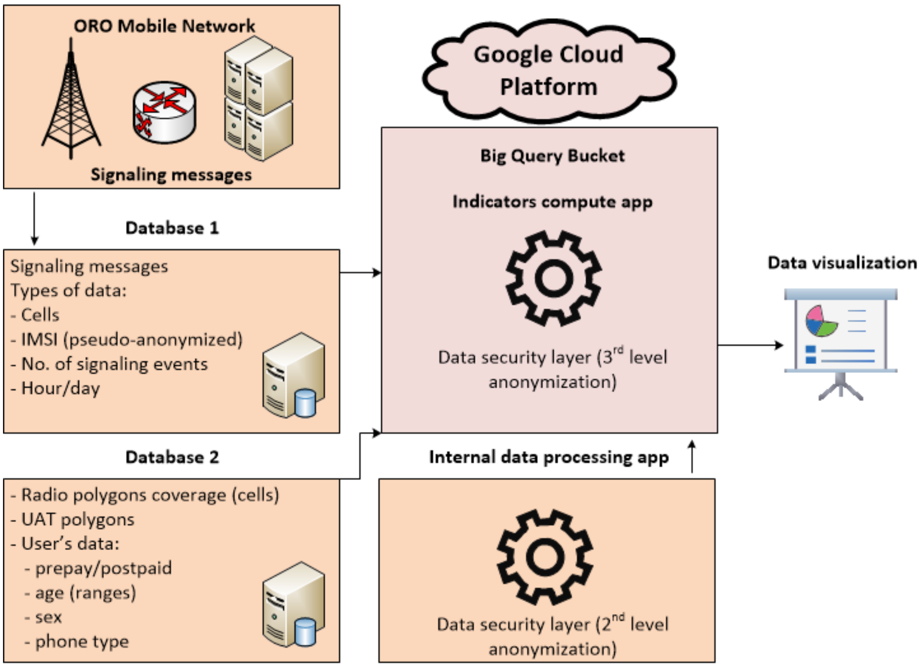

Figure 53. Cellular data analytics application architecture.

The mobile data statistical analysis application also uses other data sources such as radio cell coverage areas, administrative areas at the level of the local administrative units, locality or district in the case of cities. Also, for refining statistical indicators, information regarding the characteristics of mobile devices, the gender, and age categories of users who have subscriptions can be used.

Different algorithms for eliminating errors in the statistical calculation of indicators are also used, such as removing Machine-to-Machine (M2M) devices from the statistical calculation, using the penetration rate of mobile devices at the national level, extrapolating the company's market share at the regional level (e.g., county, city), and distributing users in an analyzed area based on the coverage percentages of radio cells. In the near future, an algorithm for distributing users based on population density information in accordance with land use standards associated with different geographical areas will be implemented.

The only personal data used is the IMSI. Personal data is stored in Google Cloud after it has been pseudonymised through an irreversible algorithm. Also, the stored data can only be accessed after the authorization of persons/machines processing the data, and access can be audited at any time. Data processing can only be initiated following a work order, and data extraction for processing is carried out through a process that includes a second stage of pseudonymisation using a key generated at the issuance of the work order, on a machine independent of the processing platform, which cannot be recovered and is destroyed after the completion of the study or after a maximum of 4 months.

Statistical indicators are calculated independently without being extracted from a general intermediate table. For each statistical indicator, a calculation procedure is used. A third level of security is used, thus, for each procedure that calculates a statistical indicator, a key composed of alphanumeric and special characters is generated at the beginning of the procedure and is destroyed at the end of it. The approximate runtime for procedures that calculate statistical indicators is < 120 seconds for those without recurrence (a day's data), < 300 seconds for two weeks of recurrence.

For additional protection of personal data, administrative roles have been divided among different teams that manage different processes or maintenance activities, so that access to anonymization keys, operating systems, or databases is distributed among several operational teams. Reports can be viewed in an interactive format through Google Data Studio or Tableau applications and downloaded in a csv file format.

Finally, it is important to note that all the data processing steps and methods are following the guidelines and procedures described in WP1 deliverables D1.2 [51] and D1.5 [52].

3.2.3.2 Test on field

Concomitantly with the Sf. Parascheva field trials that were performed using video cameras connected over 5G and were already detailed in D3.2 [5], ORO also recorded the cellular signaling data for this document as a preliminary test of the solution. Based on the analysis of the gathered data, using the procedure described in the previous section, ORO designed a comprehensive report regarding the people dynamics for the period between 8 and 22 October 2023, including the religious event that was officially celebrated on 14th of October.

One of the statistics obtained from the report is the quantitative presence of national visitors that is showcased in Section 2.4.4. As can be seen in Figure 54, during the religious event celebration day and the surrounding days the number of visitors in the Stefan cel Mare pedestrian alley area was much higher than normal, peaking at about 100.000 unique mobile users of 14th of October.

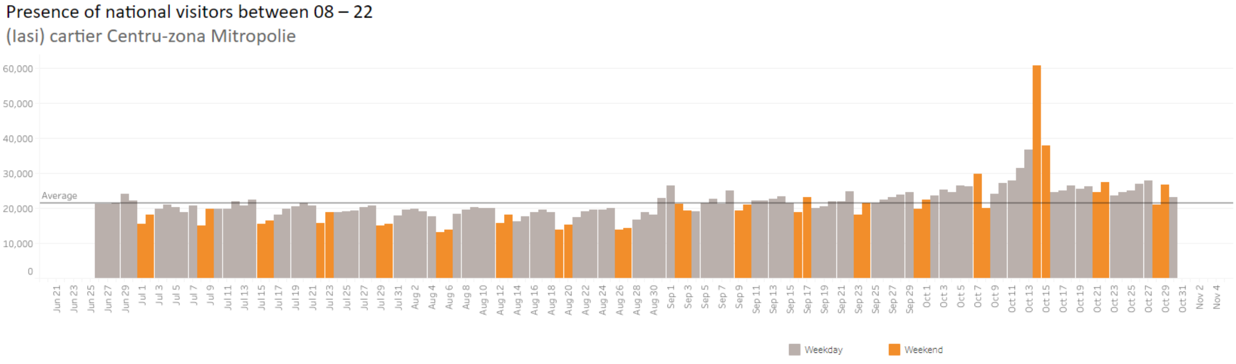

Figure 54. Presence of national visitors report.

Another analysis that was performed is related to the amount of time mobile users stay in the area of interest, as showcased in Figure 55. As can be seen, the number of users that stayed more than 1 hour in the Stefan cel Mare pedestrian alley area on the 14th of October way overpasses the daily averages from the rest of the month, this fact being correlated with the religious proceeding that happened in the morning when the crowds of people stayed almost motionless listening to the priests' speech.

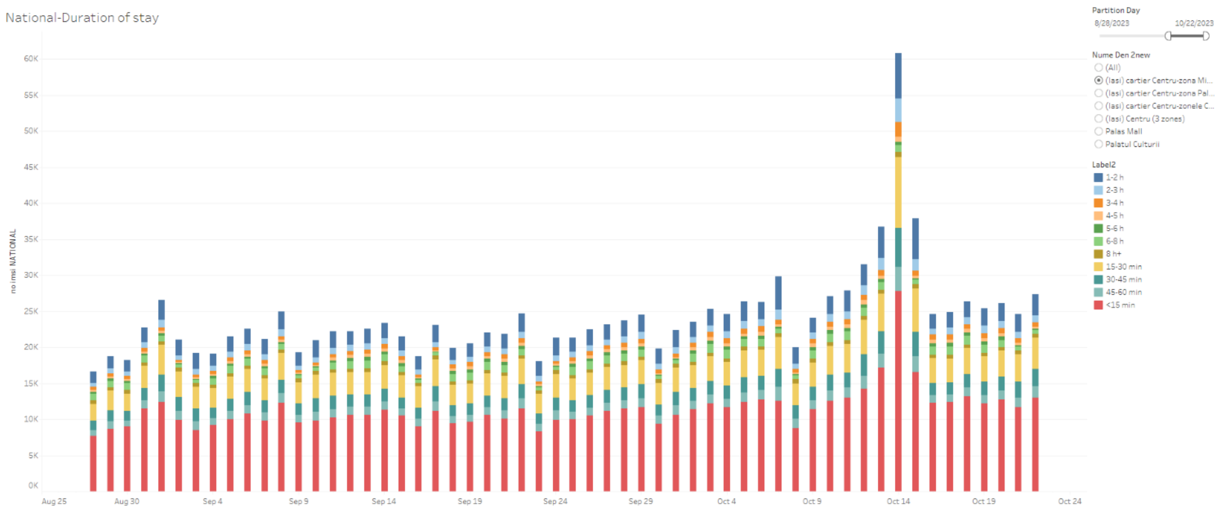

Figure 55. Duration of stay report.

Moreover, based on the signaling data, ORO could also compute the home location of the mobile users, as shown in Figure 56. This report revealed interesting data, showing that 92% of the visitors were coming from Romania, while about 8% were coming from abroad. From this 8% of international users, 22% of them were from Italy, while from the 92% of national users, 62% were coming from Iasi. Looking at the Iasi local visitors, the cellular data analytics also computed the home neighborhood from which they were coming from.

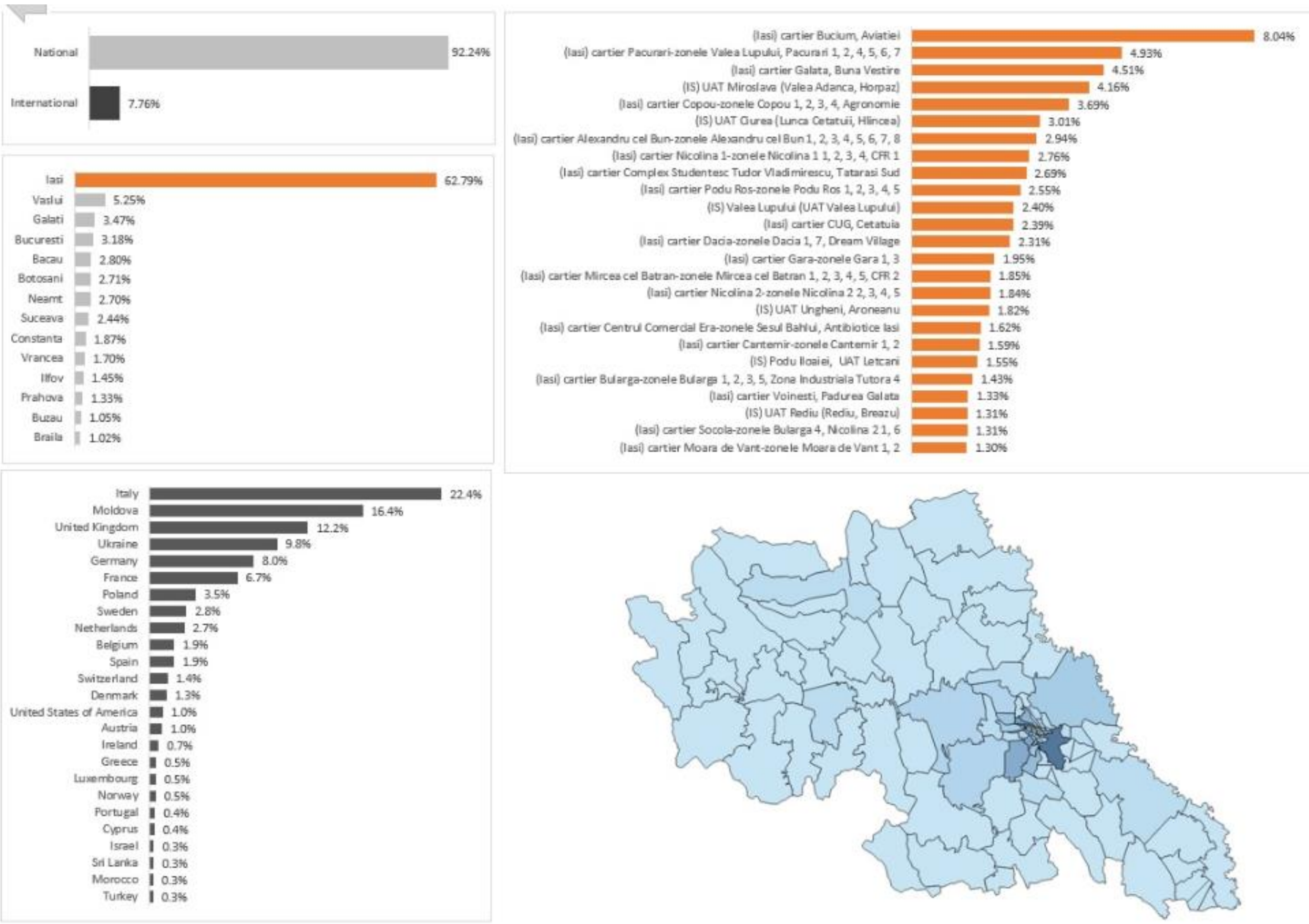

Figure 56. Home location report.

Finally, it is important to note that this analysis could be extended in the future to include the signaling data coming from 5G and perform the data-processing tasks in real-time. Furthermore, based on these advancements, the solution could be also used for the UC4.

4 Sustainability aspects from the infrastructure perspective

TrialsNet identified a set of Key Values (KVs) that are candidates for being impacted by the project technologies [18]. In particular, the project identified values along three different axis (societal, economic, and environmental) which are paramount for the overall handprint of the proposed UCs. Although the innovations in WP2 focus on specific network domains to support the UCs rather than directly benefiting the end-user, they hold the potential to affect different KVs. For example, in relation to the innovations described in Section 3, the following high-level consideration can be made:

- **Zero-touch service management:** This innovation can target the KVs of trust, security, and user experience since it can ensure stable performance and automated management of networks, which can enhance user trust and security while improving the overall experience by reducing latency and increasing reliability.
- **B5G applications framework**: Addressing the KVs of user experience, security, and environmental sustainability, this design keeps data processing at the edge of the network, which can improve application responsiveness, reduce the risk of data breaches, and lower the energy consumption associated with data transmission.
- **Usage of cellular network information for crowd and traffic monitoring**: this innovation addresses the Key Value of Trust, as it is a fundamental functionality for the use cases related to crowd monitoring. Most importantly, this Key Value is enforced through noninvasive techniques that always guarantee the end user privacy.
- **AI mechanisms for diagnostics and resource efficiency**: These mechanisms can improve trustworthiness, security, and economic sustainability by ensuring efficient resource use and preventing failures, thereby optimizing performance and reducing operational costs.
- **Automatic orchestration of network slices to ensure QoS in mobility**: This targets the KVs of service, user experience, and societal sustainability by dynamically assigning network resources to provide consistent quality of service, which is crucial for applications such as smart ambulances and real-time clinical data communication.
- **Digital Twin applied to next generation mobile network**: Targeting the KVs of resilience and environmental sustainability, this innovation involves using a digital replica of a physical network to simulate and predict performance. This can improve resilience by allowing for better planning and response to network issues, and contribute to sustainability by optimizing network resource usage.

Acknowledging the importance of all the KVs, in the context of the definition of the 6G, the sustainability plays a relevant role between the others. The TrialsNet project, focused on technology trials for beyond 5G networks, holds therefore a unique position in advancing the understanding of next-generation mobile networks in terms of this KV. Especially, the work carried out in WP2 allows tackling sustainability under a number of different perspectives, such as:

- **Understanding Sustainability for TrialsNet network technologies**: The overall goal of the TrialsNet project is not just advancing the state of the art of beyond 5G networks by fulfilling the envisioned usa cases providing the required KPIs, but it is also critical assessing how the technology that needs to be developed in order to support the performance requirements can be obtained in a sustainable way, analyzing it from many points of view, especially the environmental and the societal aspects. From the perspective of this WP, this entails evaluating how new network technologies can reduce resource consumption, which in turn minimizes the carbon footprint. The activities carried out in this project hence offer a unique vantage point to understand this impact at scale.
- **Large-scale impact analysis**: The scale at which TrialsNet operates allows for a comprehensive analysis of sustainability impacts. By conducting large-scale trials, the project can gather extensive data on how new network technologies affect energy consumption, resource utilization, assessing the overall environmental impact.
- **Economic and societal impacts**: While the primary focus of WP2 is on the environmental aspect of sustainability, its contribution to the use cases implementation allows to evaluate societal and economics aspects as well. This will be further explored in the WP3, WP4, and WP5 activities in the next months. This involves not just minimizing negative impacts but also discussing how the network can be used to improve these aspects.

In the following, the overall approach followed by WP2 with respect to achieving sustainability is discussed, as well as the discussion of some specific solutions aiming at improving the sustainability in the network.

4.1 Approach to sustainability

In the context of TrialsNet, WP2 plays an important role in addressing sustainability in three distinct yet interconnected domains:

- The first area targets the large-scale trialing of new, energy-efficient infrastructure. This involves the evaluation of the A5G network technologies for reduced energy consumption. The overall objective is to bring into practice effective technologies for the increased resource efficiency in real-world scenarios, whose outcomes may be used to understand the impact on paving the way for more sustainable network operations on a global scale.
- Another area of research that WP2 can leverage is the data-driven evaluation of novel, large-scale network configurations. By modeling the data, the WP2 work may be used to understand and optimize network setups for enhancing sustainability. This includes strategies like dynamic resource allocation that aligns network resources with real-time demand to minimize resource suboptimal utilization. In general, this activity could also be combined with the creation of Digital Twins of the network system, as discussed in Section 3.1.3.
- Finally, the cross-cutting activities carried out by WP2 in conjunction with WP3, WP4, and WP5 allows to extend the focus on sustainability to the domain of application architectures, supporting the development of applications that are natively designed with these criteria. This involves the creation of applications that consume less energy or promote sustainable practices, contributing to an overall reduction in the environmental footprint of network technologies.

Together, these three areas represent a comprehensive approach to embedding sustainability at the heart of next-generation network development.

4.2 Sustainability solutions

4.2.1 Self-sustainable energy harvesting

Self-sustainable energy harvesting allows a completely off-the-grid autonomous solution to power sensors and electronic devices (e.g., processing units, telecommunication systems) used in several use cases of TrialsNet. Using environmental energy instead of non-rechargeable batteries can be beneficial for devices that need low power, long duration and self-reliance. Also, this can help to avoid the need for long wires and power loss during transmission. This scenario is typical of IoT devices and sensors installed outdoor, as in the UC5 where Wi-Fi sensors are installed in the Valentino park in Turin.

Given a system to power, the actual sustainability of self-sustainable energy harvesting in terms of economic and ecological issues must be carefully considered since energy harvesting solutions have not only capital costs, but also operation costs (e.g., maintenance, replacements due to thieves) that must be carefully evaluated. For this reason, there is a need for a simple model tailored to the deployed devices and to the specific environmental conditions in which it is located. In the following, scenarios compatible with sensors (e.g., Wi-Fi sniffers, cameras) which are installed outdoor and exhibit a low-power consumption (e.g., less than 10-20 W) are considered.

Paper [53] offers a survey of the current state of the art regarding energy harvesting for small low power devices, like the sensors who are deployed in TrialsNet scenarios, focusing on the battery and transducers technology. In [54] is devoted specifically to photovoltaic self-powered devices, with several examples of proposed solutions in different use cases and the proposed methodology will be considered for some specific use cases in TrialsNet. Finally, [55] is another survey about the energy harvesting solutions for wireless sensor networks.

Self-sustainable self-charging power units are based mainly on three components, shown in Figure 57: (i) an energy harvesting system which exploits some “natural” environmental power supply (e.g., solar irradiation, wind, mechanical vibrations) to generate electric energy, (ii) an energy storage which allows to store the excess electric energy to power the system when the energy harvesting is not sufficient, (iii) a power management system which coordinates the energy storage and the system powering, in order to maximize the use of the power supply and to minimize the outage probability due to the unavailability of energy to power the devices.

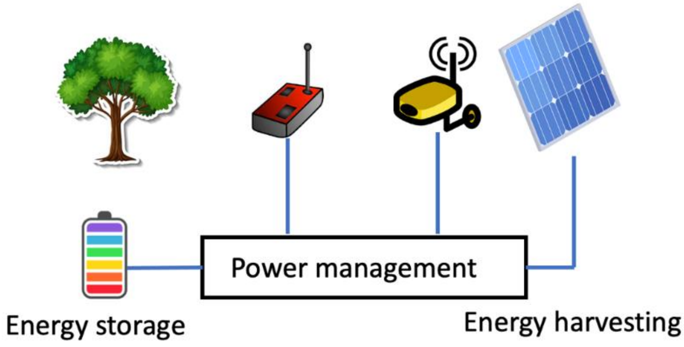

Figure 57. Self-sustainable energy harvesting system.

The energy storage technologies are classified into electromechanical, electrical, thermal, and mechanical. All variants are not considered in this work, for which a large literature is available, but only as example the adoption of electrochemical energy storage, is considered i.e., commonly referred as re-chargeable batteries, since they are the commonly used in scenarios similar to the ones in TrialsNet use cases. Many drawbacks affect their adoption. First, the lifespan of batteries is limited and depends on their actual use, which is hard to be predicted in advance and not always measurable with real-time measurements. As a consequence, planning their replacement is typically done with very conservative approaches, with also an impact from an ecological point of view. High capital cost appears to be one reason for the delay in the commercialization of self-sustainable technology.

The storage efficiency and capital cost of the energy storage component must be considered when developing a self-sustainable technology. In [53], considering Li-On batteries, the following storage capabilities have been reported:

- **Energy density**: 75-250 Wh/L (where L means liter as volume unit)
- **Power densities**: 150-315 W/L
- **Efficiency**: 65-75%
- **Lifetime**: 5-15 years

Furthermore, the following costs have been reported:

- **Power**: 220-1990 USD/kWh
- **Energy**: 16-110 USD/kWh
- **Life cycles**: 700-3000

When disposed of, batteries have still a high impact on the environment, unless modern battery recycling schemes are adopted.

By considering an energy harvesting scheme based on solar power, it is important to evaluate the actual solar irradiation estimated in the area considered for the installation, also considering the position and possible obstacles. The Photovoltaic Geographical Information System (PVGIS) by EU community [56] allows an overall estimation of the average produced energy per day per 1kWp of installed PV panels, by which it is possible to start the initial dimensioning of the harvesting system. Notably, due to variability of the atmospheric condition, the knowledge of the average production is not sufficient to minimize the required storage, but still the value can be used as initial dimensioning of the harvesting system. Thanks to recent crowd-sourcing services (as PVOutput [57]) that monitor the actual production of the PV plants installed in private houses, it is possible to better optimize the dimensioning of the storage system. Furthermore, in the operational costs, the maintenance must be considered (e.g., cleaning of solar power, replacement due to thieves, damages).

Finally, a profiling of the power consumption of the adopted sensors and electronic devices is required. As an example, in UC5, a Raspberry Pi is adopted to implement the sensor and process on-board the Wi-Fi signals. The Raspberry Pi, depending on the model [58] [59], consumes max 5W (950mA) under stress, normally between 1.5W and 2W. The power supply is 5.1V. Universal Serial Bus (USB) devices might draw up to 1.2A, the board will consume typically 400-500mA. In particular, the adopted Raspberry 4B consumes typically

600mA, with a maximum current of 3A, corresponding to around 15W of maximum power. Profiling can be performed through simple measurements and can allow an accurate and less conservative approach to dimension the self-sustainable system.

In the following months, a practical analysis of some scenarios relevant for TrialsNet use cases will be performed to understand the sustainability of an energy harvesting system.

4.2.2 Energy utilization Edge versus Cloud

In the context of the TrialsNet project, ORO, together with IMEC, will develop an E2E orchestration platform that will take advantage of ORO's 5G Labs private network topology, that implements the Edge-Cloud continuum concept by including both central cloud and edge facilities, to monitor and act on the energy utilization data coming from the edge-deployed applications for both UC1 and UC4.

4.2.2.1 Edge-Cloud continuum concept

Over the past few years, cloud and edge computing technologies have rapidly gained prominence as the primary computing paradigms. They cater to a wide range of application domains with varying QoS demands. Traditional cloud and edge computing systems have introduced numerous orchestration solutions to address specific infrastructure aspects (such as cloud-native computing systems and IoT device management) as well as application requirements (such as ultra-low latency and high computational intensity). The current centralized computing model is being expanded to accommodate the needs of applications deployed across different parts of the computing infrastructure. This expansion is taking the form of distributed computing continuum systems. The computing continuum refers to the integration of computing and communication resources and services across an end-to-end infrastructure. The computing continuum expands the existing cloud computing infrastructure by incorporating edge computing and IoT computing devices. The goal is to facilitate the deployment and management of distributed applications throughout the continuum while ensuring energy efficiency, high performance, and meeting security and privacy requirements. The shift towards distributed computing continuum systems presents several challenges that need to be addressed. One of these challenges is the requirement to approach management aspects from a holistic standpoint, where different systems must collaborate towards a shared objective. Operating within the computing continuum involves complex systems, as applications leverage multiple computing tiers and orchestration stacks to function effectively. Within the computing continuum framework, the challenges encountered necessitate the distribution of orchestration across distributed cloud environments, with the aim of approaching management from a comprehensive standpoint. Each stakeholder within the computing continuum, whether they are application, network, infrastructure providers, or end users, establishes their own objectives for application operation. However, a crucial question arises regarding how these objectives can be accurately measured or enforced. Moreover, the computing continuum encompasses a multitude of resources spread across different computing tiers and layers of abstraction. Ensuring the fulfilment of objectives while maintaining system equilibrium presents another unresolved inquiry, that should be addressed through the TrialsNet project developments.

4.2.2.2 Federated architecture with central and edge cloud components

Looking at the high-level architecture showcased in Figure 58, ORO's federated 5G Lab sites will be integrated within the same umbrella, under a Virtual eXtensible Local Area Network (VXLAN) topology, leveraging on the IP FABRIC [60] network paradigm capabilities and the disaggregated nature of the 5G Stand-Alone architecture allowing for the distribution of the UPF units close to the locations where the traffic is generated.

In this approach, the Bucharest 5G Lab site will act as the central cloud facility, while the Iasi 5G Lab site will act as the associated edge facility, both sites being interconnected over a high broadband link of 100Gbps, allowing for the lowest possible latency needed for live migration of VMs or containers.

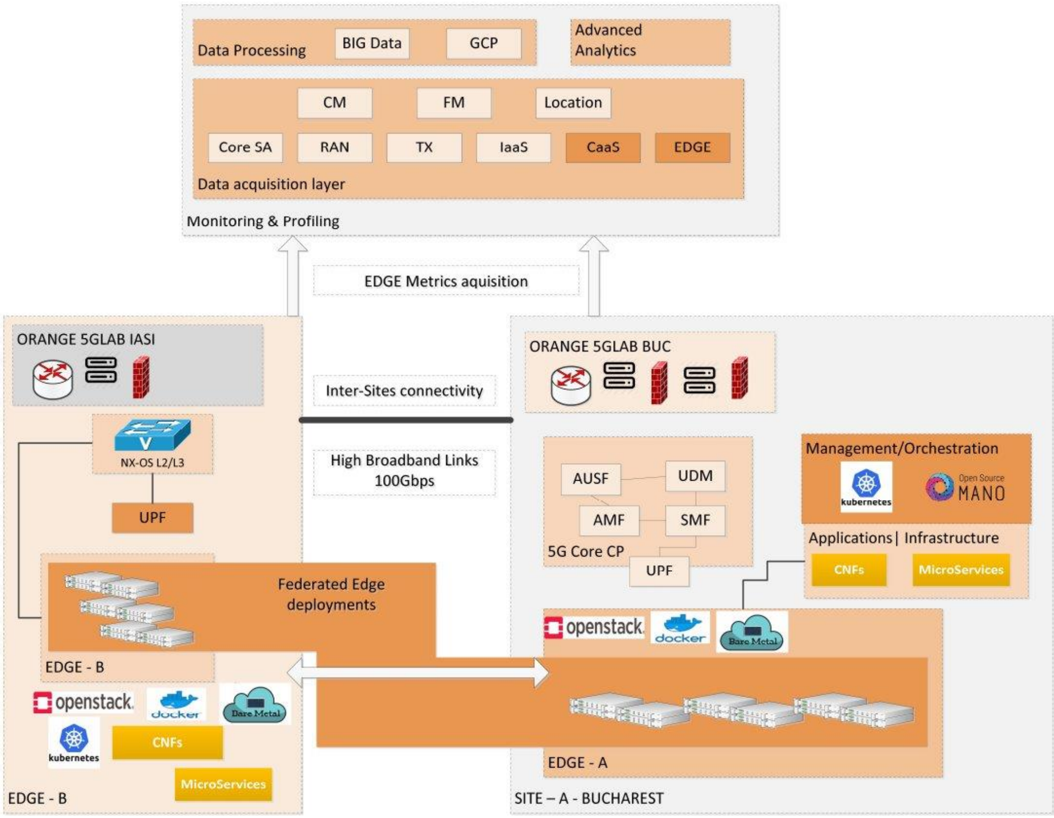

Figure 58. ORO's 5G Labs federated testbeds architecture.

4.2.2.3 Onboarding and intelligence application over the virtualization platforms

The VITAL-5G platform developed within the homonym project [61], already deployed on ORO's 5G Lab infrastructure as well as on IMEC's lab environment, will be also implemented in the scope of the TrialsNet project. From the current VITAL-5G platform architecture the VNF orchestration layer, leveraging on Open-Source MANO (OSM) will be bypassed, leaving space only for the VMs and containers orchestration tools.

In this approach, Kubernetes will be the main orchestration solution that will be integrated in the onboarding platform through specific APIs, while OpenStack will also remain available as an alternative solution to Kubernetes for non-containerized applications.

Both Kubernetes and OpenStack will be cross integrated between the two sites, all the supporting servers appearing as being part of the same cluster. Therefore, Kubernetes' master node, which is already deployed in the Bucharest 5G Lab will be integrated with the worker nodes present in both Iasi and Bucharest, while OpenStack will include two availability zones based on the location, one in Bucharest and one in Iasi, allowing for the migration of VMs between the two regions.

4.2.2.4 Developments towards the target setup

In this approach, IMEC will develop the logic and intelligence allowing for UC1 and UC4 associated applications instantiation and mobility between the two sites in an edge-cloud continuum scenario (Iasi site representing the near edge, while the Bucharest site representing the far edge), based on the network and compute metrics already available through ORO's VITAL-5G monitoring platform and application-level energy consumption data available through specialized tools implemented by ORO in the testbed. The final intelligence application will be integrated in the to-be-developed TrialsNet onboarding platform and will also include ZSM capabilities. Therefore, the two novel architectural components of the platform are described in the paragraphs below.

Energy monitoring using Scaphandre and Kepler

Kepler is an open-source tool founded by Red Hat in collaboration with IBM Research, which captures the power usage metrics from a Kubernetes cluster. In particular, the CPU, Graphics Processing Unit (GPU) and RAM power consumption information is collected using the extended Berkeley Package Filter (eBPF) in the

Linux kernel and RAPL (Running Average Power Limit) metrics, and data can be exported as a set of Prometheus metrics and displayed in dashboards using Graphana. One of the aims of Kepler is the integration of carbon intensity data.

Scaphandre is another open-source metrology agent that shows the power consumption of a single process. It has been developed using RUST. It uses the PowercapRAPL sensor, which reads the values of the energy counters from powercap. It then stores those values, and does the same for the CPU usage statistics and for each running process on the machine at that time. From there, it is possible to compute the ratio of CPU time actively spent for a given PID (Process ID) over the CPU time actively spent doing something, in order to estimate the subset of power consumption that is related to that PID on a given timeframe. As in Kepler, and data can be exported as a set of Prometheus metrics and displayed in dashboards using Grafana.

In particular, ORO's focus will be on Scaphandre, since it was already integrated in the 5G Lab testbed and can measure power consumption on bare-metal hosts or virtual machines from the hosts, as well as exposing power consumption metrics of a virtual machine, to allow manipulating those metrics in the VM as if it was a bare metal machine, whereas Kepler exposes only power consumption metrics of Kubernetes, which is quite constrictive for ORO's aims.

Decision making logic

The decision-making tool that would be developed allowing for energy aware applications and ZSM capabilities is described in Section 4.2.3 and Section 3.1.1, respectively.

4.2.3 Energy aware application design

The edge computing environments are not only specific for the scarcity of computing resources, as discussed in Section 3.1.1, but also for the challenges related to energy consumption and preservation. For instance, some edge devices operate on limited power sources, such as batteries, which is quite specific for aerial edge deployments such as Unmanned Aerial Vehicles (UAVs) that are gaining momentum when it comes to ad-hoc provision of network and edge computing capabilities. This constraint poses challenges for energy efficiency, as these devices need to balance processing capabilities with power consumption to extend battery life.

Despite the extensive analysis of various design aspects of B5G Applications framework (Section 3.1.2), the energy-awareness is not yet considered when it comes to innovations in the architectural framework. Nevertheless, this specific feature seems to be relevant given the abovementioned challenges in B5G networks, in particular Edge networks, that need to be sustainable. In addition to batteries and other types of energy sources for EdgeApps and Edge computing environments in general, the type of hardware used in edge devices significantly influences energy consumption. For example, more efficient processors, low-power components, and optimized architectures are essential to minimize energy usage while maintaining performance. The application design should focus on the last one, i.e., optimized architecture for minimizing energy usage while maximizing service performance.

However, the energy-awareness of EdgeApps does not necessarily mean that the application itself becomes fully responsible for its own performance. On the contrary, in this TrialsNet activity, the focus is on creating Zero-touch Service Management (ZSM) framework with the role of event-driven network and service programming that apart from performance requirements and real-time monitoring data, also consumes energy reports from each EdgeApp instance to fine tune the orchestration decisions. Therefore, the quality-aware and energy-aware functions are integrated into both ZSM layer and EdgeApp functions, whereas both layers are encouraged to apply optimization techniques that will minimize energy consumption. The application functions running within an EdgeApp are communicating events with an overarching ZSM orchestrator (Figure 59), which is in charge of reprogramming network and services to fulfil identified quality and energy needs. In Figure 59, different colors indicate energy consumption of specific EdgeApps, creating energy awareness of ZSM layer on energy consumption of particular EdgeApp instances. The goal is to achieve trade-off between energy efficiency and performance (mainly reliability and end-to-end latency), by also exploiting resource sharing between AI/ML modules running in ZSM functions and user applications.

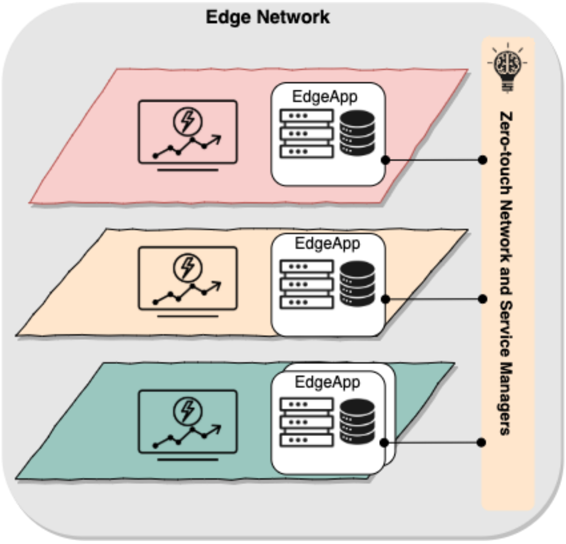

Figure 59. A simplistic view on interaction between energy-aware EdgeApps and ZSM units.

Some optimization techniques that EdgeApps could deploy to improve their energy consumption, i.e., to reduce their energy footprint, are algorithms for efficient task scheduling, data compression, and power-aware computing to ensure that energy is used consciously and responsibly. These specific optimization techniques will be further studied in the context of this sustainability aspect, and as such will be evaluated in the trials. On the other hand, further optimization of energy consumption will be handled by ZSM, as dynamic resource allocation strategies play a crucial role in energy-efficient edge computing. This means that ZSM layer will employ techniques that involve adjusting computational resources based on workload demands, allowing devices to scale up or down their processing power as needed. Finally, as per discussion in Section 4.2.1, renewable sources and energy harvesting devices will be closely studied in relation with EdgeApps that might be deployed on them. The integration of renewable energy sources could be a design strategy for edge infrastructures, thereby harnessing solar, wind, or other sustainable energy sources to provide power to edge nodes.

During the trialling activities, especially in the context of UC4, the main KPIs that will aim improvements will be energy efficiency, end-to-end latency, and reliability. On the other hand, the KVIs that will be tackled in the UC4 context and energy aware applications are environmental and societal. A summary of sustainability aspects that are tackled and will be evaluated in the context of energy-aware EdgeApps is provided hereafter:

- **Societal sustainability**: Reduced emergency/support response time for VRUs (indirect impact from ZSM reflected in terms of efficient application service placement and improved reliability), increased operational efficiency of providing support to VRUs in urban and rural environments (indirect impact from ZSM reflected in terms of energy-aware deployment of new instances of services at required locations); Improved reliability of smart traffic management services that prevent injuries of VRUs, positive impact on VRU safety
- **Environmental sustainability**: Reduced energy consumption will be brought at distributed and ubiquitous edge computing nodes in dense urban environments and highways. This reduction in energy consumption will be considered in two layers:
    - **EdgeApp functions**: Energy-aware functions in EdgeApps refer to the following: If AI/ML model needs to be retrained and function restarted, an event is detected by ZSM orchestrator that further triggers reconnection of user to another instance of service, whereas the one that requires more energy is getting migrated to another environment (e.g., cloud). In that case, ZSM orchestrator is reprogramming the communication flow between user and application service in order to maintain energy consumption and strict latency and throughput performance requirements
    - **ZSM service functions**: Intelligent placement of user application functions to reduce energy footprint (reduced number of instances deployed in distributed ecosystem while pertaining QoS levels). Decisions made by orchestration entities are constructed to reduce energy consumption

In summary, the energy consumption of EdgeApps is a critical consideration due to the unique challenges posed by the distributed nature of edge computing. This TrialsNet activity will be focused on optimizing energy consumption by edge deployments, developing efficient algorithms at both EdgeApp and ZSM layers, and possibly integrating renewable energy sources to address these challenges and make edge computing more sustainable.

5 Conclusions

This deliverable examined the most recent advancements in the evolution of the platform and network solutions in TrialsNet, transitioning from Bs5G to A5G. It also delved into the strides made in both horizontal and vertical network-related TrialsNet innovations and provided an analysis of various network sustainability aspects.

The document provided a snapshot of current progress of the infrastructures deployment and configuration for each of the trial sites within the four main clusters, i.e., Italian, Spanish, Romanian, and Greek, where the different use cases will be trialled in the context of the activities of WP3, WP4, and WP5. In the case of the Turin site (Italian cluster), the commercial 5G network deployment is available as Bs5G, whereas the A5G capabilities are reflected in the VNF orchestration properties, which will be employed on the service level (i.e., service orchestration), i.e., to improve the overall service performance for UC5, UC12, and UC13. The Pisa site (Italian cluster) is being built from scratch, supporting both Release 16 (network slicing) and 17, with new frequency spectrum bands (26 GHz, mmWave range) and advanced orchestration capabilities (A5G), which will ensure required performance levels for UC7, UC8, and UC9. The Spanish cluster is reporting updates towards Release 17 and 5G high-band as part of A5G, towards improving uplink throughput capabilities that are necessary for UC1, UC6 and UC10 implementation and testing. Concerning the Romanian cluster, 5G SA deployment in Iasi is now enriched with edge computing capabilities that enabled edge deployments for core functions such as UPF in support to the implementation of UC1 and UC4. Within the same cluster, experimental facilities located in Antwerp are enriched with the PoC deployment of ZSM framework and intelligent orchestration functions as A5G, which will be transferred to Iasi in the next phase in the context of UC4. Finally, the Greek cluster reported advancements on top of the commercial 5G network in Athens International Airport and public venues in the city of Athens, with a private 5G testbed with distributed edge computing capabilities in the context of A5G used for testing and validation of UC2, UC3, UC6, UC11, and UC13.

The current status of all trial sites is supported by the preliminary performance measurements and related evaluation reported in D3.2 [5], D4.2 [6], and D5.2 [7], which will help to understand the capabilities of the current deployments and possible enhancements that need to be made to fulfil the use case requirements defined in D3.1 [2], D4.1 [3], and D5.1 [4]. It is important to note that all trial sites have been made available for hosting new UCs from the Open Call, whereas the additional details about specific tests and infrastructural upgrades needed to support them will be provided in the next deliverables D2.3 and D2.4.

Concerning the TrialsNet innovations from the network perspective, this deliverable reported the latest insights into the design and implementation progress of both horizontal (network innovation applicable to all verticals) and vertical (specific to a vertical or particular use case) innovations. The additional vertical innovation related to using the cellular network information for crowd and traffic monitoring has been also introduced. The report included a more specific reference to their validation strategies related to specific UCs developed by the project.

Given the importance of sustainability for any Beyond 5G and 6G trial activity, this deliverable dived deeper into the network sustainability perspectives, focusing on understanding the sustainability aspects for TrialsNet network technologies, large scale impact, and economic and societal benefits. The topics such as self-sustainable energy harvesting, energy utilization at edge versus cloud, and energy-aware application design, have been studied and reported.

Acknowledgment

TrialsNet project has received funding from the European Union's Horizon-JU-SNS-2022 Research and Innovation Programme under Grant Agreement No. 101095871.

References

[1] TrialsNet, “Deliverable D2.1 - Preliminary design aspects for Platforms and Networks solutions”, June 2023.

[2] TrialsNet, “Deliverable D3.1 - Use Cases definition for Infrastructure, Transportation and Security & Safety (ITSS) domain”, April 2023.

[3] TrialsNet, “Deliverable D4.1 - Use Cases definition for eHealth and Emergency (eHE) domain”, April 2023.

[4] TrialsNet, “Deliverable D5.1 - Use Cases definition for Culture, Tourism, and Entertainment (CTE) domain”, April 2023.

[5] TrialsNet, “Deliverable D3.2 - First results of Use cases implementation for ITSS domain”, March 2024.

[6] TrialsNet, “Deliverable D4.2 - First results of Use cases implementation for eHE domain”, March 2024.

[7] TrialsNet, “Deliverable D5.2 - First results of Use cases implementation for CTE domain”, March 2024.

[8] H2020, “5G-EVE,” [Online]. Available: https://www.5g-eve.eu/. [Accessed 2024].

[9] H2020, “5G-TOURS,” [Online]. Available: https://5gtours.eu/. [Accessed 2024].

[10] Dropper, “Dropper,” [Online]. Available: https://www.dropper.ai. [Accessed 2024].

[11] Ericsson, “Ericsson Antenna System Catalog,” Ericsson, 2024. [Online]. Available: https://www.ericsson.com/en/antenna-system/forms/ericsson-antenna-system-catalog-2023 . [Accessed 2024].

[12] Ericsson, “Multi-band booster: Enabling 5G transport everywhere with E-band,” Ericsson, 2024. [Online]. Available: https://www.ericsson.com/en/mobile-transport/multiband. [Accessed 2024].

[13] “Poweredge R640 specification,” DELL Technologies, [Online]. Available: https://i.dell.com/sites/csdocuments/Product_Docs/en/poweredge-r640-spec-sheet.pdf. [Accessed 2024].

[14] VITAL-5G, “D2.2 VITAL-5G experimentation platform - Early (testing),” 2023. [Online]. Available: https://www.vital5g.eu/wp-content/uploads/2023/05/VITAL5G-D2.2_VITAL-5G_experimentation_platform_Early_testing_drop_v2.0_final.pdf.

[15] Nokia, “FastMile 5G Receiver 5G14-B,” [Online]. Available: https://www.nokia.com/networks/fixed-networks/fastmile/5g-receiver-5g14-b/.

[16] Teltonika, “RUTX50,” [Online]. Available: https://teltonika-networks.com/products/routers/rutx50.

[17] Kubernetes, “Kubernetes: Production-Grade Container Orchestration,” [Online]. Available: https://kubernetes.io/. [Accessed 2024].

[18] TrialsNet, “Deliverable D6.1 - First report on validation and dissemination activities”, Novemeber 2024.

[19] Open5GS, “Open Source implementation for 5G Core and EPC,” [Online]. Available: https://open5gs.org/.

[20] Free5GC, “Free5GC,” [Online]. Available: https://free5gc.org/.

[21] SRS, “Open Source RAN,” [Online]. Available: https://www.srslte.com/.

[22] Ettus, “USRP,” [Online]. Available: https://www.ettus.com/all-products/ub210-kit/.

[23] Intel, “Intel NUC,” [Online]. Available: https://www.intel.com/content/www/us/en/support/products/98414/intel-nuc.html. [Accessed 2024].

[24] Quectel, “Quectel 5G Module,” [Online]. Available: https://www.quectel.com/5g-iot-modules. [Accessed 2024].

[25] C. Wireless, “MK5,” [Online]. Available: https://www.cohdawireless.com/solutions/hardware/mk5-obu/. [Accessed 2024].

[26] ETSI, “Intelligent Transport Systems (ITS); Access layer specification for Intelligent Transport Systems operating in the 5 GHz frequency band,” [Online]. Available: https://www.etsi.org/deliver/etsi_en/302600_302699/302663/01.02.00_20/en_302663v010200a.pdf. [Accessed 2024].

[27] Peplink, “Peplink 5G Routers for Robust Wireless Connectivity,” [Online]. Available: https://www.peplink.com/technology/what-is-5g-with-peplink/. [Accessed 2024].

[28] 5GTours, “5G Tours website,” [Online]. Available: https://5gtours.eu/.

[29] N. Slamnik-Krijestorac, M. Camelo, C. Y. Chang, P. Soto, L. Cominardi, D. de Vleeschauwer, S. Latre and J. M. Marquez-Barja, “AI-empowered management and orchestration of Vehicular Systems in the Beyond 5G Era,” *IEEE Network*, 2023.

[30] ETSI, “Zero-touch network and Service Management (ZSM); Terminology for concepts in ZSM,” [Online]. Available: https://www.etsi.org/deliver/etsi gs/ZSM/001 099/007/02.01.01 60/gs ZSM007v020101p.pdf. [Accessed 2024].

[31] ETSI, “Zero-touch network and Ser vice Management (ZSM); Reference Architec-ture. ETSI GS ZSM 002 V1.1.1 (2019-08),” [Online]. Available: https://www.etsi.org/deliver/ etsi gs/ZSM/001 099/002/01.01.01 60/gs ZSM002v010101p.pdf. [Accessed 2024].

[32] E. Coronado, R. Behravesh, T. Subramanya, A. Fernandez-Fernandez, M. S. Siddiqui, . X. Costa-Perez and R. Riggio, “ero touch management: A survey of network automation solutions for 5G and 6G networks,” *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, p. 2535–2578, 2022.

[33] M. Liyanage, Q.-V. Pham, K. Dev, S. Bhattacharya, P. K. R. Mad dikunta, T. Gadekallu and G. Yend, “A survey on zero touch network and service management (ZSM) for 5G and beyond networks,” *Journal of Network and Computer Applications*, vol. 203, 2022.

[34] Zenoh, “Zenoh Eclipse,” Zettascale, [Online]. Available: https://zenoh.io/. [Accessed 2024].

[35] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. -C. Liang and D. I. Kim, “Applications of Deep Reinforcement Learning in Communications and Networking: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, p. 3133–3174, 2019.

[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland and G. Ostrovski, “Human-level Control Through Deep Reinforcement Learning,” *Nature*, vol. 518, no. 7540, p. 529–533, 2015.

[37] C. D. Alwis, A. Kalla, Q. Pham, P. Kumar, K. Dev, W. Hwang and M. Liyanage, “Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research,” *IEEE Open Journal of the Communications Society*, vol. 2, pp. 836-886, 2021.

[38] ETSI, “White Paper No. 55: MEC support towards Edge Native Design,” [Online].

[39] N. Slamnik-Kriještorac, F. Z. Yousaf, G. Y. Yilma, R. Halili, M. Liebsch and J. M. Marquez-Barja, “Edge-Aware Cloud-Native Service for Enhancing Back Situation Awareness in 5G-Based Vehicular Systems,” *IEEE Transactions on Vehicular Technology*, vol. 73, no. 1, pp. 660-677, 2024.

[40] G. Association, “GSMA official website,” [Online]. Available: https://www.gsma.com/. [Accessed 2024].

[41] 3GPP, “3GPP SA6,” [Online]. Available: https://www.3gpp.org/technologies/sa6-app-enable.

[42] 3GPP, “CAPIF,” [Online]. Available: https://www.tech-invite.com/3m23/toc/tinv-3gpp-23-222_d.html.

[43] S. Barmpounakis, “Data Analytics for 5G Networks: A Complete Framework for Network Access Selection and Traffic Steering,” 2018.

[44] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan and T. Woo, “CloudIQ: A Framework for Processing Base Stations in a Data Center,” in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, ser. Mobicom ’12. New York, NY, USA: Association for Computing Machinery*, 2012.

[45] G. Garcia-Aviles, A. Garcia-Saavedra, M. Gramaglia, X. Costa-Perez, P. Serrano and A. Banchs, “Nuberu: Reliable RAN Virtualization in Shared Platforms,,” in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, ser. MobiCom ’21. New York, NY, USA: Association for Computing Machinery*, 2021.

[46] J. A. Ayala-Romero, A. Garcia-Saavedra, M. Gramaglia, X. Costa-Perez, A. Banchs and J. J. Alcaraz, “VrAIn: A Deep Learning Approach Tailoring Computing and Radio Resources in Virtualized RANs,” in *The 25th Annual International Conference on Mobile Computing and Networking, ser. MobiCom ’19*, 2019.

[47] D. Bega, M. Gramaglia, X. Costa-Pérez and P. Rost, “CARES: Computation-Aware Scheduling in Virtualized Ra- dio Access Networks,” *IEEE Transactions on Wireless Communications,* vol. 17, no. 12, p. 7993–8006, 2018.

[48] L. Bonati, M. Polese, S. D'Oro, S. Basagni and T. Melodia, “OpenRAN Gym: An Open Toolbox for Data Collection and Experimentation with AI in O-RAN,” in *IEEE Wireless Communications and Networking Conference (WCNC),* 2022.

[49] srsRAN, “Software Radio Systems (SRS), “srsRAN.”,” [Online]. Available: https://www.srslte.com/.

[50] N. Apostolakis, M. Gramaglia and P. Serrano, “Design and Validation of an Open Source Cloud Native Mobile Network,” *IEEE Communications Magazine,* vol. 60, 2022.

[51] TrialsNet, “Deliverable D1.2 - Ethics Assessment Plan,” 2023.

[52] TrialsNet, “Deliverable D1.5 - Data management Plan,” 2023.

[53] K. Calautit, D. S. Nasir and B. R. Hughes, “Low power energy harvesting sys-tems: State of the art and future challenges,” *Renewable and Sustainable Energy Reviews, Elsevier,* 2021.

[54] D. Hao, L. Qi, A. M. Tairab, A. Ahmed, A. Azam, D. Luo, Y. Pan, Z. Zhang and J. Yan, “Solar energy harvesting technologies for PV self-powered applications: A comprehensive review,” *Renewable Energy, Elsevier,* 2022.

[55] A. J. Williams, M. F. Torquato, I. M. Cameron, A. A. Fahmy and J. Sienz, “Survey of Energy Harvesting Technologies for Wireless Sensor Networks,” *IEEE Access,* 2021.

[56] “Photovoltaic Geographical Information System (PVGIS),” [Online]. Available: https://re.jrc.ec.europa.eu/pvg_tools/. [Accessed 19 04 2024].

[57] “PVOutput,” [Online]. Available: https://pvoutput.org/. [Accessed 19 04 2024].

[58] “Raspberry Pi Power supply requirements,” [Online]. Available: https://github.com/raspberrypi/documentation/blob/develop/documentation/asciidoc/computers/raspberry-pi/power-supplies.adoc. [Accessed 19 04 2024].

[59] “Power Consumption Benchmarks,” [Online]. Available: https://www.pidramble.com/wiki/benchmarks/power-consumption. [Accessed 19 04 2024].

[60] R. 8. standard, “Segment Routing Architecture,” July 2018. [Online]. Available: https://datatracker.ietf.org/doc/rfc8402/.

[61] VITAL-5G, “Project webpage,” [Online]. Available: https://www.vital5g.eu/.

Annex A

This Annex A provides more detailed information about the intent-based network management as part of the ZSM framework, described in Section 3.1.1.

Intent-based network management

According to the ETSI ZSM Group, intent-based interactions in autonomous network management frameworks are part of ZSM, and the goals and expected behavior are intent-defined as agnostic instructional language that enables services to interact with each other independently of the diverse technologies from the vendors, providers and operators. The management of intents includes the interpretation of the requests but also an optimal coordination for its fulfillment to satisfy the users' Quality of Service (QoS). As the ZSM reference architecture [30] is service-based, intent-driven management has to perform cooperatively like an Intent Management Entity (IME) within the ZSM framework to handle the intent requests from the Management Service (MnS) producer who is the intent owner, and then generate the intent requests (MnS consumer) to a further level across the domains. IMEs can also contain knowledge-driven decision-making processes [31].

A basic intent interpretation is used in the current model, which motivates the goal to expand the capabilities of the current PoC to optimally coordinate the interpretation and execution of intents. To accomplish this, an IME is proposed to be in charge of the interpretation that can be used by different intent owners like demanding users/operators or applications placed in a high-level domain and another for the coordination of the intents at the domain where the decision making process takes place. Currently, a try/fail approach is used to ensure the fulfillment of the request in case of failure but with a pre-defined number of trials based on priority. The goal is also to reduce the consumption of resources dedicated to coordinating the execution of the intents, by the use of time-based criteria to reduce the overflowing queues [5], where waiting intents can be sorted by the use of advanced prediction methods that are based on AI/ML techniques.

An IME component is essential for the implementation of Zero-Touch Service (ZTS) within the PoC framework. It plays a crucial role in translating high-level objectives into actionable directives for the network infrastructure. The intricate nature of ZTSs, designed for automated management of diverse network functions, demands mechanisms to interpret and execute user intentions. IMEs have a knowledge base that contains the intent ontology [3] and are also found to be queue-based as the requests are coordinated based on priority and availability of resources [2]. The lifetime of an intent can depend on its priority, which determines the number of tries for its fulfillment. These trial/error approaches tend to increase the consumption of resources as the number of attempts and timing are predefined before they are queued along with the rest of the requests and consequently are fixed. These approaches can benefit from a more accurate determination of the lifetime of the intent by extending the criteria for priority classification. Machine reasoning plays a key role in intent management, with its capability to understand abstract concepts from diverse domains and provide precise, specialized conclusions based on precedent as it can contribute to the quantification of risk and uncertainty, making decisions in circumstances of conflicting goals and new situations [3]. In this section, two crucial features of IME are addressed, namely intent interpretation and intent coordination. In the case of interpretation operations, they are mainly located in the first line of the ZSM architecture (see Figure 35) where the interaction of the users or operators is made through high-level instructions. Regarding intent coordination, it can be located near the inner domains of the ZSM architecture, managing the resulting outputs of the decision-making processes where more intents are generated with more specific instructions and targets.

Intent interpretation

Intents can serve various purposes, such as configuring or overseeing the performance of the network or service, and more. To cope with these purposes, an enhanced intent interpreter algorithm needs to be designed to provide an interface that receives and delivers instructions from high-level intents and compiles them into more detailed and technically elaborated intents that can be used by the destined orchestrators/controllers that will be applying changes and configurations to comply with requirements from the intent. Following the specifications from ETSI [3], the design of the intent interpreter should ensure that intents are expressed in terms of the intent owner, which does not specify how the desired service is to be realized. A pre-process can be employed to deconstruct natural language intents where the intents are formatted into the structure that is expected by the IME. These intents are later submitted in the network as standardized information objects. Among related works that propose different approaches for the development of intent interpreters, a more generic proposal comes from McNamara et al., as they present a flexible interpreter for intent realization [7]. Firstly, they identify a mathematical and a model-based as a means of representation for intents. Secondly, a flexible interpreter takes care of the aforementioned representations through the following stages: functionality templates, intent matching and translation.

Intent coordination

B5G and 6G future networks are essentially intent-driven and proper coordination of the realization of the intents contributes to their success and eventually the performance of autonomous network management frameworks. In the search for practices in intent coordination, there is an increasing attention to the subject from researchers and network operators as well. In this subsection, different approaches that are used to manage the execution of intents are presented and discussed, and are illustrated in Figure 60.

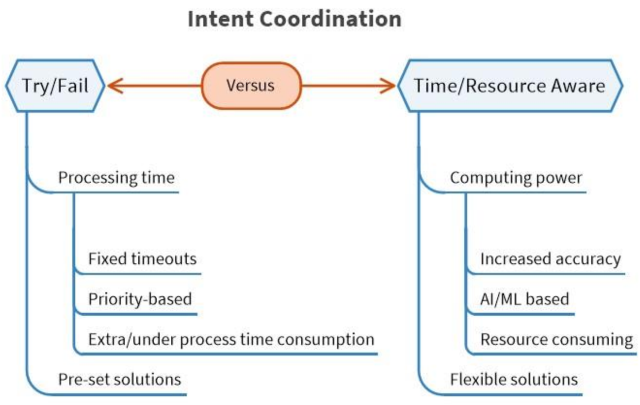

Figure 60. Approaches for intent-driven coordination.

TianZhang He et al [9], propose an approach to orchestrate networking and computing resources based on user requirements while filling the gaps in existing Vehicular Edge Computing (VEC) applications. This involves developing necessary algorithms that take into account both the computing and networking needs of the applications. As a result of their work, two algorithms for intent resolution were designed:

- **Priority-Aware Intent Installation**: This algorithm periodically checks the intents associated with a suspended event. If an intent cannot be satisfied, the algorithm switches the state of the event to 'Failed' and schedules it for reinstallation, with a threshold of 3 attempts.
- **Location-Aware Mapping**: Consists of a mapping microservice dedicated to determining the best path for allocating virtual nodes by taking location contains from the intents. The algorithm later sorts the searched nodes based on the depth and the distance, which is particularly useful for mobile end-users like vehicles and VRUs where the location of the virtual node is expected to change in time. This would minimize the impact on node reallocation.

In the case of the Priority-Aware Intent Installation algorithm, the retry threshold can be dynamically adjusted based on current edge conditions. In dynamic environments, a low threshold may lead to a high failure ratio, while a high threshold can increase intent reinstallation and processing time. Additionally, a larger retry threshold may enhance the acceptance rate for high-priority intents. Retry attempts are traditionally set at a fixed number, such as three tries. However, there is an opportunity for more intelligent determination of retry parameters, including timeouts, by incorporating AI/ML support. This would allow for dynamic adjustments based on rules rather than relying solely on fixed numerical values, although is not exempt from challenges like extra computing efforts for prediction and mapping. Time-based coordination in intent-driven management means that if intent processing is not optimized, then the expected effect over performance would be contradictory. Instead of trying for re-compilation, the intent manager stands by for notification or for a queue of executions for when the conditions are suitable.

The monitoring/update application can collect the requests and give them a timeout that can be pre-set or determined based on AI-based time estimation. In Figure 60, two different approaches regarding execution threshold in intent coordination are compared. AI/ML techniques can be used to prioritize orders in a queue by implementing a priority queue system. The prediction tool that can be used to estimate when a particular service is going to be executed based on a queue of pending actions and the current status of available resources is an AIpowered queue management system. By incorporating AI/ML into intent coordination, specifically through an order queuing tool that estimates waiting times, more accurate sorting of pending intents can be achieved. This enables the determination of estimated response times, ultimately reducing costs associated with holding unnecessary resources. The goal is to identify a suitable prediction tool within AI/ML that can estimate when a specific service will be executed based on a queue of pending actions and the real-time status of available resources. This approach aims to optimize service execution by providing accurate estimates based on dynamic factors, thereby improving overall efficiency in the queuing system.

A systematic assessment of the proposed time-based intent coordination, in conjunction with the previously discussed coordination approaches, will be undertaken. This evaluation aims to provide valuable insights into the effectiveness and performance of time-based intent coordination, offering a nuanced understanding of its impact on the overall orchestration and management of vehicular network services within the evolving landscape of advanced network management frameworks.

Annex B

This Annex B contains more background information about different application frameworks presented in Section 3.1.2.

3GPP Architecture for enabling Edge Applications

In the architectural overview presented in Figure 46 (Section 3.1.2), the communication between different architecture elements is defined via a set of reference points, i.e., EDGE-1-EDGE-7. In particular, EDGE-1 and EDGE-4 are important reference points as they enable creating interfaces between edge applications residing in the edge data network, and e.g., client applications in vehicles. EDGE-1 is used for registration/deregistration between EEC and EES, retrieval of configuration data of EAS, discovery of EAS, etc. Similarly, EDGE-4 creates bridge between EEC and EES, where ECS can be preconfigured on the client side, and is used by vehicle to retrieve information about edge data network.

Concerning the relationship with ETSI MEC, EAS is compatible with MEC application. To be able to discover and consume EASs, i.e., MEC applications or EdgeApps, client application first needs to register on the edge data network, thereby becoming edge-aware and capable to exploit edge services. For example, in the bootstrap phase: i) EEC registers at ECS or EES, ii) EAS registers at EES, and iii) EES registers at ECS. In the client registration, EEC provides its Security credentials, Edge Enabler Client context ID, Application Client Profile, Source Edge Enabler Server ID (Identifier of the EES that provided Edge Enabler Client context ID) and Endpoint (i.e., URI, IP address). After registration, if admitted to edge data network, client in the vehicle is able to discover different services.

After registration, the client can create an EAS discovery request, and in 3GPP there are two modes available, i.e., pub/sub and request/response. EEC sends a request to EES, or subscribes to discovery topics, by providing its authorization credentials needed, filter: application type, etc., and EES receives it and based on the topological/geographical service area, it determines whether client should be connected to EAS or not. If authorized to connect to this edge data network, vehicle gets response from EES, i.e., the information about the discovered EAS endpoints and additional information such as KPIs, application client locations that are supported by application server, and endpoint of EAS that is used for routing outgoing application traffic. There is an additional option to dynamically discover EASs that are important for vehicle, by creating an EAS dynamic information subscription. Thus, any change that happens on the server side (e.g., i) change in MEC applications due to reconfiguration, or change in software version and image, ii) change in availability, iii) change of endpoint, etc.), or any change that happens on the client side (e.g., change in location due to mobility, change in application type preference in the discovery request, etc.), reflect on the EAS availability for client, and using EDGE-1/EDGE-4 vehicle gets informed about new EAS availability list.

When user consuming the service moves to a new location, different EASs can be more suitable for serving the application clients in the vehicle. Thus, it is important to support service continuity to minimize service interruption while replacing the serving EAS, called source EAS, with a target EAS. As EAS, or EdgeApp, is associated with an application-context, this context needs to be transferred from source EAS to target EAS, before vehicle reconnects from one to another. In the same 3GPP standardization track, there are three situations in which application-context needs to be transferred in order to enable service continuity: i) client's mobility (predicted or expected), ii) overload situations in EAS or edge data network in general, and iii) maintenance of EAS. There are three roles that elements in edge data network can have during application-context relocation procedure: i) detection entity, detects a probable need for application state relocation by monitoring various aspects (e.g., UE location), ii) decision entity, determines whether relocation is needed, and if yes, instructs execution entity to perform relocation, and iii) execution entity, which executes context relocation.