5G smarT mObility, media and e-health for toURists and citizenS

Deliverable D3.3
Technologies, architecture and deployment advanced progress

Project Details

Call	H2020-ICT-19-2019
Type of Action	RIA
Project start date	01/06/2019
Duration	36 months
GA No	856950

Deliverable Details

Deliverable WP:	WP3
Deliverable Task:	Task T3.1 – T3.5
Deliverable Identifier:	5G-TOURS_D3.3
Deliverable Title:	Technologies, architecture and deployment advanced progress
Editor(s):	Cédric Thiénot (EXP)
Author(s):	Marco Gramaglia, Alberto Garcia-Martinez, Ginés Garcia-Aviles, Jesus Perez-Valero (UC3M), Linas Maknavicius, Bessem Sayadi, Stéphane Betgé-Brezetz (NBLF), Silvia Provvedi, Mara Piccinino, Giancarlo Sacco (ERI-IT), Sofiane Imadali (ORA-FR), Eleni Rigani, George Mitropoulos (NOKIA-GR), Ignacio Labrador (ATOS), Dorota Inkielman, Iwona Wojdan, Zbigniew Koper (ORA-PL), Carlos Barjau, Álvaro Ibáñez, Aarón Montilla (UPV), Ioannis Belikaidis, Vera Stavroulaki, Aimilia Bantouna, Ioannis Dimitriadis (WINGS), Christophe Burdinat, Cédric Thiénot (EXP), Pietro Scalzo, Andrea Buldorini, Giorgio Calochira (TIM), Velissarios Gezerlis (OTE)
Reviewer(s):	Linas Maknavicius (NBLF), Giorgio Calochira (TIM), Fons de Lange (PRE), Nikolaos Papagiannopoulos, Vania Giasla (AIA), Silvia Provvedi (ERI-IT), Belkacem Mouhouche (SRUK), Albert Banchs (UC3M), Nelly (Eleni) Giannopoulou (WINGS)
Contractual Date of Delivery:	30/04/2021
Submission Date:	29/04/2021
Dissemination Level:	PU
Status:	Final
Version:	1.0
File Name:	5G-TOURS_D3.3 Technologies, architecture and deployment advanced progress_v1.0.docx

Deliverable History

Version	Date	Modification
V1.0	*29/04/2021*	*Initial version, submitted to EC through SyGMa.*

Table of Content

List of Acronyms and Abbreviations

3GPP	3rd Generation Partnership Project
5GPPP	5G Public-Private Partnership
4G	4th Generation mobile network
5G	5th Generation mobile network
5GC	5G Core
5G EVE	5G European Validation platform for Extensive trials
5G-Xcast	Broadcast and Multicast Communication Enablers for the Fifth Generation of Wireless Systems
5G-MoNArch	5G Mobile Network Architecture for diverse services, use cases, and applications in 5G and beyond
5GT	5G-Transformer
5GT-VS	5GT Vertical Slicer
ABR	Adaptive Bitrate
AI	Artificial Intelligence
AF	Application Function
AMF	Access and Mobility Function
AL-FEC	Application Layer Forward Error Correction
API	Application Programming Interface
AR	Augmented Reality
AAS	Advanced Antenna Systems
BBU	Baseband Unit
BM-SC	Broadcast Multicast Service Centre
BSCC	Broadcast Service and Control Centre
BSS	Business Support Systems
CMAF	Common Media Application Format
CN	Core Network
CNN	Convolutional Neural Network
COTS	Commercial Over The Shelf
DoS/DDoS	(Distributed) Denial of Service attack
DVB	Digital Video Broadcasting
E2E	End to end
EE	Execution Environment
eMBMS	Enhanced Multimedia Broadcast Multicast Services
ENI	Experimental Network Intelligence
EPC	Evolved Packet Core
ESMLC	Evolved Serving Mobile Location Centre
FEC	Forward Error Correction
FDT	File Description Table
GMLC	Gateway Mobile Location Centre
GPRS	Generic Packet Radio Service
GSM	Global System for Mobile communications
GUI	Graphical User Interface
HPHT	High Power High Tower
HSS	Home Subscriber Server
IaaS	Infrastructure as a Service
IoT	Internet of Things
IRP	Integration Reference Point
IWL	Interworking Layer
KPI	Key Performance Indicator
KQI	Key Quality Indicator
LPLT	Low Power Low Tower
LTE	Long Term Evolution
M-ABR	Multicast Adaptive Bitrate
MB-SMF	Multicast Broadcast Bitrate Session Management Function
MB-UPF	Multicast Broadcast User Plane Function
MBS	Multicast Broadcast Services
MBSF	Multicast Broadcast Service Function
MBSU	Multicast Broadcast Service User Plane
MANO	Management and Orchestration
MBB	Mobile Broadband
MDAF	Management Data Analytics Function
MEC	Mobile Edge Computing
mMTC	Massive MTC
MooD	Multicast on Demand
MAE	Mean Absolute Error

ML	Machine Learning
MME	Mobility Management Entity
MTC	Machine Type Communication
NEF	Network Exposure Function
NS	Network Service
NSD	Network Service Descriptor
NFV	Network Function Virtualisation
NFVI	NFV infrastructure
NFV-O	NFV Orchestrator
NR	New Radio
NRF	Network Repository Function
NSA	Non-Standalone
NSO	Network Service Orchestrator
NSSF	Network Slice Selection Function
NWDAF	Network Data Analytics Function
PCRF	Policy and Charging Rules Function
PM	Performance Monitoring
PSO	Particle Swarm Optimisation
O-DU	O-RAN Distributed Unit
O-RAN	Operator Defined Next Generation RAN Architecture and Interfaces
O-RU	O-RAN Radio Unit
ONAP	Open Network Automation Platform
OSM	Open Source MANO
OSS	Operation Support System
OVS	OpenVSwitch
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RNN	Recurrent Neural Network
SA	Standalone
SA	System Architecture
SDN	Software Defined Networking
SDO	Standards Developing Organisation
SDR	Software Defined Radio
SFP	Small Form-factor Pluggable
SLA	Service Level Agreement
SMF	Session Management Function
SPGW	Serving Packet Data Network Gateway
TOI	Transport Object Identifier
TR	Technical report
TS	Technical Specification
UC	Use Case
UDM	Unified Data Management
UMTS	Universal Mobile Telecommunications System
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications
VDU	Virtual Deployment Unit
vEPC	Virtual EPC
VIM	Virtualized Infrastructure Manager
VM	Virtual Machine
VNF	Virtual Network Function
VNFM	Virtual Network Function Manager
VR	Virtual Reality
WEF	Wireless Edge Factory

List of Figures

List of Tables

Executive Summary

This report provides the third update out of four planned for the 5G-TOURS project presenting overall work performed in WP3 (network architecture and deployment). The scope of this deliverable is to report the progress made in the 5G-TOURS network architecture for the 5G network deployment at an advanced stage within the three trial sites (Turin, Rennes, and Athens).

This report focuses on the following aspects:

- An update of the overall project network architecture based on 5G EVE, relevant standards components and 5G-TOURS innovations, as well as architecture instantiation per trial site;
- The status of the network deployment on each trial site;
- The technology evolutions conceived in the 5G- TOURS project to address the use cases and their implementation status.

This document provides substantial updates in all the three areas, as discussed in the following.

An update of the 5G Network Architecture is provided, resulting from the continuous effort devoted to the refinement of the network architecture. In particular, we consolidated the architectural view by:

- describing the requirements coming from the NEST templates generated by WP2,
- introducing the different technology evolutions and
- defining the insertion strategies.

Concerning network deployment: as the project transitions from phase 1 deployment to phase 2 deployment we report the new infrastructure that has been installed in the different sites. Most notably we remark, a better integration of the 5G-TOURS project into the 5G EVE platform.

Technology evolution highlights 5G-TOURS work focused on a number of specific evolutions for the 5G network technologies: Enhanced MANO, Big Data and AI Orchestration, Broadcast and Service Layer. In this document, we present the instantiation of the novel technologies in the different sites, discussing how some of them are necessary enablers for many of the use cases that will be presented by the project. More specifically, 5G-TOURS provides solution for the Enhanced MANO by using also AI elements as well as a novel Broadcast Solution. Finally, all of them are made available to the verticals through the Service Layer.

More specifically, 5G-TOURS developed solutions for the integration of MANO with specific AI algorithms, In this document, we mostly deal with two aspects that are related to:

- the needed extensions of the baseline MANO architecture and implementation towards the integration of AI (e.g., the extension needed to support AI agents) and
- the design and evaluation of specific AI algorithms into this architecture.

The Service Layer, instead, is detailed into many flavours that specifically tackle an area of the vertical to network interactions. Among others, we specify implementation of the service layer for the direct interactions between the vertical and the AI governing the system (as also demonstrated in the ETSI ENI PoC developed by the 5G-TOURS partners and we cover the multicast/broadcast support in two axes: the development of the new network functions being specified at 3GPP Release 17 and the optimization of the transport delivery stack.

Future work will be focusing on the finalization of the architecture, especially for the insertion over 5G EVE architecture, both functional and physical, with the deployment of the final infrastructure in each site. This will include also the integrations of the technology evolutions towards the final demonstrations of the UCs.

1 Introduction

The 5G-TOURS project goal is to demonstrate the benefits of 5G technology in a pre-commercial environment for real users, tourists, citizens and patients and the respective vertical players, by implementing 13 representative use cases in 3 different types of cities:

- Turin, the touristic city (5 use cases),
- Rennes, the safe city (4 use cases) and
- Athens, the mobility-efficient city (4 use cases).

The exceptional situation worldwide has highlighted the importance, role and the shortcomings of the telecommunications network ecosystem. It is undeniable that 5G is here and has an important role to play in the telco world, as a service itself and as a complementary wireless network for private business. The number of announced 5G devices continues to rise and has passed the 600 mark in March 2021; there are now 628 announced types of 5G devices, which is an increase of 21% over the beginning of 2021. The number of 5G devices commercially available has also grown rapidly, exceeding 400 types of devices for the first time [52]. 428 operators in 132 countries/territories are investing in 5G mobile or 5G FWA/home broadband networks [51].

In parallel to 5G-TOURS, 5G EVE also experienced progress during these months. While continuously upgrading their physical infrastructure, the most relevant change is the launching of the application portal to the public, which interfaces the 5G equipment distributed in the European nodes to the interested verticals, allowing them to design, deploy and monitor their own 5G experiments [53]. The developments of 5G EVE are being closely monitored by 5G-TOURS and their implementation requirements are being factored in the design philosophies for the 5G-TOURS innovations.

The first deliverable of WP3 D3.1 [2] described the initial design ideas for the architecture, technologies and implementation. The deliverable marked the first progress update on the aforementioned topics, alongside a section on 5G EVE on-boarding and technological integration strategies.

The second deliverable of WP3 D3.2 [3] illustrated the progress on the architecture (as was designed to accommodate all the novel technologies developed during the project to support and enhance the proposed use cases) and physical deployment in 5G-TOURS. A description of the capabilities, including use case on-boarding and 5G EVE integration was also covered, alongside the plans for future enhancement. The progress on the development of 5G-TOURS technologies were described. The status in standardization fora was covered as well as the implementation and integration of the subjacent NFs or algorithms into 5G-TOURS architecture.

This current deliverable D3.3 covers the progress to date of WP3 ongoing work.

Section 2 provides the latest update on the high level 5G-TOURS architecture and the interaction with the portal and the interworking layer of 5G EVE. This section details the work performed in Task 3.1 “Overall architecture and Security”. Subsection 2.1 reviews and updates the requirements for the logical architecture, based on the collaboration with WP2 (use cases design), and feedbacks from the network deployment with WP4 (Touristic city), WP5 (Safe city) and WP6 (Mobility-efficient city).

Section 3 details how the 5G-TOURS innovations are integrated with the 5G EVE infrastructure, corresponding to the progress made within the Task 3.5 “Assessment and expansion of the network deployment”. Some extension or entry points to integrate 5G-TOURS in 5G EVE platform have need identified and are described in this section.

Section 4 describes the progress and status of the network deployment for 5G-TOURS. For each of the trial sites, it indicates how the 5G-TOURS architecture is instantiated to implement the targeted use cases, and the physical infrastructure deployment status. The physical infrastructure of the three sites is described for the 2 phases of trials, the first phase depending on the current 5G EVE infrastructure capabilities and the second phase including the future 5G capabilities and 5G-TOURS innovations.

Finally, Section 5 gives the latest update by the 5G-TOURS technological innovations, their on-going implementation status and their promotion to the relevant Standardisation Organisation group and/or open source

projects such as ETSI ENI [61] and OSM [36]. Several major enhancements have been identified to achieve the use cases designed in WP2:

- Enhanced Management and Orchestration (MANO), focused on Open Source MANO (OSM) and service level assurance monitoring (see Section 5.2),
- Artificial Intelligence (AI) orchestration, providing intelligent solutions and algorithms developed inside the project for the network slice management (see Section 5.3),
- Broadcast support, detailing the inclusion of broadcast in the 5G-TOURS infrastructure and covering the progress made in 3GPP for broadcast communications (see Section 5.4),
- Service Layer description with the requirements, expectations and implementation methods (see Section 5.5).

2 Progress on 5G-TOURS Architecture

In this section we will revise the architecture definition provided in the previous WP3 deliverable (D3.2 [3]) in the light of the ongoing implementation activities that have been performed during the last months. As already discussed also in D4.2 [1], D5.2 [7] and D6.2 [8], the different sites are moving towards the second phase of the implementation, so the architecture had to pass an intermediate reality check with its instantiation in Turin, Rennes and Athens.

This will also be an input towards the technology integration and the insertion strategies, as thoroughly discussed in Section 3. Moreover, also other WPs further progressed in their work, in Section 2.1 we will re-iterate on the requirements set by WP2 through the usage of NEST Templates, which will be propaedeutic to fill the service blueprints as requested by the 5G EVE portal.

2.1 Requirements

When designing the 5G-TOURS overall architecture, among other inputs such as the state of the art, architecture proposals and the 5G EVE architecture, the requirements coming from Work Package 2 were also considered, not only in terms of pure KPIs provided by the network (which are more relevant to the discussion on the specific sites, see Section 3), but also from the functionality that the network has to provide in terms of management, configurations, procedures, etc. WP2 worked on defining the UC requirements in a standard form, and for that purpose, partners provided UC requirements definition using NEST [9] templates. In the following, some relevant parameters for the NEST for UC1a (presented in Table 1) are taken as reference and its implication on the Architectural design is discussed.

Table 1. Selection of items for the NEST template of UC1a.

Sub-UC	Attribute ID	Attribute	Sub-parameter	Allowed value	Relevant 5G-TOURS Technology
1a	1	Performance Monitoring	Monitoring	**per second,** per minute, per hour	Service Layer
1a	2	Performance Prediction	Availability	Throughput, Latency, **Service Success Rate**	AI Orchestration
			Monitoring	per second, **per minute**, per hour	Service Layer
1a	3	Root Cause Investigation	-	Active / Passive	Security (partially)
1a	4	Slice Quality of Service	Resource Type	Non GBR	Service Layer (for the monitoring)
			Priority Level	CQI 80	
			Packet Delay Budget	0.01	
			Packet Error Rate	10-6	
			Jitter	N/A	
			Maximum Packet Loss Rate	N/A	
1a	5	User Management Openness	-	Supported / Non Supported	Service Layer

We can see that 5G-TOURS innovations are able to tackle 3 areas of the NEST templates:

- **Service Layer:** which allows for an easy configuration of many of the parameters related to the network slice. While most of such functionality is already provided by the 5G EVE portal, some of it will need an extension (through the Service Layer). Especially, the continuous monitoring of the KPIs (per second), is a planned extension as discussed in Section 5.5.1.
- The **AI Orchestration** algorithms, as discussed in Section 5.3.1, allow for the seamless re-configuration of the system. However, as 5G-TOURS plans for a continuous closed-loop interaction between the vertical and the underlying system, we are not exposing the actual operation directly, but rather allow for a "knob"-alike configuration (i.e., only one parameter instead of the many needed by the AI algorithm, which are known by the operator) for the vertical.
- Finally, we may target the **Root Cause Analysis** feedback for the errors, especially in the security by design aspect.

While most of our solutions are going to be integrated as much as possible into the 5G EVE infrastructure, some of them are implemented in some ad-hoc flavours, as discussed for the Service Layer implementation, to fulfil the needs of the use cases.

2.2 Overall Architecture description

The 5G-TOURS architecture is structured as a set of layers, each of them targeting a specific domain of the network, as depicted in Figure 1 below. This is described in the next subsections.

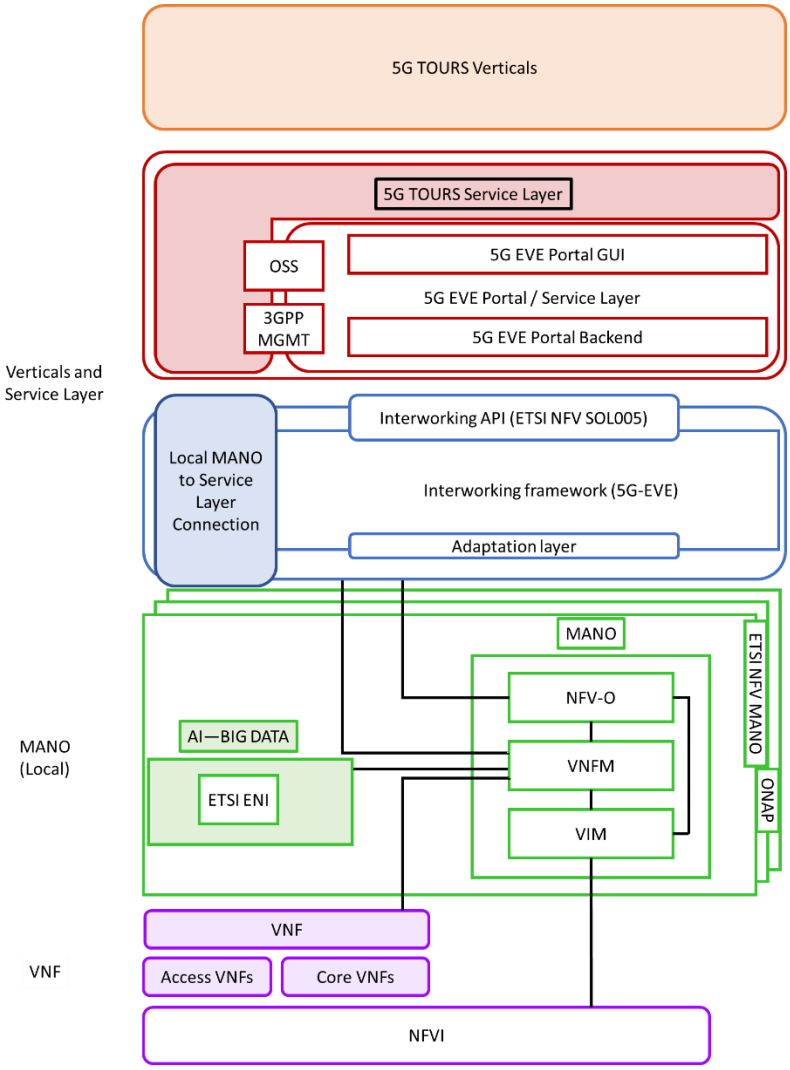


Figure 1. The Overall 5G-TOURS Architecture. Colour filled boxes represent native modules of the 5G-TOURS architecture, while non colour filled ones are inherited from 5G EVE.

2.2.1 Architecture description

The upmost layer is represented by the Verticals and their applications providing the different services envisioned in the project. For instance, the museum running Augmented Reality and Virtual Reality network functions to provide the enhanced museum visit service.

These functions are onboarded in the network through the 5G-TOURS service layer, which is an extendable interface that helps to ease the interaction between the verticals and the network operator. The extent of this layer depends on the specific use case, as also defined by the requirements set by WP2 (e.g., some of them require a continuous KPI feedback from the network). As such requirements were not fully fulfilled by the 5G EVE portal (which, to some extent, covered the functionality envisioned by the Service Layer, such as the network slice onboarding), the following actions were taken:

- Performed an exhaustive gap analysis (detailed in [3]). The performed gap analysis was conducted largely driven with the support of the vertical partners present in the 5G-TOURS consortium that gave feedback on the specific needs for the interface related to the specific use cases of 5G-TOURS.
- Designed extensions of the 5G EVE portal for some specific functionality (see Section 5.5).
- Leverage a direct connection between the vertical service layer and the local MANO to provide some enhanced functionality that cannot be provided without a full capability exposure between the Vertical

and Local MANO domains. This connection integrates the interworking layer of 5G EVE, which provides the glue among the centralized 5G EVE portal and the different sites. The rationale behind this direct connection is to enhance the Interworking Layer provided by 5G EVE (which was designed as open loop towards the vertical, which can onboard experiments and check the result upon their completion), providing new functionality such as continuous KPI monitoring or exposure of AI-related capabilities.

The local MANO is native for each site, which is implemented using different software releases by 5G EVE. For instance, the Turin site is relying on the Open Source Mano (OSM) [36] and the EVER Orchestrator [10] for the Core and Radio parts. The French side, instead, relies on the ONAP platform, while the Greek one employs the OSM Orchestrator as well.

Locally, we empower the MANO through the usage of Artificial Intelligence on both the Orchestration Platform itself (the AI agents included in the OSM platform, as discussed in Section 5.3) or through the application of the ETSI ENI concept [61] into the orchestration. These items allow to flexibly manage the network resources according to the loads, as also showcased by the different proof of concepts that have been proposed by the partners of 5G-TOURS (see Section 5.5.1 for the ETSI OSM PoC and Section 0 for the ETSI ENI PoC).

Finally, the lowest layer in the 5G-TOURS Architecture is represented by the NFV Infrastructure, the virtualized platform (e.g., the 5G EVE Infrastructure as a Service) and the Virtual Network Functions that provide the connectivity towards the terminals. These network function could either be virtual (as the core ones) or physical (as most of the base stations). Some network functions are directly provided by 5G EVE (for instance, the Core network functions in the Turin site), others are extended to support 5G-TOURS specific functionality (such as the probes deployed in the Athens site), while others are additions from 5G-TOURS to the 5G EVE infrastructure, as the radio deployment in the CHU Rennes Hospital, the Palazzo Madama museum in Turin, and the Athens Airport.

2.2.2 Architecture Instantiation

The instantiation of the architecture in the different test sites is done according to the two phases methodology discussed in [3]. The purpose of this split is to pipeline the activities in the network deployment and the use case application development. That is, by allowing an early deployment of the 5G Services in a rapid way, verticals could get a first impression of the achievable performance and react accordingly, while the full deployment continues for the phase 2 in which the fully fledged network will be developed, with the integration of the technical innovation discussed in the project.

In the following subsections, we describe this transition towards phase 2 discussing how such technologies are going to be integrated into the specific sites. Then, in Sections 3.1, 3.2 and 3.3, the implementation details of each site are discussed.

2.2.2.1 Turin site

The transitions of the Turin site (WP4) from phase 1 to phase 2 is depicted in Figure 2 below.

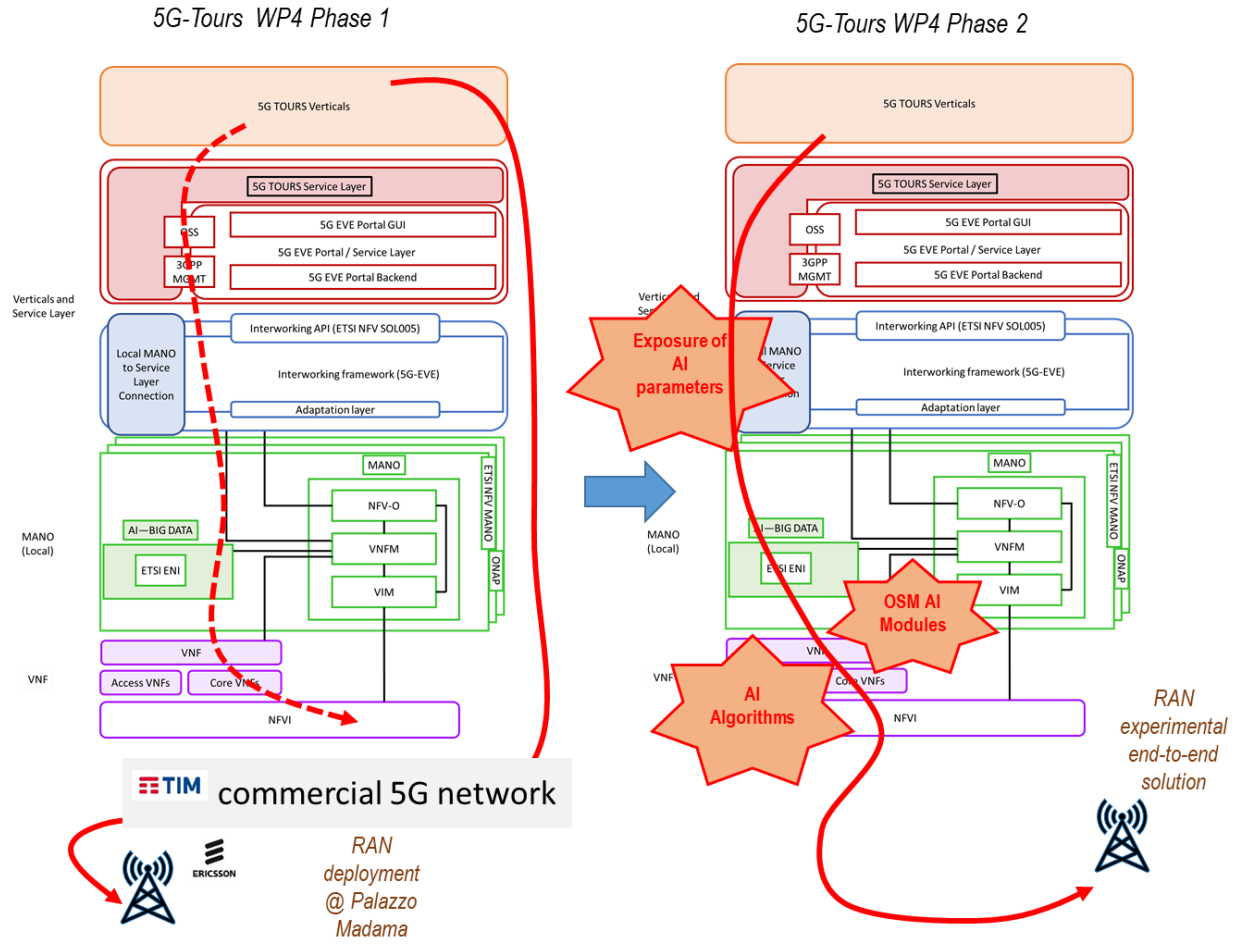


Figure 2. Phases outline in the Touristic City Site.

During phase 1, the 5G-TOURS partners focused mostly on providing the wireless coverage at Palazzo Madama with 5G NSA solutions, relying on the commercial 4G Core Network of TIM (solid red line). The obtained results are already in line with the expected performance required by the different implemented use cases (see Section 3 for more details). In addition, several enhancements will be integrated for phase 2 on the architecture solution described in Section 4.1.1.

- The integration of the AI agents innovation into the next release of the Open Source MANO installation in the 5G EVE site, this will allow for a flexible orchestration of the system, especially for the part related to the application VNFs.
- The introduction of specific AI algorithms in the network such as the ones already demonstrated in the ETSI ENI PoC (see in Section 0), or further implementation of AI algorithms such as the ones described in Section 5.3.
- The exposure of some AI configuration parameters to the vertical. While the AI algorithms deployed in the network take care of the low level details of the configuration of VNFs, a 'knob' to configure them is available to the verticals through the service layer (see in Section 0).

2.2.2.2 Rennes site

The transitions of the Rennes site from phase 1 to phase 2 is depicted in Figure 3 below.

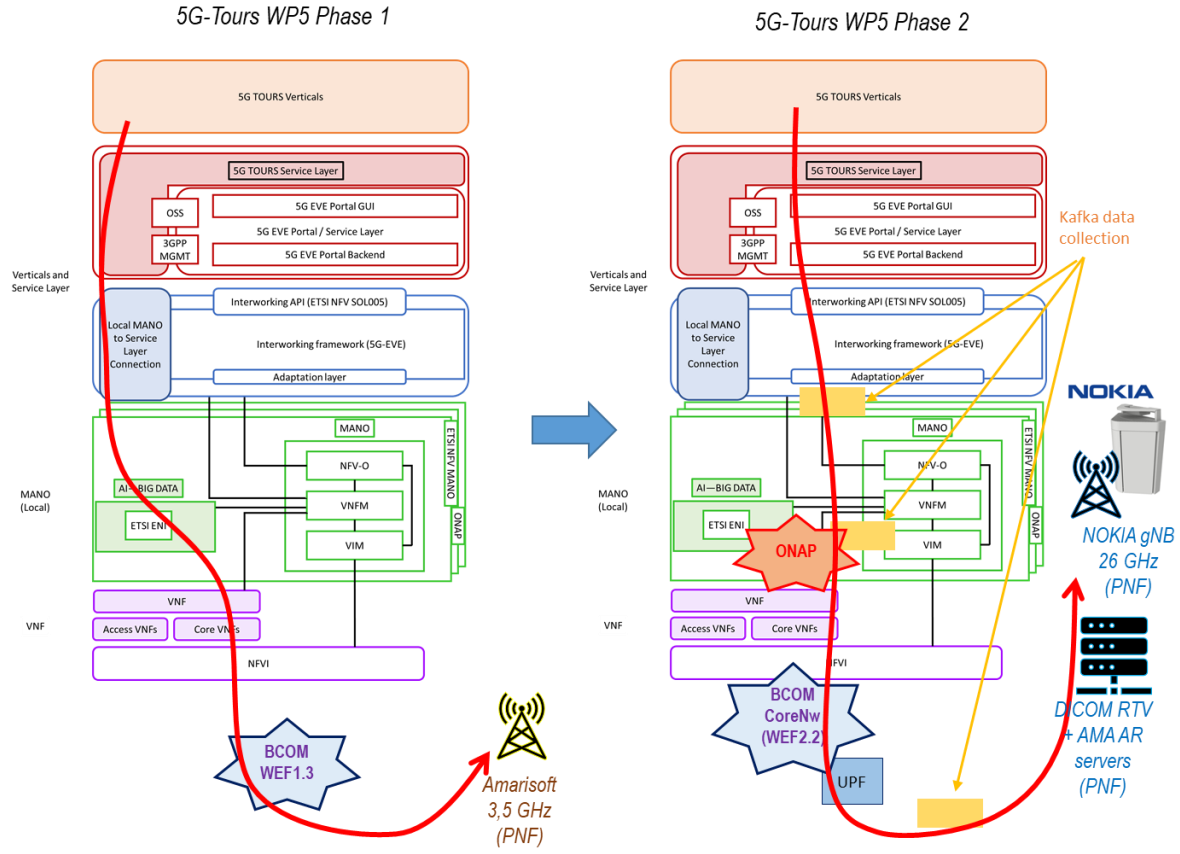


Figure 3. Phases outline in the Safe City Site.

Phase 1 in the Rennes site is provided through a small scale setup of the RAN and Core infrastructure. Specifically, the RAN infrastructure is provided through an Amarisoft eNB [62] implementation, connected to a BCOM WEF v1.3 Core Network. This setup allowed for an initial prototyping of the use cases, which will be further integrated for phase 2, as follows (more details in Section 4.2):

- The 5G new-radio RAN deployment will be over the 26 GHz band, to guarantee very high bitrates over a well-defined area (one of the first to be deployed in the context of 5G PPP projects).
- The core network (based on the BCOM WEF 2.2) will be orchestrated using ONAP on the 5G EVE infrastructure, in a remote location (the Orange Labs in Chatillon). On the other hand, edge functions will still be located closer to the real applications.
- Specific network metrics probes will be deployed in the network orchestrator, at many levels in the hierarchy, to enhance the amount of data that can be exposed to the vertical.

2.2.2.3 Athens site

The transitions of the Athens site from phase 1 to phase 2 is depicted in Figure 4 below.

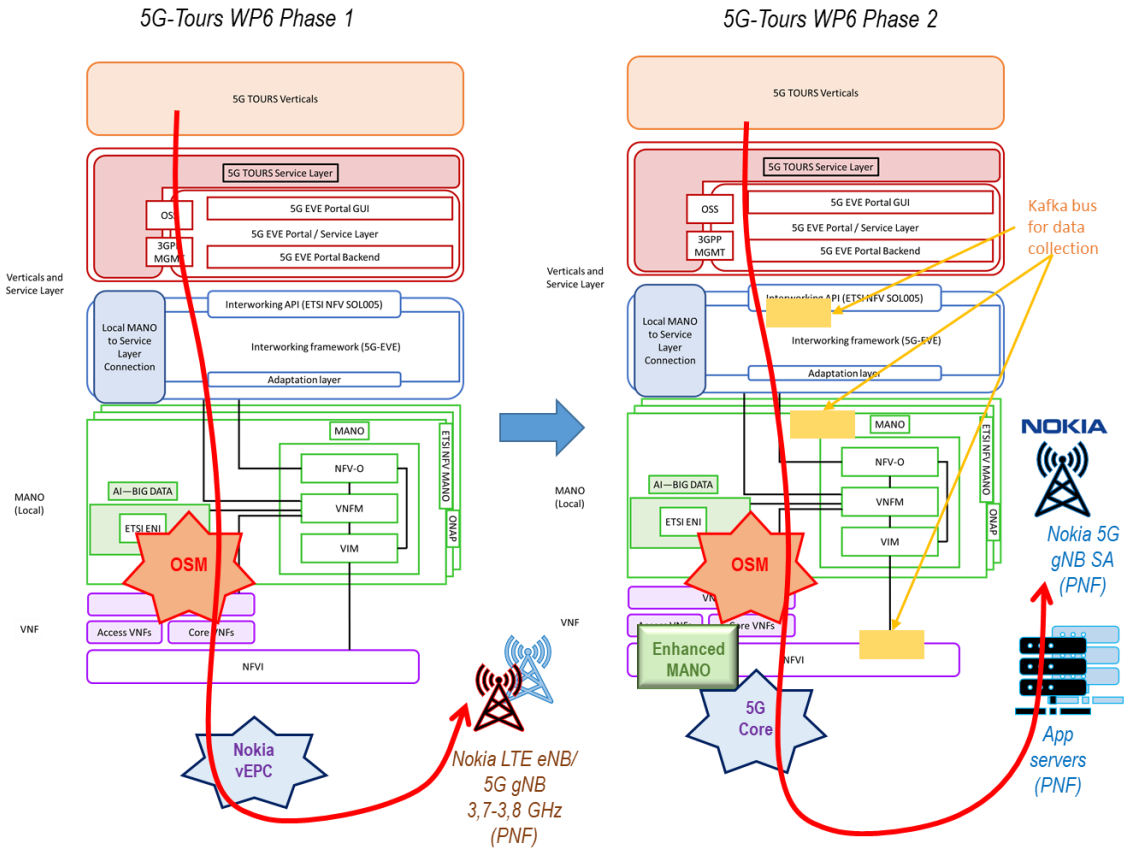


Figure 4. Phases outline in the Mobility-Efficient City Site.

Phase 1 in the Athens site is already providing 5G connectivity using NSA technology connected to an OSM-orchestrated vEPC provided by Nokia. The whole deployment is a mixed LTE, 5G-NR on the 3.5 - 3.6 GHz band, maximizing the coverage area. During phase 2 the following enhancements will be integrated in the network:

- A 5G NR SA solutions, relying on the 5GC implemented by Nokia (more details can be found in Section 4.3.1).
- The exposure, through the service layer, of specific network metrics gathered through the probes that will be deployed in the network (more details in Section 5.2).

2.2.3 Enhanced Exposure Functionality

The ongoing quest for the tight integration between network operation and network service provisioning initiated with the introduction of 5G often clashes with the capacity of current network architectures to provide means for such action. Owing to the traditional design of mobile networks, which barely required a tight interaction, network elements offer capabilities for their continuous optimization just within their domain (e.g., access, or core), allowing for a "silo-style" automation that falls short when aiming at closed-loop automation that embraces all the actors involved in the network, from network functions up to the service-provider network functions as. for instance, we do with the service layer to link vertical and the network.

5G networks introduce a giant leap forward compared to the rather monolithic structure of legacy mobile networks, which has been basically designed to provide mobile broadband services over one physical network instance. In order to manage such a possibly large number of network instances, Big Data and Artificial Intelligence (AI) techniques are considered by 5G-TOURS as potential enablers for autonomous management of the network. Features such as auto-scaling, self-optimization, or intent-based networking clearly fit well with a data-driven approach to the network environment, in which elements offer ways to produce and consume data,

but can also be configured according to the service-related policies in a jointly optimized way. Achieving autonomous network management is thus a challenging task that entails overcoming a number of technical complexities, such as data heterogeneity and temporal scale heterogeneity. In this context, having a flexible exposure of data among different network elements is paramount.

Traditionally, procedures related to network management, network orchestration, and network control have been developed by different tracks of the standardization activities. Hence, the functions that execute these procedures (i.e., OSS functions, element managers, orchestrators, or radio and core NFs) have been designed in a domain-specific and, in some cases, even proprietary manner, with possible optimizations happening only in a "per domain" way, leaving the interaction limited to peer-to-peer reference points within a domain, e.g., between control plane (CP) and user plane (UP).

In such reference-point-based setups, optimizations are either open-loop (i.e., no feedback among different modules in the system) or require very expensive human engineering procedures to close the loop. This approach has been deemed as valid within legacy networks, due to their limited number of possible configurations. However, in a 5G environment, this approach clearly falls short. Also, as legacy NFs either have function-specific data acquisition and processing procedures or no procedure at all, simple configurations or rulesets are usually sufficient to achieve optimization goals.

In 5G, network slicing, among others, has imposed a more modular design of NFs, allowing these NFs to be shared and re-used across slices in a more fine-grained and targeted manner. That is, a single Network Slice Subnet Instance (NSSI) and its constituent NFs may be used across several slices and services (e.g., common radio access NFs across slices). Thus, interfaces devoted to the data acquisition and processing from NFs or even to feed and push data to specific AI modules shall be designed. In order to "close the loop" in an automated manner, by adding AI and big data solutions, new interfaces and functionalities are needed:

- **Flexible data exchange across domains**. Different network domains shall be able to exchange information among them using, e.g., a publish-subscribe methodology.
- **Reliable and scalable configurations.** Besides producing and consuming the data, the NFs in all domains shall offer ways to allow authorized configuration of their relevant parameters, possibly with different levels of authorization, e.g., depending on the enforced resource provisioning scheme. For instance, some service providers may have full configuration capabilities while others may only have limited visibility of configurable parameters.

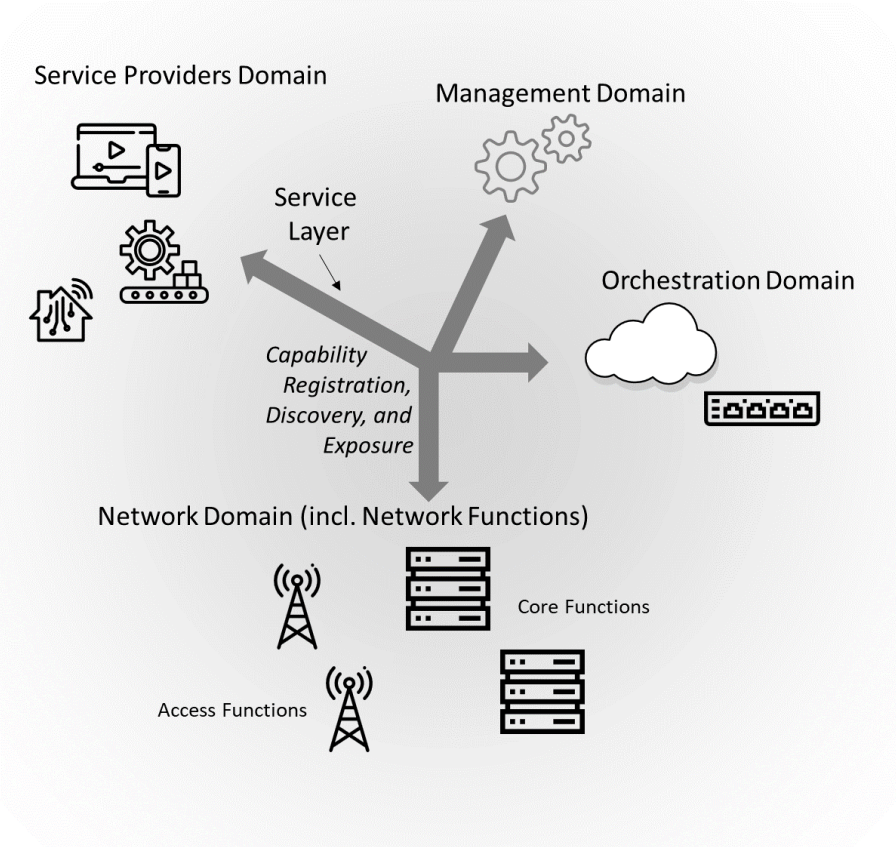


Figure 5. Different network domain in the current 5G Architecture.

As depicted in the above Figure 5, several domains can be producer and consumer of capabilities, which include, but are not limited to, the exposure of monitoring data, and can also include the capability of configuring specific parameters that are relevant in each domain. In 5G-TOURS we mostly focus on the exposure of these capabilities to verticals, as accomplished by the service layer, discussed in Section 5.5.

2.2.4 5G security-by-design for verticals

In this section, a so-called "security-by-design" approach to 5G security is explained, security improvements from 4G to 5G are briefly presented, and a threat analysis of Mobile Edge Computing (MEC) vulnerabilities is given, since several 5G-TOURS use cases intend to use MEC facilities to some extent. In the next (and final) deliverable of 5G-TOURS architecture and deployment, further analysis on security implications for 5G verticals will be performed, especially concerning:

- how to combine inherent 3GPP security mechanisms with specific tenant constraints (e.g. safe-city/eHealth);
- how to devise means to express 5G security requirements for the use of the verticals and to translate them into automatic 5G infrastructure configuration (e.g., by extending/adapting of network service descriptors/blueprints/slice templates, as defined by ETSI NFV ISG or GSMA).

Wireless communication is inherently vulnerable and needs specific protection against interception and tampering. Consequently, ever since GSM, the second generation of mobile networks, encryption has been used on the radio interface to secure the user communication. In the following two generations of mobile networks, UMTS and LTE, respectively, the security architecture was significantly enhanced. Besides encryption of user traffic, these networks have also provided mutual authentication between mobile terminals and the network, as well as integrity protection and encryption for all control and management traffic. Overall, UMTS and LTE security features ensure not only high level of security and privacy for subscribers, but, very importantly, also assure the resilience required to combat various forms of attacks against the integrity and availability of the services these networks provide. This raises the question: Are new security concepts required for the next mobile network

generation? The answer is yes. On the one hand, the support of a variety of new use cases and, on the other, the adoption of new networking paradigms have made it necessary to reconsider some current elements in the approach to security. Figure 6 visualizes the main drivers for 5G security.

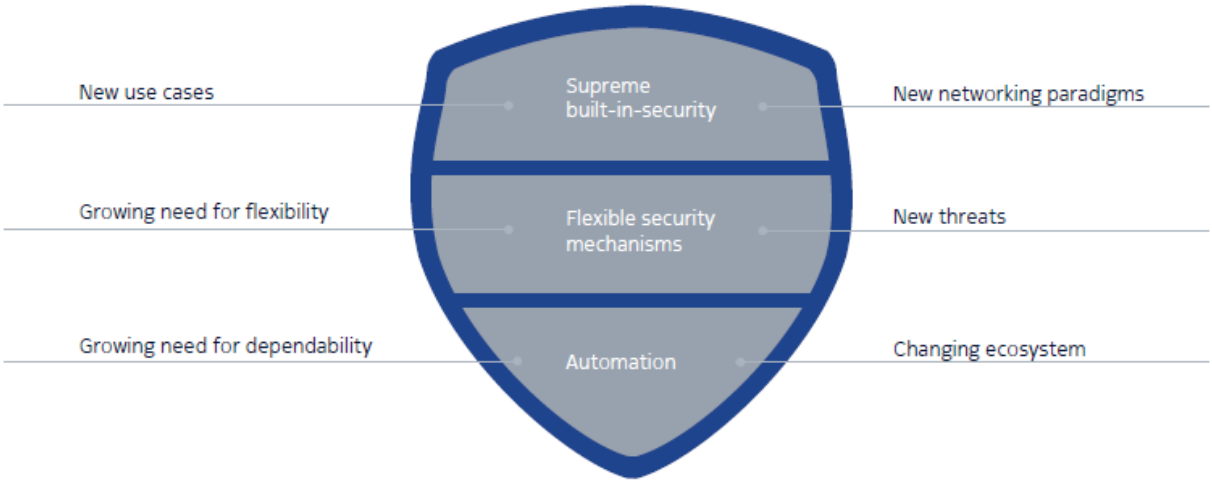


Figure 6. 5G security drivers.

While LTE was designed primarily to support the mobile broadband use case (i.e., broadband access to the Internet), 5G targets a variety of additional use cases with a variety of specific requirements. These cases include support of an enormous density of mobile devices or the need for ultra-low latency in the user communication. Use cases, such as vehicular traffic control or industry control, place the highest demands on the dependability of the network. Indeed, human safety and even human lives depend on the availability and integrity of the network service.

To support each use case in an optimal way, security concepts will also need to be more flexible. For example, security mechanisms used for ultra-low latency, mission-critical applications may not be suitable in massive Internet of Things (IoT) deployments where mobile devices are inexpensive sensors that have a very limited energy budget and transmit data only occasionally.

To efficiently support the new levels of performance and flexibility required for 5G networks, it is understood that new networking paradigms must be adopted, such as Network Functions Virtualization (NFV) and Software Defined Networking (SDN). At the same time, though, these new techniques also bring new threats. For example, when applying NFV, the integrity of virtual network functions (VNFs) and the confidentiality of their data may depend to a larger degree on the isolation properties of a hypervisor. More generally, they will also depend on the whole cloud software stack. Vulnerabilities in such software components have surfaced in the past quite often. In fact, it remains a major challenge to provide a fully dependable, secure NFV environment. SDN, for its part, bears the threat that control applications may wreak havoc on a large scale by erroneously or maliciously interacting with a central network controller.

Another driver for 5G security is the changing ecosystem. LTE networks are dominated by large monolithic deployments—each controlled by a single network operator that owns the network infrastructure while also providing all network services. By contrast, 5G networks may see a number of specialized stakeholders providing end-user 5G network services, as illustrated in Figure 7.

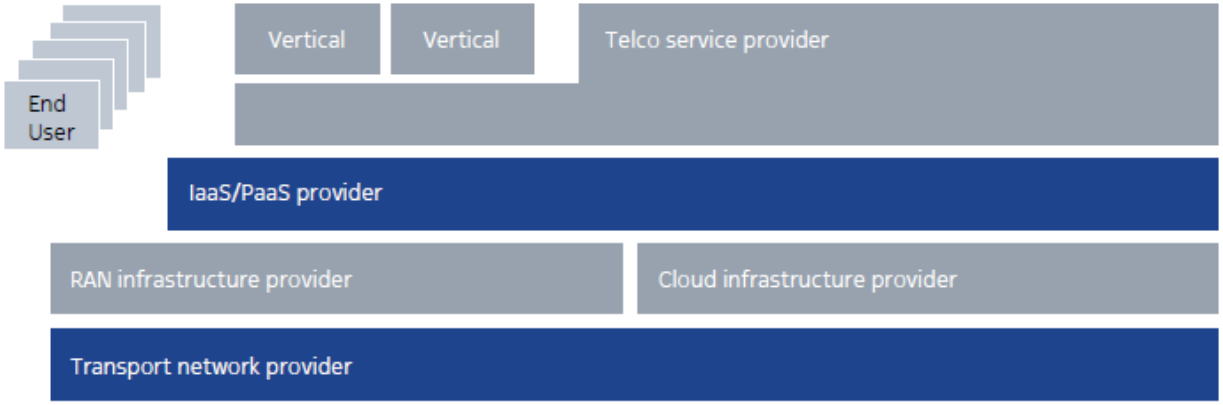


Figure 7. Example of multiple stakeholders involved in providing end-user 5G.

In particular, there may be dedicated infrastructure providers decoupled from telco service providers that host several service providers as tenants on a shared infrastructure. Cloud as the infrastructural choice on its own brings a new set of important 5G security considerations and dilemmas to be solved, such as whether to build/utilize a private cloud (IaaS/PaaS) infrastructure or to make use of external could service providers, how to ensure secure communication in cloud, how to leverage cloud high availability and resilience etc.

In another case, telco service providers may offer not only end-user communication services, but also provide complete virtual networks or "network slices" specialized for specific applications, such as IoT applications. These may be operated by verticals. For example, a manufacturing company could run a virtual mobile network specialized for industry control applications for its own plants. The relevant security issue here is the building and maintenance of new trust relationships among all stakeholders. The aim would be to ensure a trusted and trouble-free interaction resulting in secure end-user services.

Obviously 5G networks must support a very high level of security and privacy for their users (not restricted to humans) and their traffic. At the same time, networks must be highly resistant to all kinds of cyber-attacks. To address this two-fold challenge, security cannot be regarded as an add-on only; instead, security must be considered as part of the overall architecture and **built into the architecture right from the start *("security-by-design")***. Based on a secure architecture, secure network function implementations are also essential in order to ensure a high security network. Security assurance methods are therefore essential so that operators can ensure the required security level for different network functions.

5G security must be flexible. Instead of a one-size-fits-all approach, the security setup must optimally support each application. This entails the use of individual virtual networks or network slices for individual applications, as well as the adjustment of the security configuration per network slice. Security features subject to this flexibility may comprise the mechanisms for identifying and authenticating mobile devices and/or their subscriptions, or for determining the way that user traffic is protected. For example, some applications may rely on security mechanisms offered by the network. These applications may require not only encryption, as in LTE, but also user plane integrity protection. However, other applications may use end-to-end security on the application layer. They may opt out of network-terminated, user-plane security because it does not provide additional security.

2.2.4.1 Security improvements from 4G to 5G

Implementing the security architecture for mobile network functions as standardized by 3GPP is an essential pillar for building highly secure and reliable 5G networks. 3GPP TS 33.501 is the key document providing a detailed description of 'security architecture and procedures for 5G system' [63]. The specification defines a model of a security architecture, consisting of six security domains, as depicted in Figure 8:

- **Network access security (I)** – security features that enable a user terminal to authenticate and access the network by providing protection on the radio interfaces.
- **Network domain security (II) -** security features that enable network nodes to exchange signalling and user data securely.

- **User domain security (III)** - security features that enable the secure user access to mobile devices.
- **Application domain security (IV)** - security features that enable user and provider domain applications to exchange messages securely. 33.501 specifications do not cover application domain security.
- **Service Based Architecture (SBA) domain security (V)** - a new set of security features that enable network functions of the SBA to communicate securely within serving and other network domains.
- **Visibility and configurability of security (VI)** - security features that enable the user to be informed regarding which security features are in operation or not.

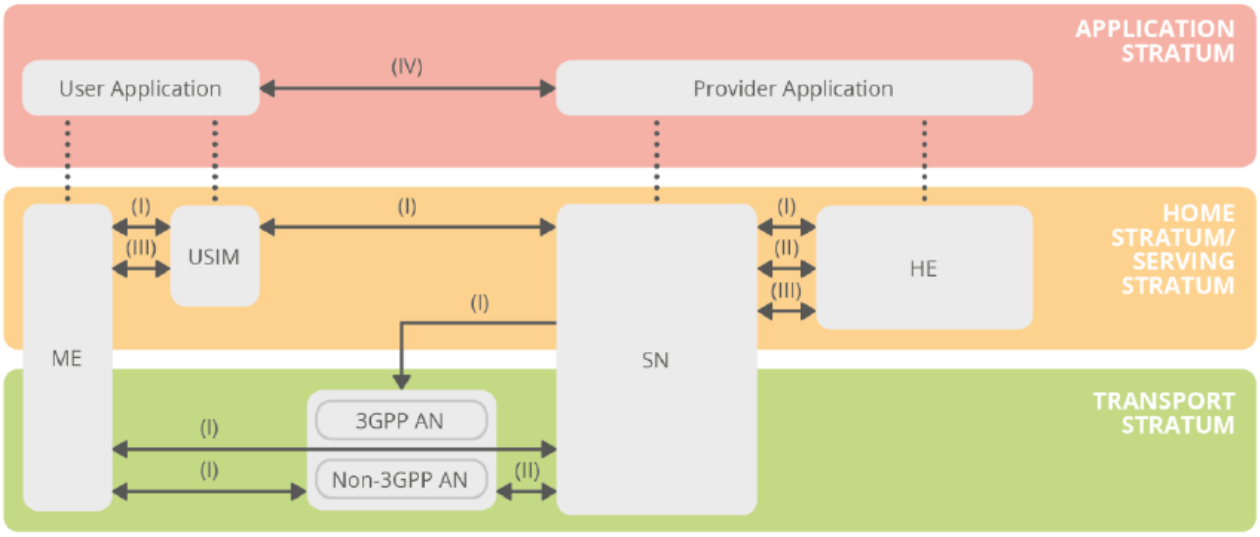


Figure 8. Security architecture model as defined in TS 33.501 [63].
(acronyms used are: ME=Mobile Equipment, SE=Serving Network, HE=Home environment)

5G builds on the well-proven security concepts of 4G, such as maintaining separated security associations for the access stratum and the non-access stratum, using a choice of well scrutinized crypto algorithms, using temporary identifiers to protect user location privacy and so on. However, 5G also introduces some significant enhancements and improvements.

At a glance, the new 5G security features are [69]:

- New access-agnostic authentication framework with improved home network control for roaming
- Enhanced subscription privacy (protection against "IMSI-catching")
- User plane integrity protection
- EAP-based "secondary authentication"
- Security for service-based interfaces
- Enhancements for interconnection security

Table 2 provides an extended overview over the new 5G mechanisms and a comparison with the respective 4G mechanisms [69]. At the same time, however, some of these security controls are defined as optional or there is a degree of flexibility left to suppliers on how to implement and for operators on how to interpret and utilise the controls.

Table 2. 3GPP Security Features – 5G versus 4G.

4G (LTE) Security	5G Security
UE is authenticated by two different authentication methods depending on the access network type (EPS AKA on LTE access and EAP AKA'on Wi-Fi access).	**UE is authenticated using either 5G AKA or EAP AKA'**, irrespective of access type (access agnostic). The home network decides on the method.
A serving network can fraudulently request authentication info for a UE from the UE's home network, even if the UE is not at all roaming in the serving network. This info can then be abused.	**Improved home network control**: When a serving network authenticates a roaming UE, the home network gets a proof that the UE is indeed present in the serving network.
The UE subscription identifier (IMSI) is in some cases transmitted as plain text without encryption over the air; fake base stations can force UEs to reveal the IMSI ("IMSI-catching").	The Permanent Subscription Identifier (SUPI) is not sent in clear over the air in any network procedures; instead, the **Subscription Concealed Identifier (SUCI)** is used, which is an encrypted form of the SUPI.
No integrity protection of user plane traffic; this allows certain attacks although the data are encrypted.	**Integrity protection of user plane traffic** is mandatory to support by the UE and the network (optional to use).
The network provides very restricted support for authentication of UEs to connected packet data networks ("PCO-based authentication").	**EAP-based "secondary authentication"** provides a flexible and strong concept for authentication between UEs and connected data networks.
Packet core network control signaling uses Diameter, and IPsec protection is mostly not applied for it.	The **5G Service Based Architecture (SBA) uses HTTP/2 with IETF Transport Layer Security (TLS)**, and the OAuth 2.0 framework to authorize access to restful APIs.
No flexible security for interconnection of different PLMNs via an IP Exchange (IPX) network. Use of E2E IPsec tunnels is specified, but this is in conflict with the operational requirements of an IPX, so it is mostly not used.	The **new function SEPP (Security Edge Protection Proxy)** protects the edge of the network and provides flexible interconnection security. It allows to protect selectively sensitive information while making other information visible to entities in the interconnection network, as required.

2.2.4.2 MEC threat analysis

Mobile Edge Computing (a.k.a. Multi-access-Edge Computing in the ETSI MEC Industry Specification Group) [64] complements the 5G architecture and allows applications to be executed close to the Radio Access Network (RAN) and in proximity of the User Equipment (UE). This is critical for applications requiring ultra-low latency. While latency to a core network (executed in a regional data center) is about 20ms to 40ms, latency for an edge node is less than 10ms (and even down to 200µs for a far edge). Moreover, as UEs such as IoT have very limited computation and processing capacities, the execution of many tasks needs to be performed in an infrastructure having much powerful resources. For this purpose, it is required to locally host both the data and the compute-intensive processing. MEC is also applicable for numerous and various markets (e.g., eHealth, smart-cities, industry 4.0, transport, connected & autonomous cars) and it allows tremendous business opportunities [65], [66].

Security is of course a key challenge regarding MEC network. 5G network security has been widely investigated [67-70] and is of course applicable to MEC. Security of public MEC (e.g., MEC belonging to a telco) is expected to be more secure as it benefits from the operator security infrastructure. However, private MEC (e.g., for a hospital, an airport) will be deployed within the enterprise infrastructure which are usually not full secure and can therefore be more vulnerable to new attacks based on the MEC.

The following Figure 9 depicts a simplified view of the architecture as defined in the ETSI- MEC framework [71]. MEC architecture has two main levels: the *Mobile Edge Host Level* and the *Mobile Edge System Level*. The Mobile Edge Host Level mainly contains the local Apps executed upon a virtualized infrastructure; the Mobile Edge System Level contains the orchestration and life-cycle functions handling the apps over the different MEC hosts, as well as some other OSS and customer facing functions.

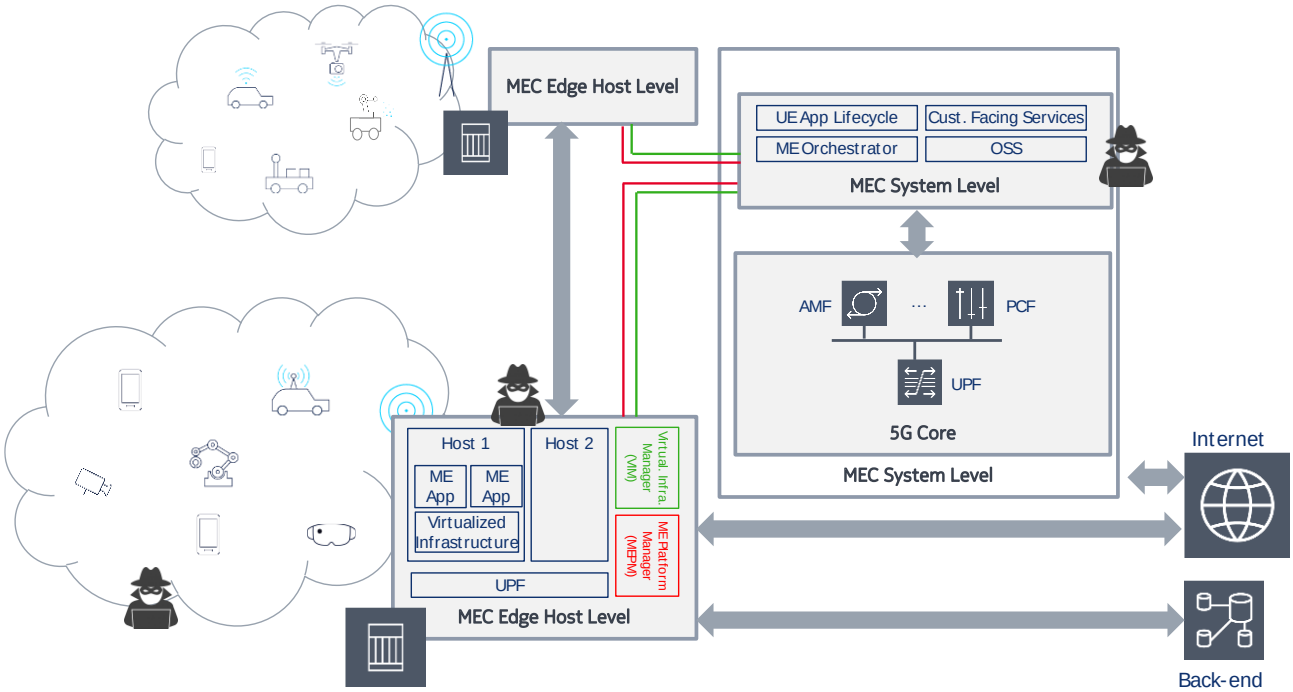


Figure 9. Network scope for vulnerability analysis.

Figure 9 also presents the areas of the main vulnerabilities [72]:

1. Vulnerabilities of the local radio MEC network.
2. Vulnerabilities in the MEC host Edge cloud (including between MEC Hosts).
3. Vulnerabilities in the MEC System Level.

1. Vulnerabilities of the local radio MEC network

It is well known that numerous UEs are highly vulnerable (e.g., IoT such as sensors, robots, AR/VR glasses) and are subject to various attacks (.e.g., physical tampering, malicious code injection, hardware trojans). A compromised UE can threaten the MEC system in various ways. An attacker can:

- Convey malicious content from a tampered UE to the MEC host [73-78].
- Deplete the resources of the UE, disrupting the normal execution the Applications/Services running in the MEC [75].
- Manipulate the data volume sent to the MEC applications ("offloading tampering") to allocate more resources and degrade the performances of the application or even the MEC itself.
- Use the potential huge numbers of IoT to perform critical large DDoS attacks on the applications.
- Eavesdrop the offloading channels from UEs may not be secure since computation tasks can be over-heard by malicious eavesdroppers. Traffic encryption between the UE (that may have limited resources) and the MEC can increase the propagation delay as well as the execution delay, thus reducing the application performance in a level not acceptable for applications.

2. Vulnerabilities in the MEC host Edge cloud (including between MEC Hosts)

Beyond the MEC applications providing by the enterprise holding the MEC, various applications from third parties or telco operators can also run on the MEC platform. This is particularly the case for verticals as they have often their independent ecosystem and developer community.

MEC network and platform API can then be a source of attacks notably unauthorized access and privilege escalation, sensitive data leakage, malicious use of the MEC NFV functions [77]. For some vertical market

(e.g., entertainment, smart cities) applications can be downloaded by consumer end-user from a marketplace. While these applications will run in the UE, it can be necessary (e.g., for time-sensitive and/or high bandwidth services) that the server part of the application will run locally in the MEC. This can be an attack vector to the MEC, notably if the user downloads the application from an official marketplace.

Large scale MEC deployment (e.g., smart cities) can be severely affected by the tampering of just one MEC. The possible heterogeneity and diversity of MEC environment also increase vulnerabilities. The different MEC applications may not have the same image version, or all updated with all the required patches. An attacker can compromise an unpatched MEC to later compromise the other MEC having the same vulnerability.

For some use-cases, the device density in MEC network could be up to 1 million devices for every square kilometre. Managing device credentials and detect their possible compromising (for credential revocation) will request scaling capabilities that have not previously been uncounted.

New attack vectors can also compromise privacy (e.g., location) and could constitute a major threat for privacy sensitive application and regarding regulation policy [78], [79].

For some use-cases (e.g., car traffic control and autonomous driving), MECs can be located along a path (e.g., highway) and an application can migrate in real-time from one MEC to another. If compromised, the application can tamper the other crossed MECs. Besides, crossing borders can induce data regulation policy violation [80].

3. Vulnerabilities in the MEC System Level

Attacks can also be performed on the link between the MEC and the core (backhaul link) [81]. MEC hosts support a much smaller traffic compared to the centralized core and are then more exposed to DoS attacks.

Moreover, the backhaul link transports critical data from the core or other servers accessed through internet. This is for instance the case of sensitive data exchanged between a MEC and a central office (e.g., company headquarter) which are exposed to attacks as eavesdropping. Communication between MECs (e.g., for automotive) also extend the possible attacks.

This link is also critical for the execution of applications as it allows migrating VM, offloading tasks that required high processing resource, transporting of critical network control information (MEC network control), transporting statistics and service logs information (from the MEC network and applications).

3 Technology Integration into 5G EVE & Insertion strategies

5G EVE and 5G-TOURS projects are working closely to guarantee the success of on-boarding the different 5G-TOURS vertical use cases. For that the hardware and the software capabilities of the 5G EVE infrastructure is monitored and assessed and, if necessary, expanded to be able to support the 5G-TOURS use cases. For some use cases, actions include possible coverage expansions.

In D3.2 [3], it has been agreed to follow the proposed methodology depicted in Figure 10:

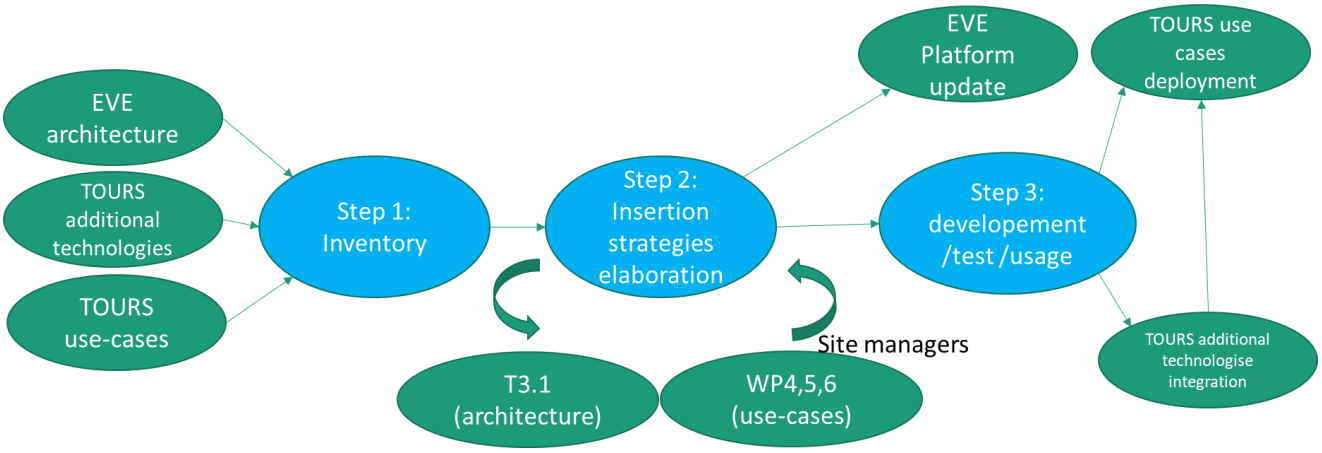


Figure 10. Insertion Methodology.

The methodology is composed by three steps: inventory, insertion strategies and development.

During the first period of the project, we implemented step 1 and step 2 from the above strategy. An inventory of the existing hardware (RAN frequencies, capabilities of the computing servers, etc.) and software (the list of VNFs including the vertical's applications) was made per use case. This helped us to fine tune how the use cases could be onboarded on 5G-EVE platform. In many cases, this onboarding requires some specific upgrade of 5G EVE infrastructure fixed by following the insertion methodology. The step 3 deals with the development, test and usage of the use-case.

We already identified the possible extension or entry points to integrate 5G-TOURS in the 5G EVE platform. They are depicted in Figure 11 and highlighted by the red pictogram. We can:

- inject new experiment blueprint in the 5G EVE Portal;
- upload new VNF to the catalogue to build the experiment blueprint in the 5G EVE Portal;
- enhance some Locations using a different frequency band;
- extend the radio coverage by adding new locations.

For more clarity, 5G-TOURS has three sites: Rennes, Athens and Turin. Each Site has different locations like airport, hospital etc.

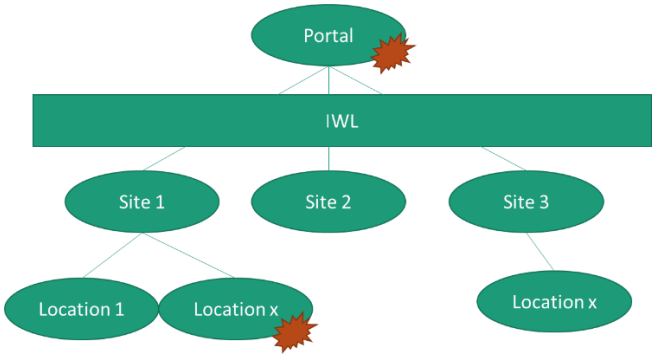


Figure 11. Possible insertion points of 5G-TOURS on 5G EVE platform.

In the following sections, we present the activities of each site.

3.1 Turin site

The 5G-TOURS use cases required a coverage extension of 5G EVE in two Locations (Palazzo Madama and GAM). The details on the current and foreseen architecture solution are reported in section 4.1.1.

The UC implementation will be verified in the field relying on network solutions that provide indoor and outdoor radio coverage, and which are connected to a commercial core network. The architecture instantiation for those verification activities is conducted using the network infrastructure described in Section 4.1.2.

5G-TOURS evolutions will exploit the end-to-end 5G NSA network solution provided by 5G EVE at the TIM laboratory.

Figure 12 shows 5G-TOURS functionalities for UC1.

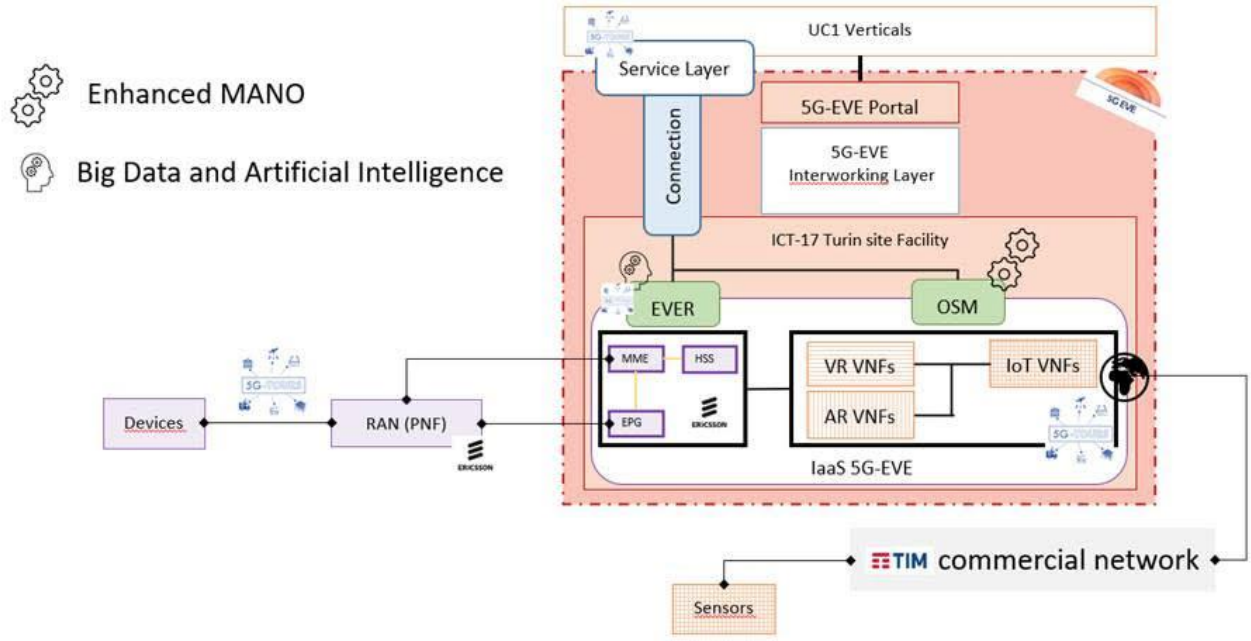


Figure 12. UC1 architecture instantiation.

3.2 Rennes site

The Rennes site hosts UC 7 and 8 which are part of the Safe City cluster in 5G-TOURS. The consortium worked on their integration in 5G EVE covering two potential insertion points: a new location in CHU Rennes, and usage of 5G EVE blueprints to instantiate UC7 and 8 through 5G EVE portal. The integration of 5G-TOURS with 5G EVE is achieved as depicted in Figure 13.

As it is depicted in Figure 13, the BCOM parking will be covered by a mm-wave 5G RAN node connected to BCOM CORE. A new location also will be deployed in CHU Rennes for UC8 to cover the surgery room. Both sites will host an EDGE node to deploy the UPF data-plane function.

The CORE will be instantiated dynamically via the 5G EVE Portal. This activity with 5G EVE consortium is ongoing.

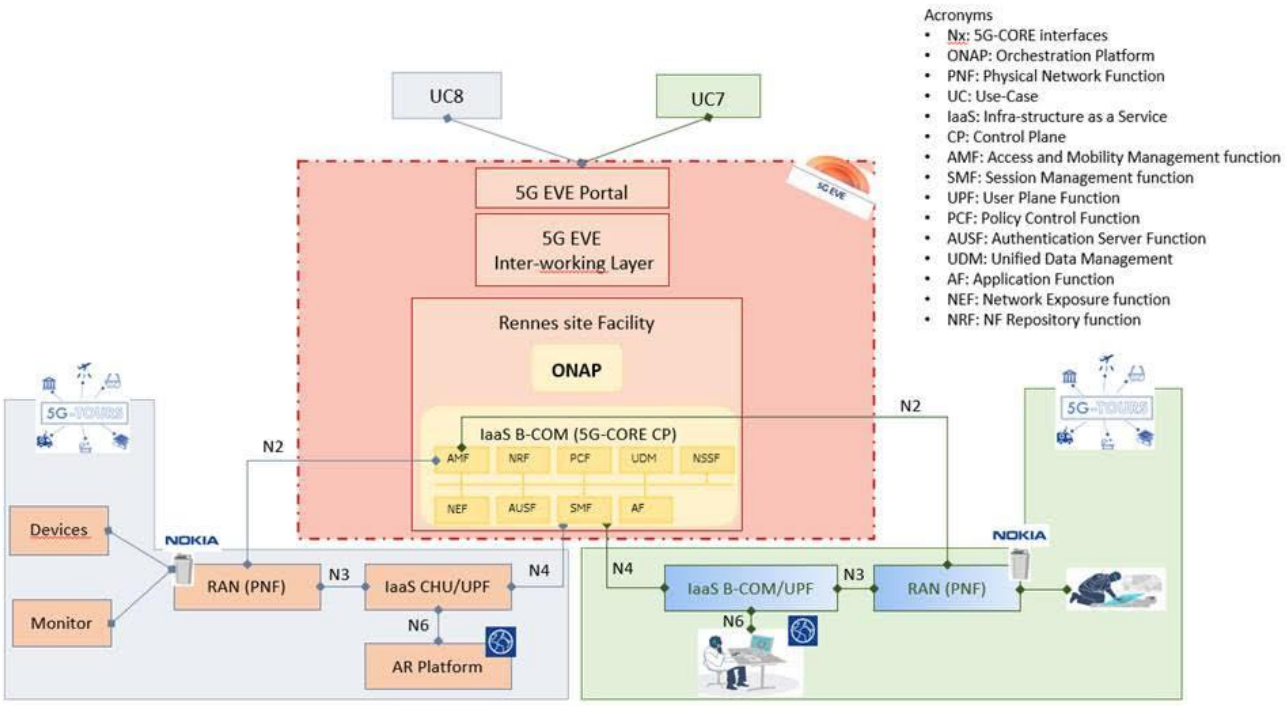


Figure 13. Use cases 7 and 8 integration in 5G EVE.

The 5G EVE Portal API enables a programmable interaction between 5G-TOURS and 5G EVE at the portal level. Such API documentation is available in 5G EVE D4.2 [32], which includes the general description and the functionalities of the first version of the portal, and in 5G EVE D4.3 [30], which includes the functionality extensions made to the first version. The 5G EVE Portal API supports experiment lifecycle management operations (e.g., instantiation, termination, polling status, etc.), whilst all the experiment design operations are available only through the 5G EVE Portal GUI.

The integration relies on the interworking capabilities of the 5G EVE platform for handling multi-site services and experiments. Following this concept, the coordination of the provisioning of the end-to-end service is entirely delegated to the 5G EVE platform.

The first step is to define the vertical service and its subcomponents and onboard the related blueprints on the 5G EVE platform, using the 5G EVE Portal GUI.

As depicted in Figure 13, the 5G-CORE control plane is part of 5G EVE platform. The 5G-CORE user plane (UPF) will be instantiated in Edge nodes deployed in CHU Rennes and in the BCOM data-center.

To make all this set-up working, a pre-provisioning of connectivity between 5G-TOURS and 5G EVE sites has been put in in place through a secure VPN. In the following Figure 14, we show some measurements of the VPN Link:

- Traffic emulation has been used for testing the VPN interconnection between French locations facilities and Orange headquarter based in Châtillon. Some initial performance has been evaluated, as shown in Figure 14. The VPN interconnection performance was evaluated between Orange Châtillon and BCOM premises that are about 300 kilometres apart in direct line. The first results have shown that the 1 Gbps tunnel was quasi filled with UDP and TCP Packets. The Jitter was very small; less than 0.06 ms and the delay is equal in average to 28 ms. The last latency value could be improved a little bit. Figure 14 illustrates the achieved values.

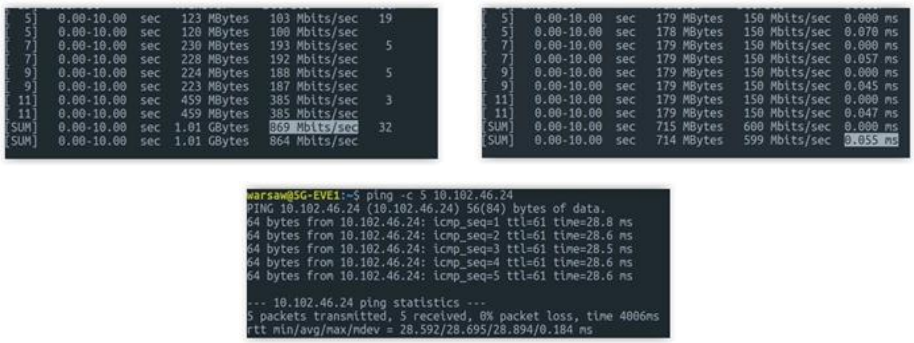


Figure 14. Orange and BCOM interconnection performance results: throughput, jitter and latency values.

With respect to the VPN connection between the Rennes University hospital (CHU Rennes) and the BCOM premise, network performance characterisation tests have been done. The following network KPIs were measured (Table 3):

Table 3. Collected KPI measurements.

Latency	~ 16 ms average.
Bitrate	From BCOM to CHU Rennes: ~ 50Mbps; From CHU Rennes to BCOM: ~ 136Mbps.
Jitter	< 0.2 ms.

3.3 Athens site

The different use cases running in Athens site, will be deployed in the AIA airport. It should also be noted that the use cases running in Athens will now also include the WP5 Safe City UC6 “Remote health monitoring and emergency situation notification” and UC9 “Optimal Ambulance routing”. As mentioned in D5.2, testing of these UCs will be performed, at least initially, at the Athens site as for these cases it would currently only be possible to use the commercial network in Rennes. In this way restrictions related to the current situation (pandemic) are overcome, while it is also useful for exploitation purposes.

The 5G coverage extension of the Athens site is already deployed, and its architecture is depicted in Figure 15. The 5G EVE is now enhanced by a new location in AIA connected to the 5G-CORE (part of 5G Eve platform) running in OTE Labs.

As illustrated in the figure, 4 indoor and 2 outdoor pair antennas (3.5-3.6GHz) are connected to 2 BBUs (BBU1 and BBU2) inside different buildings (B2 and B11). The 2 BBUs are connected directly to a switch at AIA. Also, Small form-Factor Pluggable (SFP) probes are connected into the same switch for the need of real time measurements. A Streaming Server is connected also for the need of UC11 for transmitting emergency 4K video.

The OSN switch at AIA is connected through a 10Gbps line to another OSN switch at OTE Labs, where the 5G-EVE infrastructure is developed. The network path for each UC is the following:

- UC6, UC9 (from WP5) and UC10 work with outdoor antenna in B11 building, BBU2, OSN at AIA, OSN at OTE, 5G-EVE infrastructure;
- UC11 works with outdoor antenna in B2 building, BBU1, OSN at AIA, OSN at OTE, 5G-EVE infrastructure;
- UC12 works with 3 indoor antennas in Satellite terminal, BBU2, OSN at AIA, OSN at OTE, 5G-EVE infrastructure;

- UC13 scenario A (AR) works with indoor antenna in B1 building, BBU1, OSN at AIA, OSN at OTE, 5G-EVE infrastructure;
- UC13 scenario B (VR) works with outdoor antenna in B11 building, BBU2, OSN at AIA, OSN at OTE, 5G-EVE infrastructure.

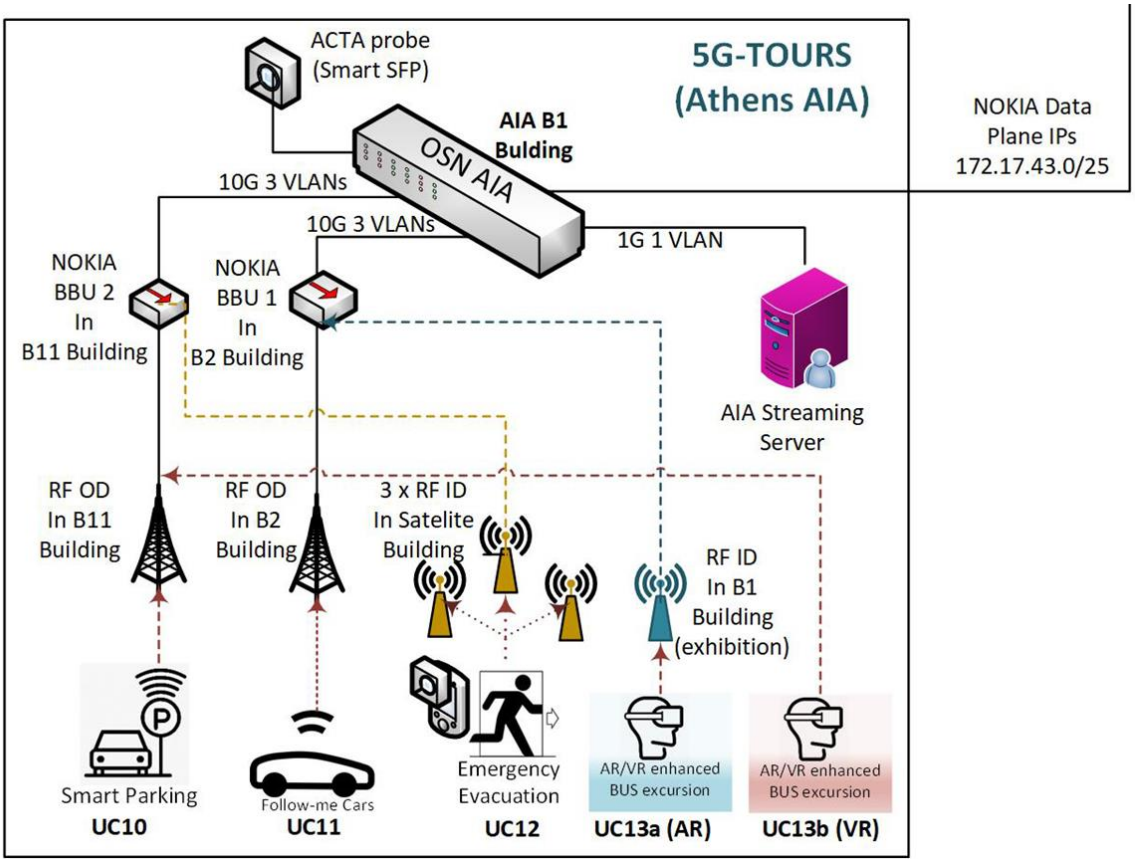


Figure 15. AIA extension location of Athens site for use cases 10, 11, 12 and 13.

In Figure 16, the interconnection for each one of the four UCs of the Athens site, are depicted in correlation with the NOKIA's 5G CORE assets of 5G-EVE infrastructure. For the needs of Athens site extension an installation of 2 Virtual Machines (VMs) at OTE Labs, one for serving the back-end content needs of ATOS/Samsung AR/VR applications and one for serving the AR and Smart Parking/Evacuation application of WINGS. Also, a Streaming server has been installed at AIA for the needs of UC11.

All the UCs use the same network path for the applications. App → 5G antenna → BBU → OTE Edge Network at AIA → Backhaul → 5G EVE EPC at OTE Labs. For the needs of UC10 the commercial 5G public network of OTE is used for sending parking spots status info to the WINGSPARK cloud at OTE Labs infrastructure.

Figure 17 and Figure 18 depict the connection and integration of the WP5 Safe city UC6 and UC9 hosted at the Athens site respectively.

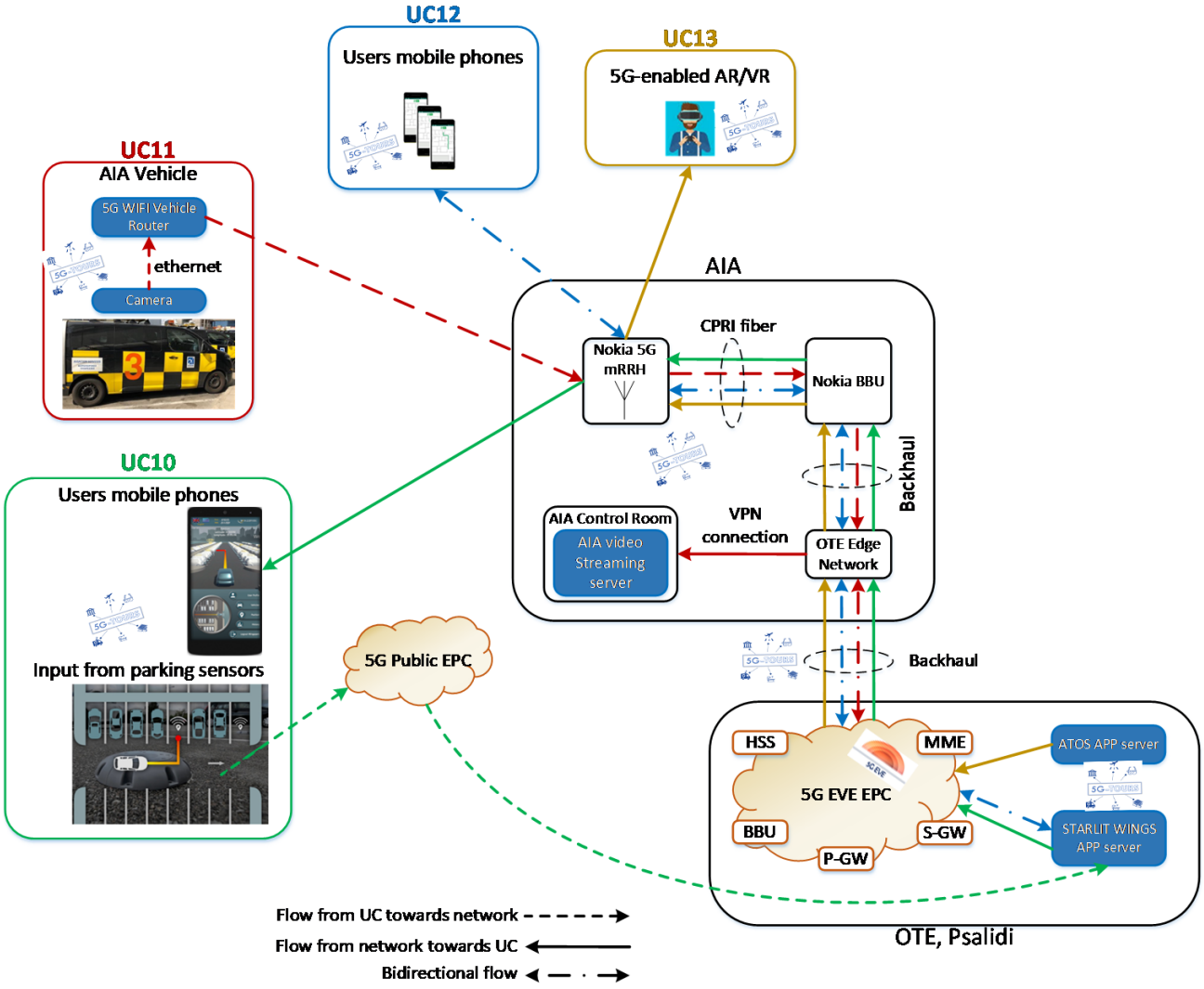


Figure 16. Use cases 10, 11, 12 and 13 integration in 5G EVE Greek Site.

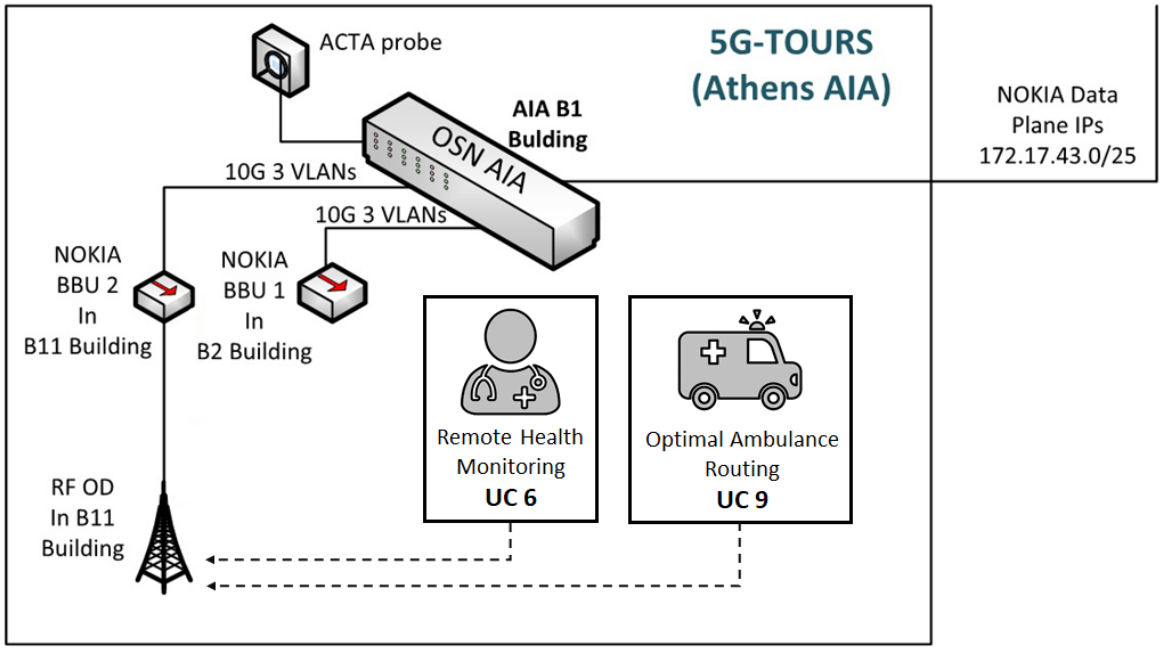


Figure 17. Connection of WP5 Safe city UC6 and UC9 hosted at the Athens site.

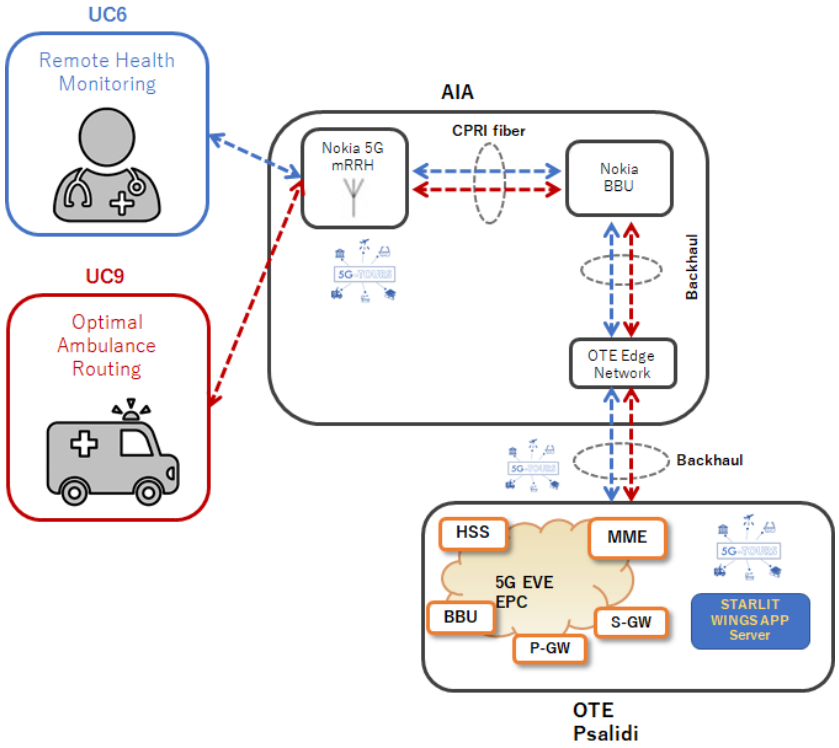


Figure 18. Integration of WP5 Safe city UC6 and UC9 hosted at the Athens site.

3.4 Insertion strategies based on 5G-TOURS technologies

5G-TOURS is aiming to build new 5G networking ideas and technologies around four main areas: enhanced MANO, broadcast, Service Layer for verticals and AI for network orchestration, see Section 5 for more details.

Referring to the 5G EVE architecture, it is difficult to inject new algorithms for orchestrations since the orchestration is not part of the experiment blueprint. 5G EVE orchestration is mainly based on the MANO stack. Two approaches are implemented based on OSM and ONAP. What could be injected is the preference of deployment of some VNFs. The preferences will target a geographical area and a specific Edge cloud as it is in Rennes site.

The AI technology developed in 5G-TOURS intends to exploit the monitoring data that is continuously collected by the 5G EVE platform. The NFVI information is assumed to provide knowledge on the computational resources' capabilities (e.g., type of CPU, memory, data plane and accelerators) and availability (status and utilisation level). Building on such information and running the AI-based algorithms, the framework can then influence and optimise placement decisions made by the Virtualized Infrastructure Manager (VIM), while ensuring that resources allocation and SLAs are adhered to. Moreover, by using this information, we can further optimise resource utilisation by:

- enabling higher density for a given set of workloads under the associated SLA;
- anticipating and reacting to changing loads in different slices and assisting the VIM in avoiding resource conflicts, and/or;
- timely triggering of up/down scaling or in/out scaling of associated resources.

More details are presented in the next sections.

To summarise, Table 4 depicts the ideal injection location to demonstrate the 5G-TOURS innovations. For example, to demonstrate the Broadcast innovation, either we mount a new location in Turin site, or we just enhance the deployed site. For the moment, the broadcast innovation is handled in a dedicated platform in UPV.

Table 4. 5G-TOURS network technology and injection into 5G EVE possibilities.

5G-TOURS Technology	5G EVE Injection place			5G-TOURS specific platform Status at March 2021
	Portal	Site	Location	
Broadcast	-	X	X	**Platform deployed in UPV labs**
AI	-	X	-	**PoC proposed by ATOS**
Service	X	X	-	**Deployed in Athens site through the OSM API**

3.5 Insertion strategies based on 5G-TOURS use cases

5G-TOURS is targeting 13 use cases grouped into three clusters. WP4 is in charge of the "Touristic city" cluster, WP5 is in charge of the "safe city" cluster, and WP6 is in charge of the "Mobility-efficient city" cluster.

In each WP, the work is focused on the definition of the required VNFs, blueprints, networking needs etc. Once these parameters are defined, we can define the use case insertion strategy.

At this stage of the analysis, the insertion strategies are presented in Table 5:

- Tourist city cluster (WP4 use cases):
  - 5G EVE portal will be used to expose new "touristic" services.
- Safe city cluster (WP5 use cases):
  - ONAP is the primary "insertion technology" for supporting new domains & services. implies the need to declare the new location in the ONAP domain by updating its multi-cloud Open-Stack to be able to instantiate VNFs in the right site.
  - 5G EVE portal is used to expose new "healthcare" services.
  - Use cases UC7 will be deployed via 5G EVE Blueprints.
- Mobility-efficient city cluster (WP6 use cases):
  - 5G EVE portal will be used to expose new "mobility" services.

Table 5. Insertion strategies per use cases cluster

Use cases Insertion point	Tourist city	Safe city	Mobility-efficient city
5G EVE portal	Yes	Yes	Yes
Orchestrator (ONAP, OSM)	Use 5G EVE orchestrator. No update from 5G-TOURS is planned	Need to update ONAP domain to add new OpenStack cluster	Use 5G EVE orchestrator. No update from 5G-TOURS is planned
Site	Turin	Rennes	Athens
Location	Yes. Need to add new location to cover Palazzo Madama and GAM Museum	Yes. Need to add new location to cover BCOM parking and the Operating Room in CHU Rennes	AIA airport

4 Network infrastructure & deployment

4.1 Touristic City Deployment Update

4.1.1 Deployment of Physical Infrastructure Phase 1

As discussed in Section 2.2.2.1, the overall physical architecture of the Turin site differs between Phase 1 and Phase 2, where the network is gradually upgraded through various deployed enhancements. In order to demonstrate 5G capabilities and early showcase the use case trials, the physical infrastructure for Phase 1, in Palazzo Madama, was implemented based on an extension of the TIM commercial network infrastructure.

The Network solution in Phase 1 is based on a 3GPP NSA Option 3 architecture in which the radio access network is composed of an LTE «anchor» layer working at 1800 MHz and a 5G layer at 3.7 GHz with 80 MHz bandwidth. The 5G indoor coverage was provided by Ericsson Radio units 4422 with Kathrein 80010922 antennas [3]; in order to provide coverage in the requested rooms, the solution deployed 4 different radio installations mounted on ad-hoc built poles to ensure the best stability and allow the mounting of all required components in terms of radio, antenna (two antennas for each radio), power supply and cabling (optical fiber and RF), see Figure 19.

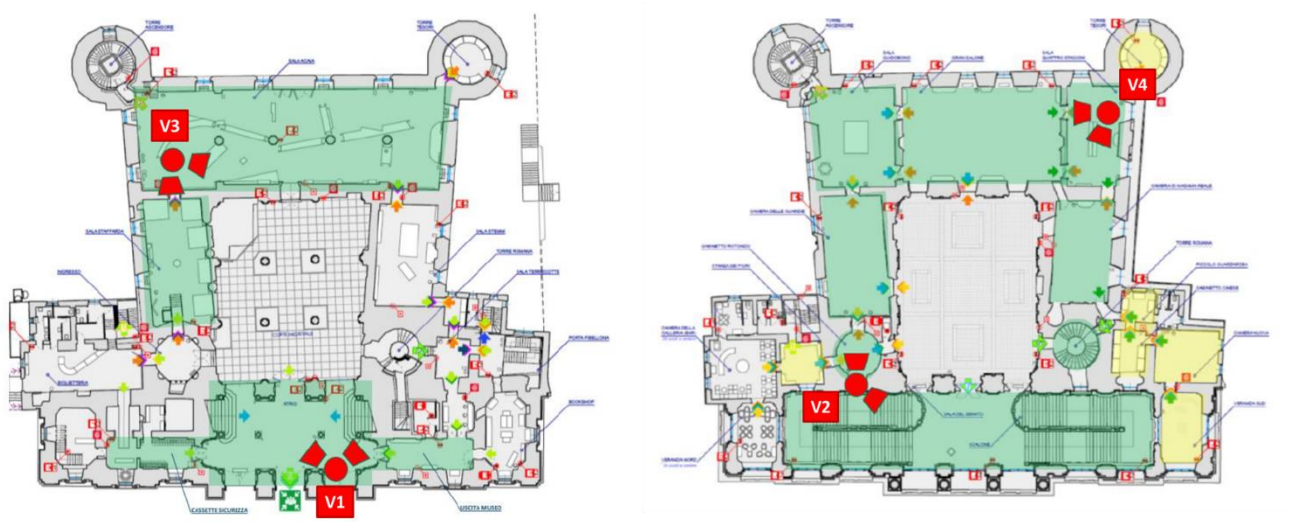


Figure 19. Location of the installations and 5G indoor coverage area (ground floor and first floor).

The baseband unit for the 5G indoor coverage was the Ericsson Baseband 6630 [3] located in the TIM network exchange point located 2.8 Km from Palazzo Madama. In order to provide the 5G fronthauling connection between the radio inside Palazzo Madama and the baseband, an ad-hoc optical fiber connection has been installed consisting of 8 couples of fibers of which 4 couples were used to connect the 4 radio units 4422, 1 couple to provide broadband Internet connection for the UC's servers installed at the museum and 3 spare couples as backup and/or future development of the 5G indoor coverage. 5G outdoor coverage was implemented deploying a dedicated 5G Baseband 6630 and an AAS radio AIR6488 [3], see Figure 20. The Indoor and Outdoor Network extension was connected to the TIM commercial Core Network.

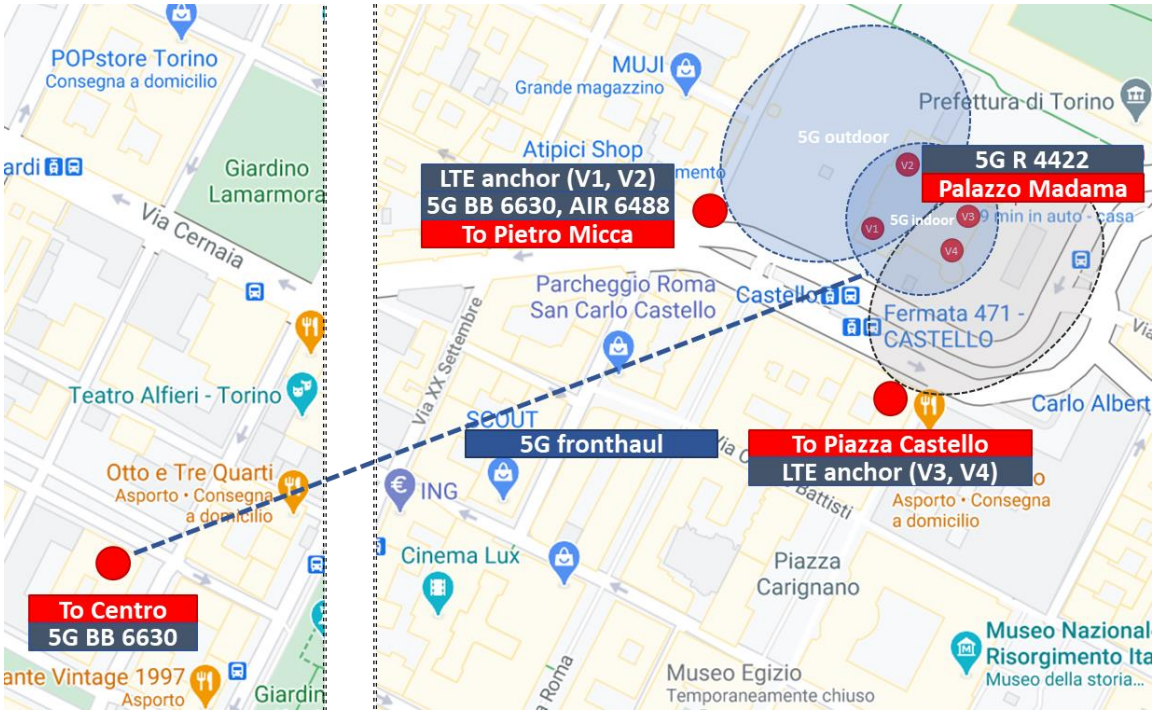


Figure 20. High level architecture of 5G-TOURS network solution at Palazzo Madama.

Figure 21 shows the installations in Sala Quattro Stagioni and Sala Acaia for the 5G indoor coverage and the rooftop antenna that provided the 5G outdoor coverage.

Figure 21. Installation for 5G indoor coverage in Sala Quattro Stagioni and Sala Acaia and rooftop antenna for the 5G outdoor coverage.

4.1.2 Deployment of Physical Infrastructure Phase 2

The phase 2 deployment will consist of two network architecture instantiations:

- In-the-field network solution, where the 5G indoor RAN coverages will be connected to the TIM commercial core network. In terms of museum infrastructure integration, this can be considered as an evolution of Phase 1.
- The TIM laboratory Network, using the end-to-end 5G NSA network solution provided by 5G EVE (D3.1 [50]).

The phase 2 in-the-field final technical solution and positioning of the indoor equipment inside Palazzo Madama and GAM museums is being examined by the Superintendency for Cultural Heritage in order to define how to insert the radio installation inside the museum furnishings for the whole duration of the project. It is expected that an agreement with all the stakeholders will be made soon, based on the experiences obtained in Phase 1. At the moment, two alternative technical solutions are under evaluation, based on R4422 or DOT radios [3]. The solutions to be evaluated will have to meet two key requirements:

- achieve the same network performances as obtained during phase 1 with the UC5 trial and
- be in compliance with the aesthetic constraints and needs of the cultural sites hosting the trials.

The phase 2 laboratory infrastructure consists of a 5G RAN (with NSA nodes), fronthaul and backhaul and open source tools for the management and orchestration of the NFV infrastructure.

Figure 22 shows the Phase 1 and Phase 2 approaches representing the Network physical Infrastructure.

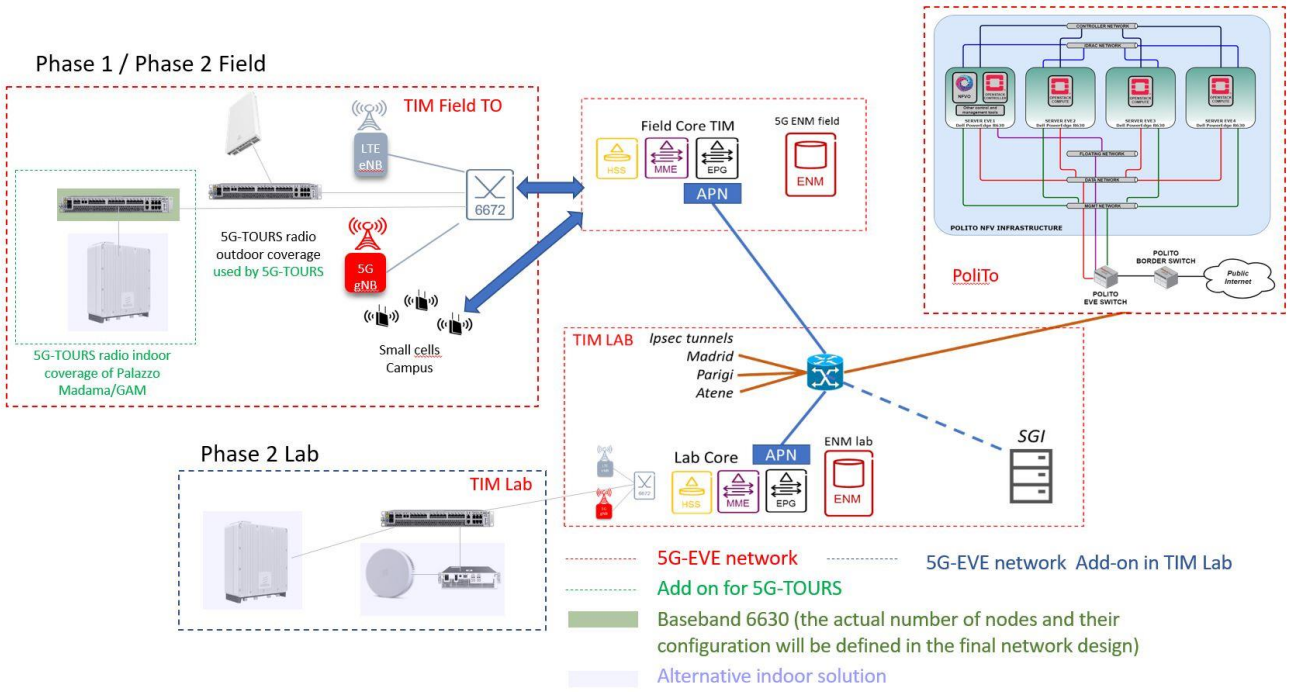


Figure 22. Deployment of Physical Infrastructure.

4.2 Safe City Deployment Update

The safe-city scenario focuses on connected and remote healthcare use cases enhanced by 5G technology. The COVID-19 crisis has, and still is accelerating the demand for connected healthcare solutions, which has further elevated the relevancy of these use cases since the start of the 5G-TOURS project.

Since the release of deliverable D5.1 [19] and internal report IR5.1 [20], all use cases have progressed in terms of the level of detail of their definition, the integration of application components, design of the application architecture, their implementation and testing.

Considerable progress has been achieved in the definition of the network infrastructure as needed for the implementation of these use cases on the experimental 5G network of the Rennes node. At this point, network deployment in terms of antenna placement, available frequency bands, base stations and edge/core networks has been defined. This includes the definition of VNFs, UPF and CPF allocation, VPNs between hospital, BCOM premise and the 5G EVE core network of Orange in their Châtillon datacenter. Also, the overall network architecture has been designed, while the above-mentioned VPNs have been characterized. Implementation and testing of the use cases on the Rennes 5G network is ongoing.

4.2.1 Deployment of physical infrastructure

From the "Safe City" work package (WP5) use cases UC7 and UC8 will be trailed in Rennes, using the mobile network infrastructure of Orange and Nokia. Phase 1 developments are exclusively done in a lab environment using BCOM's WEF v1.3, Amarisoft Software 5G RAN, and applications from AMA and BCOM (described in D3.2 [3]). The infrastructure for phase 2 currently supports a 5G mm-Wave network for enhanced mobile broadband (eMBB) communication for use cases 7 and 8, as well as a commercial LTE-M network for machine type communication (MTC) of IoT devices in the use cases 6 and 9. As already mentioned, due to only commercial network availability and the current restrictions imposed by the Covid-19 pandemic for UC6 and UC9 trialling of these use cases will be done in the Athens site. We focus here on the network deployment of the 5G experimental network.

There will be two deployments of 5G NR wireless coverage:

1. **Outdoors:** at the BCOM premises, for the connected ambulance, as shown in Figure 23. A suitable 5G NR antenna will be installed on the roof of the BCOM building, using primarily the 26 GHz frequency band.
2. **Indoors:** at the Wireless Operating Room at CHU Rennes to provide high-speed, low-latency wireless access for medical imaging equipment, using 26 GHz for data transmission and 2.6 GHz as the anchor frequency band, see Figure 24. The anchor frequency band refers to the 4G frequency used to carry the control messages in a 5G Non-Stand Alone (NSA) network. We have currently no plan to deploy a 5G Stand Alone network in the scope of 5G-TOURS.

Figure 23. 5G-TOURS 5G NR NSA/SA wireless coverage at BCOM.

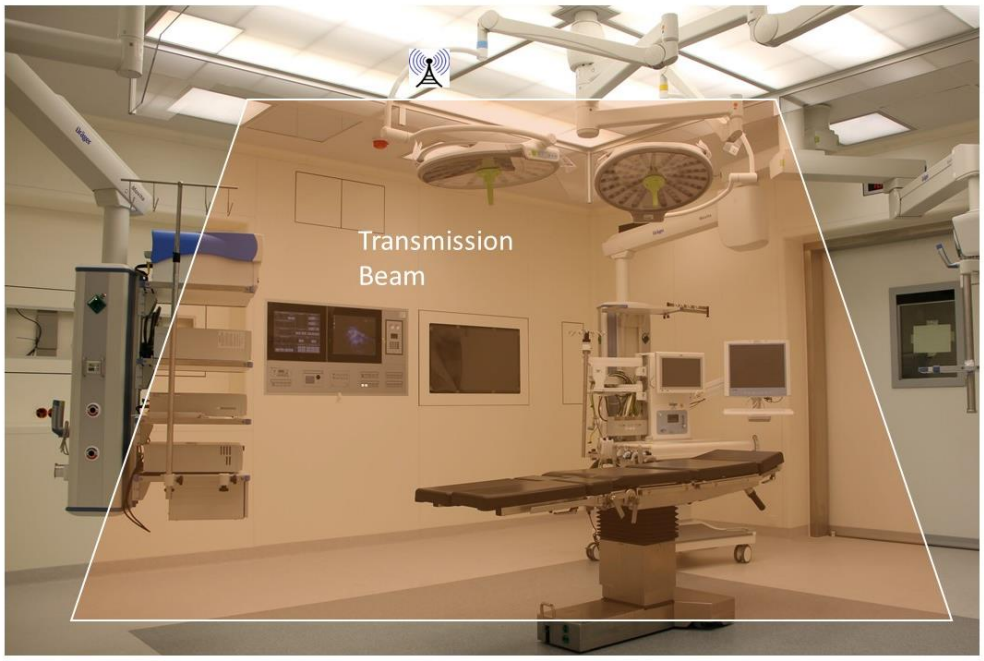


Figure 24. 5G-TOURS 5G NR wireless coverage in the Wireless Operating Room at CHU.

At the BCOM premises, there will be a 5G base station with a local virtual UPF, part of the so-called "Wireless Edge Factory" (WEF) [21]. Similarly, there will be a WEF UPF at the hospital that connects to the WEF core network hosted in the BCOM datacentre through a dedicated VPN backbone. This is depicted in Figure 25. This will enable the setting of end-to-end network performance KPIs and the prioritization of data traffic between the ambulance and the hospital to guarantee the required quality of service. Furthermore, the WEF Core Network deployed in BCOM data center will manage the WEF UPF at the hospital to connect the 5G terminals of the Wireless Operating Room.

In addition, for the non-critical overall network orchestration and automatic deployment of the WEF core network, Orange provides an ONAP orchestrator in their Châtillon data center as part of their 5G EVE infrastructure. ONAP enables the user or the experimenter to deploy and configure the WEF Core Network on demand. It could also be used to deploy the user plane part of the WEF.

The Orange data center has already been connected to the BCOM data center within the scope of the 5G EVE project. This is also shown in Figure 25.

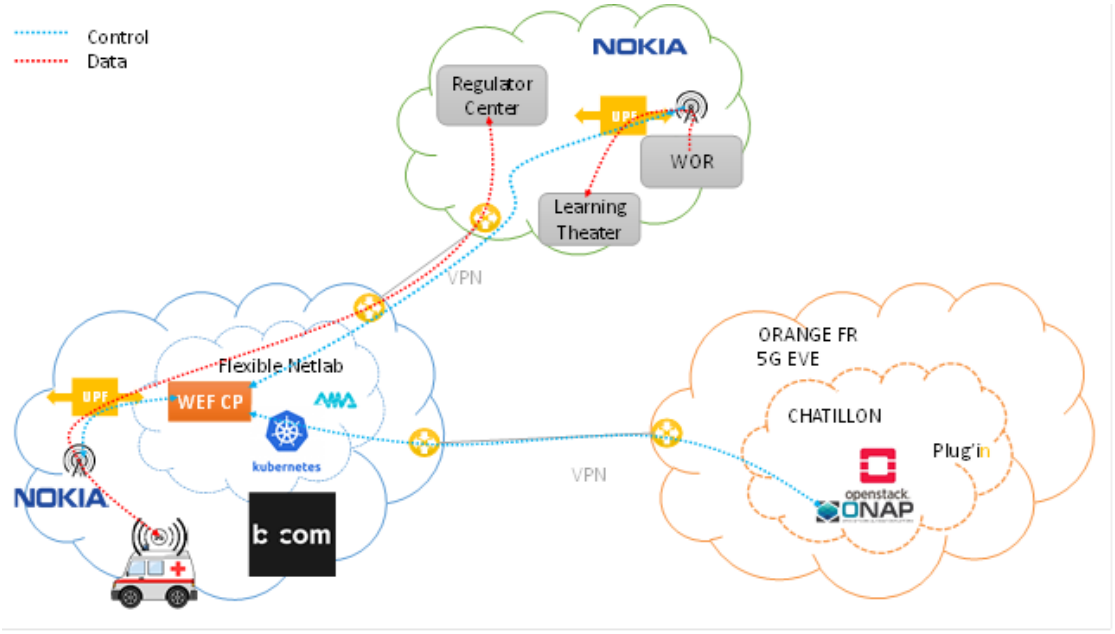


Figure 25. Overall network architecture and physical deployment of network equipment and functions.

The network equipment is described below in three steps: control plane (CP), user plane (UP), and radio access network (RAN) equipment.

4.2.1.1 Control plane network equipment

The control plane is a virtual 4G Core Network compatible with the 5G NSA standard [22]. The Control Plane is part of the WEF solution developed by BCOM. The WEF core network has also evolved from Phase 1 deployment with the version v1.3 based on pre-release 15 and an Openstack Virtual Machine deployment while version v2.2 of Phase 2 deployments will be based on 3GPP release 15 and a cloud-native container-based Kubernetes deployment. This cluster will be hosted on the Flexible Netlab platform [23] in the BCOM data center. The Control Plane will be deployed and orchestrated by an instance of the ONAP orchestrator hosted by Orange in Châtillon. Both locations are interconnected through a secure and actively maintained VPN.

4.2.1.2 User plane network equipment

The user plane equipment provides connectivity between the RAN equipment and the data network (Internet). The main component is the User Plane Function (UPF) component of the WEF provided by BCOM. Two instances of the UPF will be deployed as part of 5G-TOURS in Phase2.

The first instance will be a VNF i.e., a purely virtual UPF deployed in BCOM data center as a virtual machine hosted on an OpenStack [24] cluster provided by Flexible Netlab. This virtual machine hosts an OpenVSwitch [26] (OVS) virtual switch that acts as a tunnel endpoint for the GTP tunnels coming from the RAN equipment deployed at BCOM for use case 7 (UC7). It is thus used to connect this RAN equipment to the Rennes CHU through the VPN. The WEF Control Plane manages the virtual switch under control of the OpenDaylight [27] SDN controller that is deployed in the control plane. The second instance is a PNF i.e., an appliance built from a COTS network switch and a COTS mini-ITX PC. The PC is a KVM [28] hypervisor that hosts an OVS-based virtual machine similar to the one deployed in Flexible Netlab. It will be installed in the technical room of the Rennes CHU and will interconnect the RAN equipment deployed there with the various components required by use case 8 (UC8). The same WEF Control Plane will manage this switch through the VPN established between BCOM and the CHU.

4.2.1.3 RAN equipment

For 5G-TOURS, we will use Nokia Small Cell technology as our RAN equipment (Phase 2). Two small cells will be deployed: one at the Rennes CHU to provide coverage for the Wireless Operating Room and one at BCOM premises to cover the outside area for UC7. Both will use the 26GHz/2.6GHz bands in 5G NSA mode.

The Nokia RAN at BCOM will be deployed and operated by BCOM while the one at the CHU will be deployed and operated by Nokia. In addition, we conducted the first integration tests for UC8 in the BCOM showroom. These tests were carried out with Amarisoft Classic Callbox RAN equipment [25] during Phase 1. This equipment uses the 3.5GHz band and is also compatible with 5G Non Stand Alone (NSA) mode. At this point, we do not plan to move to a 5G Stand Alone (SA) network in the future.

All medical equipment that requires 5G wireless connectivity has been connected to this RAN equipment through a commercial "off-the-shelf" (COTS) Huawei 5G Pro CPE component [29].

4.3 Mobility-Efficient City Deployment Update

4.3.1 Deployment of physical infrastructure

The mobility-efficient city presents a set of use cases that improve the tourism and tourism-related experiences from various perspectives. The implementation of the four UCs relies on the 5G EVE Greek site infrastructure and on an extension that is implemented at the AIA premises during the 5G-TOURS project. To support the needs of the 5G-TOURS project and the implementation of the four WP6 UCs at the AIA premises, the following additional equipment has been installed to extend the existing 5G EVE Greek Site.

5G EVE main H/W installations and integrations at OTE premises:

- A fully functional 5G network installed and configured by NOKIA-GR that is up and running.
- An Orchestrator (OSM) installed and configured by WINGS that is also up and running.
- The OSM Orchestrator is fully interconnected with the Interworking Layer (IWL) of the Turin 5G EVE site through a secure tunnel and the portal of 5G EVE.
- A Kafka bus server, for the needs of keeping the metrics of the KPIs, is installed and interconnected with central Kafka server.

5G-TOURS H/W extended site installations and integrations at AIA and OTE premises:

- 2 outdoor and 4 indoor antennas designed and installed by NOKIA, that utilise the spectrum band of 3500-3600MHz are up and running.
- All the antennas are connected via optical fibre with the NOKIA's BBUs at the airport.
- OTE L2/L3 OSN switch is interconnected to the OTE IP Core, using a 10 Gbps capacity line, in order to interconnect 5G EVE Greek site infrastructure located at OTE Labs in Psalidi-Attika with AIA.
- E2E interconnection with 5G EVE infrastructure has been configured and tested successfully.
- Smart devices are used and specific innovative applications developed for the implementation of the 4 UCs (Smartphones, tablets, AR/VR headsets, IP-cameras, IoT sensors, etc.).
- For UC10, 15 parking sensors have been installed at AIA and a lot of network and application tests have taken place.
- For UC11, a streaming server has been installed at the AIA place and is interconnected on the OSN. This server steams the 4K video of the Van cameras.
- For UC12 and UC13 4 indoor antennas have been installed into 2 different AIA buildings, that have been tested E2E.
- Also, KPI measuring probes have been installed between antennas and the BBU, as well as between the BBU and OTE's Core-switch (at AIA). Finally, a probe is also placed before NOKIA's platform (ePC) at OTE-Labs. These probes are used for measuring network performance and service layer metrics in real time in order to validate KPIs of the network (see Figure 26 below).

Figure 26 below depicts the outdoor radio equipment installed at AIA to be used as part of the Use Cases demonstration.

Figure 26. Outdoor Radio equipment installation at AIA.

The up-to-date interconnection diagram/architecture of the Athens node for the needs of implementation of the mobility efficient city is depicted in Figure 27 below.

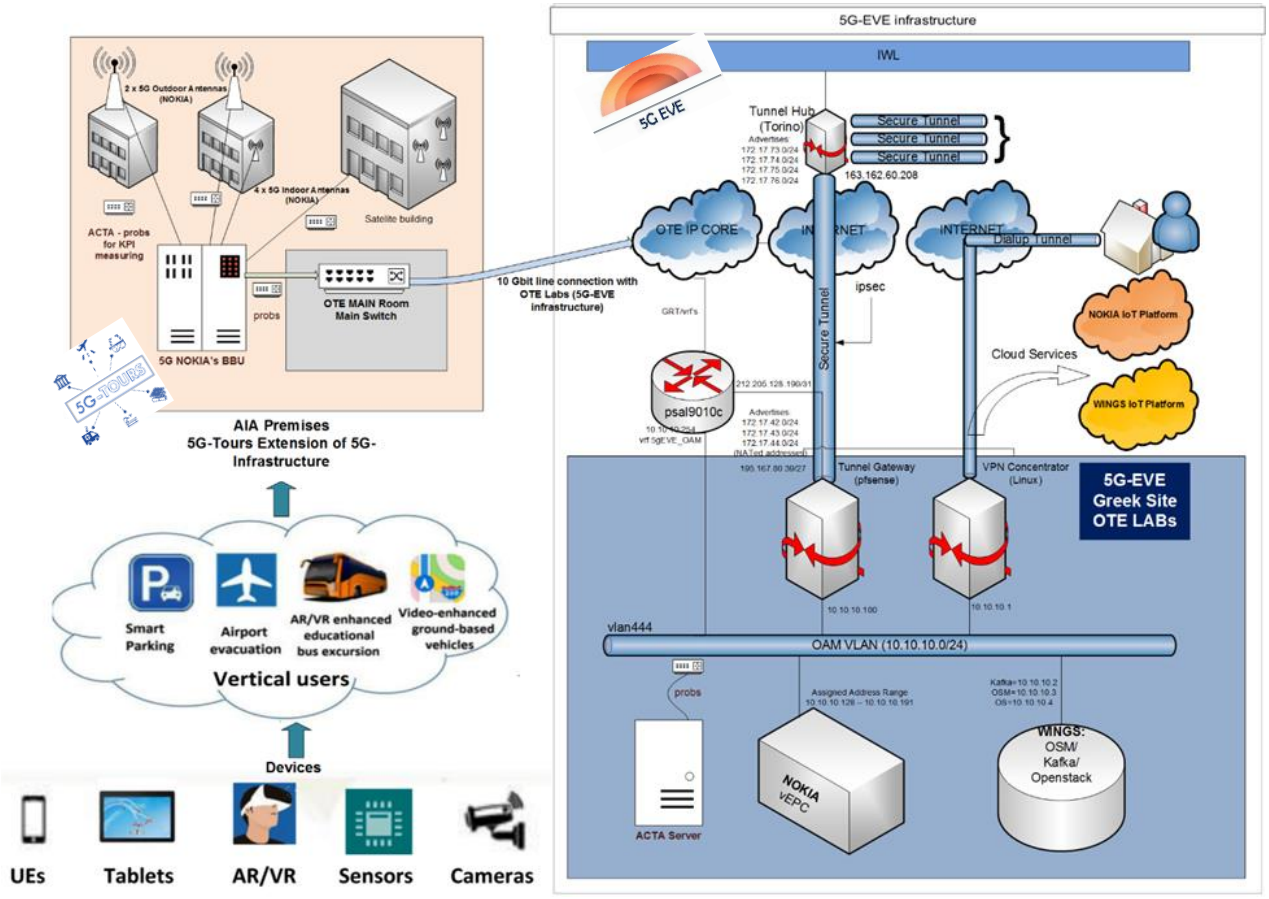


Figure 27. 5G-TOURS Athens node platform.

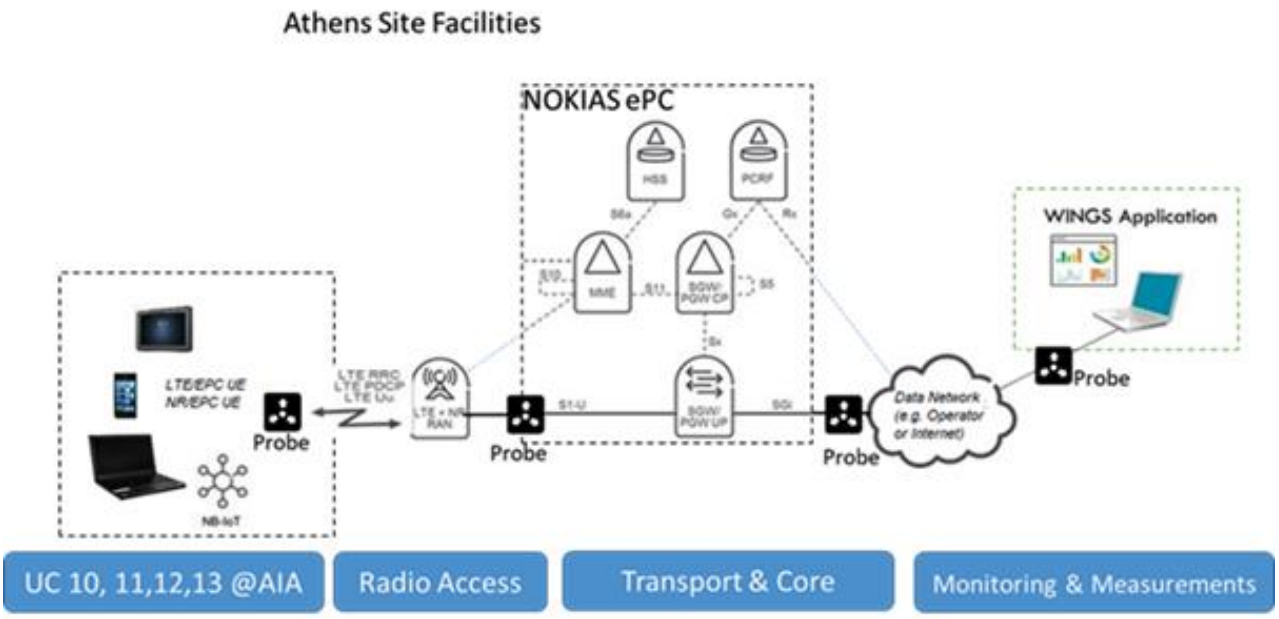


Figure 28. Probes installed to Athens node.

The Athens node network infrastructure deployment is based on a two phases' implementation approach in order to meet the coverages, the performances and Quality of Service requirements of the different use cases, their development and trials roadmaps, as well as the integration with the 5G EVE platform/infrastructure. The main aspects of the two implementation phases are summarized hereafter.

Phase 1

The phase 1 of the network implementation in Athens reuses the 5G EVE platform/infrastructure offering 4G with Option 3X Radio and Core Network capabilities. Specifically, for the Smart Parking Use Case, Cosmote' NB-IoT network will be used for registration and uplink data communication of Smart Parking sensors to the Application Server whereas the 5G EVE infrastructure will be used for the registration and data communication of the Smart Parking application users to the Application Server. For the phase 1, the roadmap foresees to have a running network by the end of 2020 / beginning of 2021 so that the first UC's trials can take place according to the respective time plans.

Phase 2

The phase 2 of the network implementation will again rely on 5G Eve platform/infrastructure offering 5G Standalone Radio and Core Network capabilities. Note that the same Radio hardware is going to be used as for phase 1, given that the latter can be configured to work either as Option 3X or as 5G Standalone RAN solution. The network implementation of phase 2 will validate the need of employing the 5G technology and demonstrate the benefits of 5G-TOURS innovations in terms of enhanced network functionalities and capabilities.

Core 5G-TOURS network solution at AIA and Psalidi (OTE's Labs)

In line with the phased approach described above, Nokia GR and OTE started to work on the network solution to provide the Radio Coverage (LTE and/or NR) at Athens International Airport (AIA).

As discussed above, the 5G-TOURS Core Network reuses the 5G EVE Core Network deployment for both Phase 1 and Phase 2 of the project. Thus, as part of phase 1 deployment, a vEPC Core Network is deployed consisting of MME, SPGW, HSS, PCRF offering 4G with Option 3X Radio and Core Network capabilities and possibly GMLC, ESMLC to support location service. Regarding phase 2 deployment, a 5GC Core Network will be deployed consisting of AMF, SMF/UPF, UDM, NRF/NSSF offering 5G Standalone Radio and Core Network capabilities and possibly GMLC, LMF to support location service (see below for abstract architecture of deployed network).

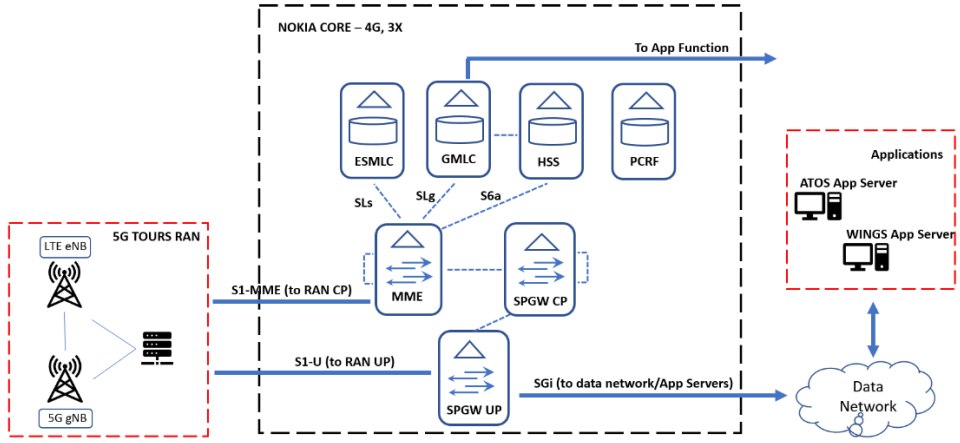


Figure 29. High level view of 5G-TOURS RAN and CORE Network infrastructure at Greek Node (Phase 1).

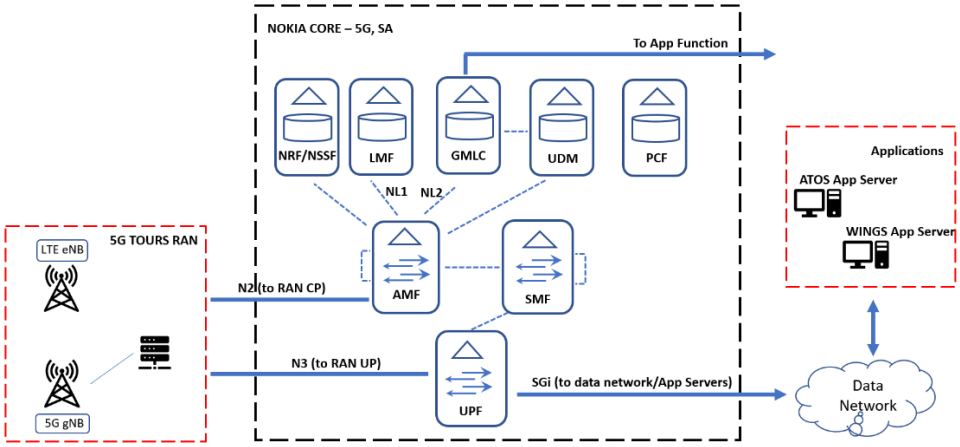


Figure 30. High level view of 5G-TOURS RAN and CORE Network infrastructure at Greek Node Phase 2).

The radio access network is deployed at AIA site (see below for details on radio deployment locations for all Use Cases and deliverable D6.2 [8] for RAN details) and then connected to either vEPC or 5GC, located at Psalidi area (OTE premises) by dedicated 10Gbit line. The 5G EVE platform is then offered access to both OTE intra-network and public internet so that it may connect with Application Server(s) needed for the Athens Use Cases.

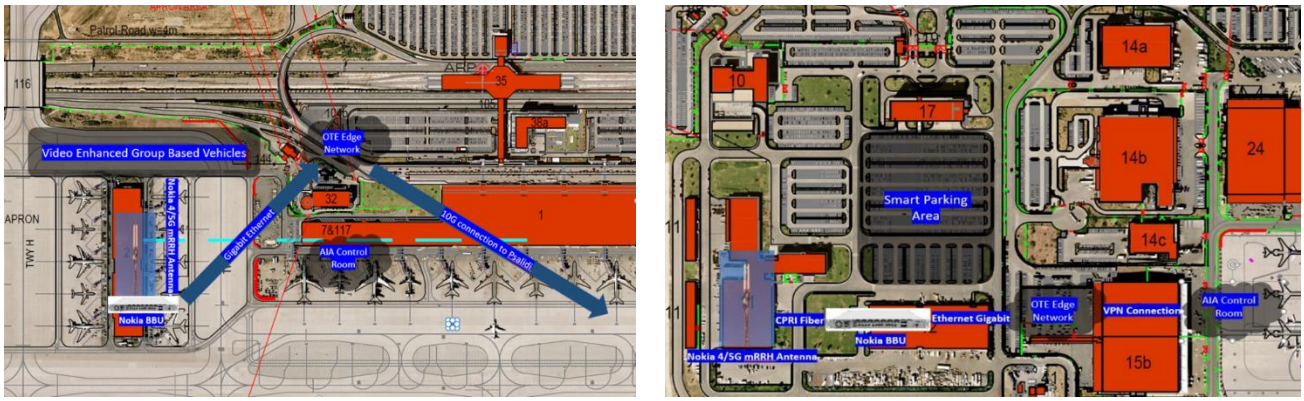


Figure 31. RAN coverage of AIA for Smart Parking UC10 and Video enhanced airfield vehicles UC11.

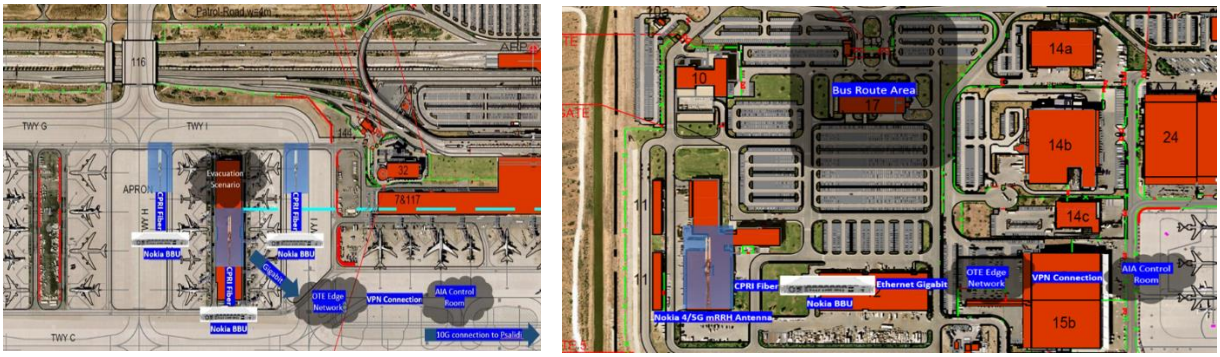


Figure 32. RAN coverage of AIA for Evacuation UC12 and AR/VR Bus Excursion UC13.

5 5G-TOURS Technology Evolution

5.1 Introduction

As already discussed in D3.2 [3], the aim of 5G-TOURS is not only to deploy bleeding edge 5G technology in the different sites, but also deploy novel mechanisms and solutions in the different sites. In this section, we will discuss the instantiation of the novel technologies in the different sites, discussing how some of them are necessary enablers for many of the use cases that will be presented by the project. More specifically, 5G-TOURS provides solution for the Enhanced MANO (Section 5.2), which is embodying the needed extensions to incorporate the usage of AI, such as the one detailed in Section 5.3. Then 5G-TOURS also developed beyond state of the art solution for the broadcast support (Section 5.4). Finally, all of them are made available to the verticals through the Service Layer (Section 5.5).

In Section 2, we already introduced how we considered the requirements to design our innovations and how they are spread over the different sites and in the overall architecture. In this section we discuss the details of the different solutions and summarize them in Table 6.

Table 6. Innovations over Site.

Technology	Site	Notes
Enhanced MANO	Turin	Integration of the OSM enhancements (see Section 5.2.1)
	Rennes	Integration of the ONAP extensions to the ORA-FR deployment.
	Athens	MANO extension for the SLA monitoring (see Section 5.2.2)
Broadcast Support	Turin	Development of the Broadcast Support solution in TIM LAB under study (see Section 5.4.2). HPHT trial in Eremo (see Section 5.4.1)
Service Layer	Athens	Implementation of the KPI exposure functionality to tenants (see Section 5.5.1)
	Turin	Dynamic SLA requirements from the tenant (see Section 5.5.2)
AI Orchestration	Turin	Proactive scaling of VNFs, as part of the ETSI ENI PoC (see Section 5.5.2). Extension to the RAN to support AI (see Section 5.2.2).

5.2 Enhanced MANO

5G-TOURS leverages on the MANO assets provided by the 5G EVE project, including OSM (for Turin and Athens) and ONAP (in the French site). However, 5G-TOURS also worked on specific extensions to the orchestration framework to accommodate the new functionality envisioned in the project, especially the data driven management of the network.

5.2.1 AI-Agent functionality

As mentioned in the previous deliverable D3.2 [3] one of the identified features to experiment as an extension of the existing MANO solutions was the deployment of Artificial Intelligence agents for network service optimisation. Also, as pointed out in that deliverable, the usage of the ETSI OSM solution [36] as NFV Orchestrator in the 5G EVE platform gave us the opportunity to enhance OSM towards providing features that could be applied in the 5G-TOURS use cases, and to create a relevant impact in the Open Source community by contributing to the ETSI in the evolution of the Open Source MANO (OSM) solution.

In order to get this, a new feature was proposed to the OSM community (Feature #9063) [34] to include artificial intelligent agents (AI-Agents) as part of the OSM platform for network services optimization. This new feature

is currently under development as part of the so-called Service Assurance (SA) module. The AI-Agents design is based on a previous feature (Feature #8157 – Executions Environments) [35] which makes possible to deploy the AI-Agents as part of OSM.

Figure 33 below shows the proposed architecture to deploy de the AI-Agents feature. White blocks (NBI, LCM, POL) are those already part of the OSM architecture. Red blocks (AI-Agent and AI Models Server) are the new architectural blocks associated to the new AI-Agents feature. Overlapped grey blocks (EE) represent the Execution Environment on which AI-Agents are deployed (Feature #8157 previously mentioned). The green circles represent the main interactions among the different functional blocks to deploy and operate the AI-Agents feature.

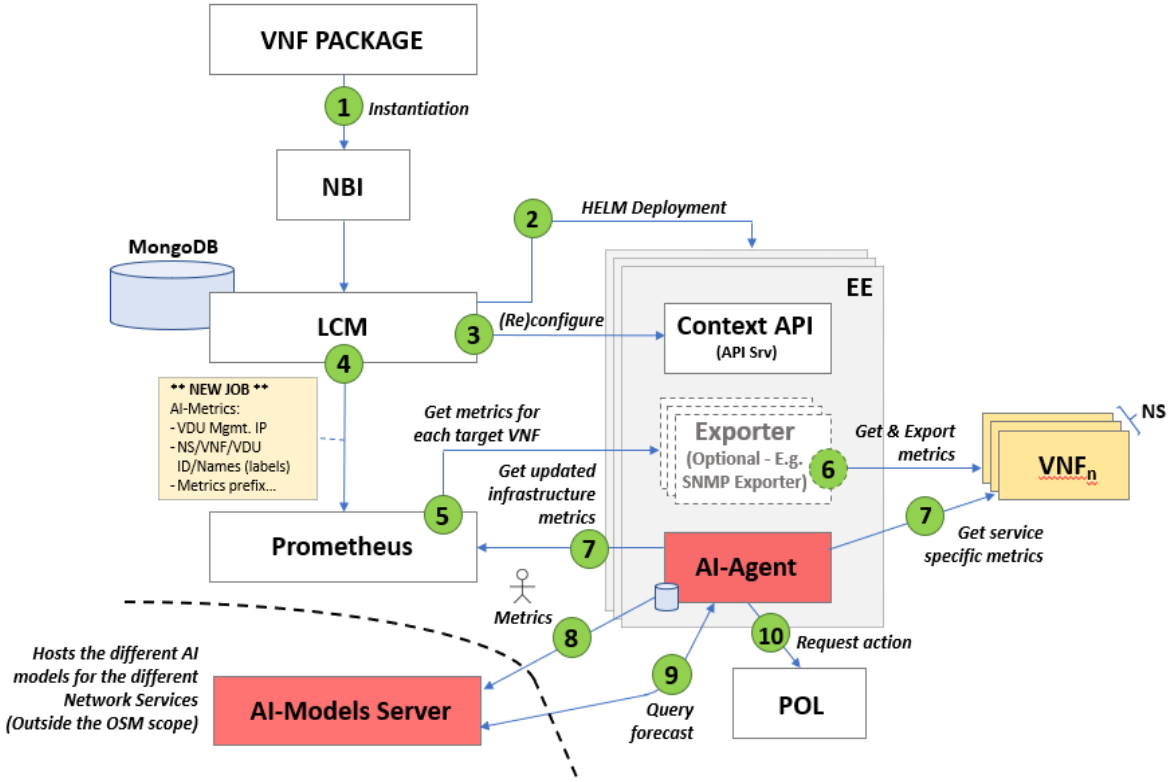


Figure 33. AI-Agents Deployment on OSM.

Focusing on the red blocks, we can see that the AI-Agent functionality is split in two: the AI-Agent itself and an external AI-Models Server. The AI-Models Server is an external component (outside the OSM scope) that can host different AI models that can be queried from the AI-Agent to perform scaling actions on the network service to which they are associated. The overall idea is that the intelligence actually resides on the external model's server, while the AI-Agent basically works as a broker with the following main features:

- It collects and normalises the relevant data from the deployed Network Services and/or the internal OSM metrics database (step 7 in the figure).
- It stores the collected data and makes it available for the AI Models Server (to enable the training of the AI models using the same data format as in the execution environment) – Step 8 in the figure.
- It accesses the AI-Models Server to evaluate possible orchestration decisions (Step 9).
- It triggers the OSM Policies Manager block (POL) to request the necessary scaling actions within the OSM scope (Step 10)[1].

[1] At the moment only scaling actions are in the scope. However, we consider other possible orchestration actions could be implemented with the same architecture (e.g., VNF placement actions).

The main advantages of this approach relying on two differentiated components (internal AI-Agent and external AI-Models Server) are the following:

- It allows an agnostic implementation regarding the associated technologies for implementing the AI models, avoiding possible vendor lock-in and allowing the NS developers to use their favourite AI framework. This agnostic approach also allows a given company already having its own AI models to easily integrate them with OSM as NFV orchestration platform.
- The external AI-Models Server allows to host AI models that could be quite complex in size and computational resources (at it usually happens in practice, especially if we consider the training stage some of them may require). The usage of the external model's server takes that complexity out of OSM.
- The solution is highly flexible, since it allows to deploy different AI/ML models regardless their design or training method. We've to consider that artificial intelligence models are diverse (e.g., there are many different topologies of neural networks and a multiplicity of learning algorithms); also, the field of the artificial intelligence technology is changing very dynamically, and what might seem adequate today may not be so in the future.
- Data normalisation is a relevant aspect in the ML domain, and it is also important this normalisation is done in the same way in the training phases of the models and in the production environment. This requirement is addressed by the AI-Agent itself, which provides homogeneous normalized data for both: the production and the training environments.

Figure 34 below shows the development model associated to the AI-Agents functionality. As we can see we consider both: the training stage (associated to the NS development stage) and the production stage. As we see the NS metrics are provided from the AI-Agent within the OSM scope. These metrics could be directly taken from the production environment. The NS development team uses these metrics to define and train the necessary AI model through an iterative process within the external AI-Models Server Framework. Once the model is ready it can be accessed from the AI-Agent to trigger scaling actions on the NFVI. As we see, these stages provide a closed control loop that can be iterated once and again to refine the AI model's performance.

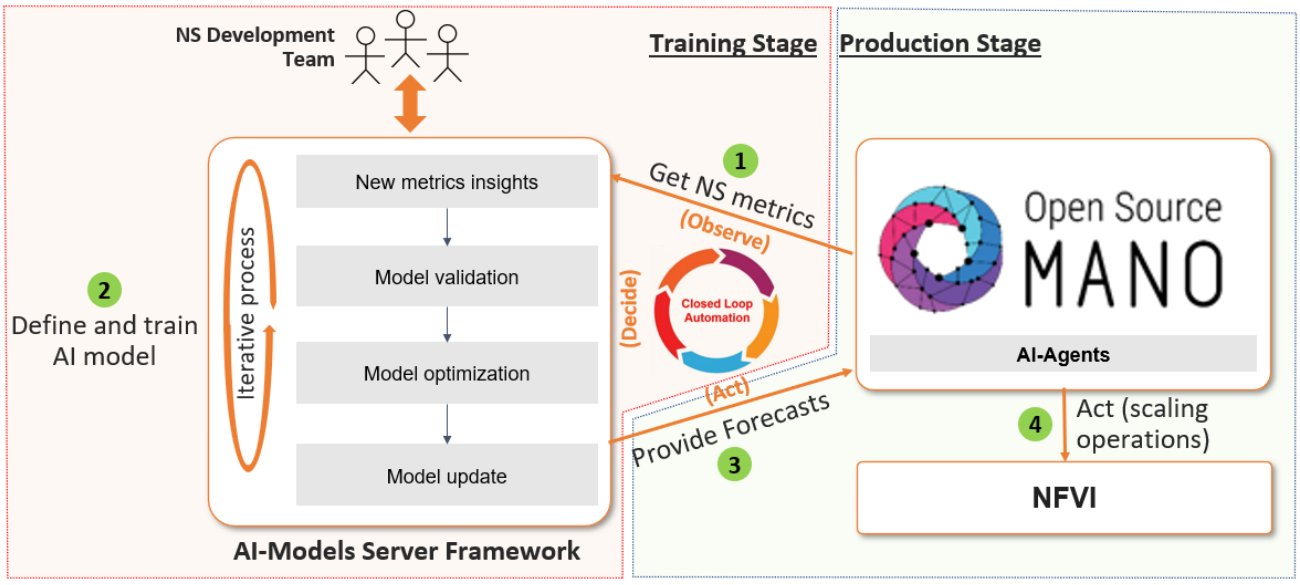


Figure 34. Development model associated to the AI-Agents functionality.

For the development of the AI-Agents functionality in the context of the OSM Open Source community a clear objective was not to replace the existing Service Assurance functionality already available in OSM (which provides VNF scaling actions based on simple threshold rule-based approach), but to be able to facilitate the deployment of the new AI-Agents functionality as a complementary function to provide a more advanced proactive behaviour.

To get this, the implementation of this AI-Agents feature has been scheduled in different stages. In the initial stage (already completed) the target was the deployment of AI-Agents having a minimal impact on the existing

OSM architecture based on the aforementioned Execution Environments feature (and also, to demonstrate the concept to the OSM community and receive their feedback). To get this, the integration of the AI-Agents block was done through the OSM NBI (North Bound Interface), instead of directly targeting the POL module. An official OSM PoC (OSM PoC #11) was done to demonstrate the usage of AI-Agents in a practical ML problem (VNF scaling based on images recognition) according to the progress done in this initial stage. The PoC is publicly available in the OSM ETSI website [37]. AI-Agents can be used in the same way described in this PoC from OSM Release 8.

The plan is to continue with the development of the AI-Agents feature in further stages. The objective would be to get a better alignment with the OSM architecture including a direct interface between the AI-Agent and the POL blocks (through the internal OSM Kafka bus); also, the updating of the OSM CLI and GUI to include the management commands associated to the AI-Agents feature, among other details.

Figure 35 below shows the implementation of the ETSI PoC mentioned before. It can be seen that in this case the external AI Models Server has been implemented using Tensor Flow Serving [38]. Of course, due to our agnostic approach, other solutions could be also suitable for the server (e.g., AcumosAI [39] or even a general-purpose HTTP server based on Flask [40]). Aligned to this, the development environment (top-left) was also based on TensorFlow [41], Keras [42] and Google Colab [43] (of course, other equivalent technologies could be used as well). As we see the AI-Agent is associated to a specific VNF (by means of its VNF Execution Environment) that is able to trigger scaling actions (through the OSM NBI) based on the analysis of different pictures collected by another VNF (Images Server). The AI-Agent normalizes the collected images and send them to the AI-Models server. The AI Models Developer can use these images to design and train the AI model that, once enabled, can answer the AI-Agent queries that make OSM able to perform the VNF scaling actions.

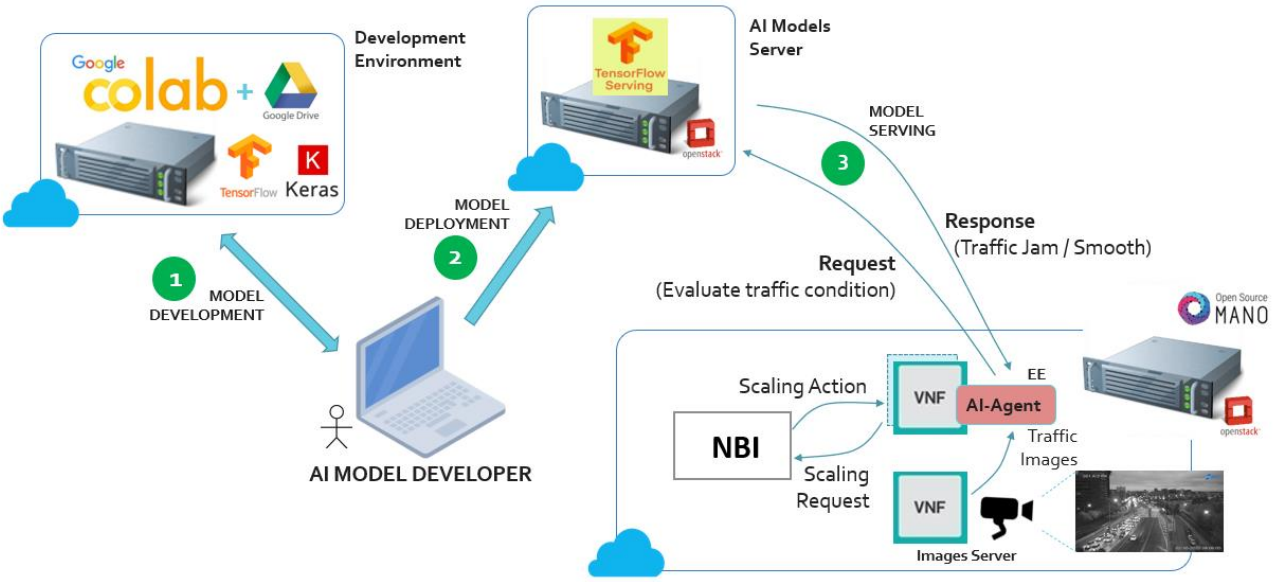


Figure 35. ETSI PoC #11 Implementation.

In the current implementation AI-Agents are deployed as VNF Helm Charts [44]. This because VNF Execution Environments on which they are based are also deployed as Helm Charts. In the following Figure 36 we can see a simplified VNF structure containing an AI-Agent attached to the Execution Environment chart.

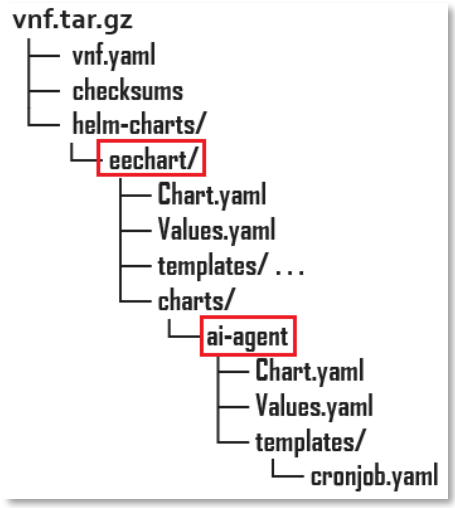


Figure 36. VNF Containing AI-Agent associated to its Execution Environment.

The 'cronjob.yaml' file contains the configuration for the periodic execution of the AI-Agent, defining how often it will be activated to collect data and make queries to the AI-Models Server. In the 'Values.yaml' file associated to the AI-Agent the configuration parameters to connect to the AI-Models Server and the model that can be queried are specified. Figure 37 below shows a basic configuration example which uses a lambda evaluator to process the answer received from the AI-Models Server.

```
# Configuration for AI-Agent
  config:
    version: 1
    # Select the AI Server to request
    AIServer:
      name: TensorFlow
      type: tensorflow
      url: "http://192.168.137.46:8501"
      version: v1
predictions:
      # List of predictions
    - model: A2-traffic-model
      active: True
      monitoring:
        url: 192.168.137.34 # URL, IP or just vdu_name
        port: 3000 # Optional
      threshold:
        function_name: evaluator
        logic: "evaluator = lambda x: True if (x['predictions'][0][0] > x['predictions'][0][1]) else False "
```


Figure 37. VNF Containing AI-Agent associated to its Execution Environment.

5.2.2 Service level assurance monitoring in 5G RAN MANO enhancement

One of the key aspects of 5G systems is the capability not only to efficiently support a variety of Quality-of-Service categories but also to aggregate different service types to offer the so-called Vertical Applications with dedicated communication resources by means of network slices.

In order to ensure customer satisfaction in this multidimensional framework, service assurance must rely on more sophisticated methodologies to be able to:

- combine multi domain observation of both network and user level performance in (near) real-time;
- identify quality degradation causes and adapt resource management policies in a closed loop manner.

Traditional PM (Performance Metrics) counters provide a statistical characterization of the network performance. They are collected by the network functions with a time granularity of minutes (typically 5-15 min). The scope of a PM counter can be the site, the cell, the sector. Despite of the coarse granularity, cell level PM counters allow a network and traffic characterization, besides the identification of trends and periodicity (daily, weekly, monthly basis). Typical KPIs derived from PM counters fall in the following areas: cell average and peak DL/Ul throughput, number of connected users, average radio resource occupation, average DL/UL latency, etc.

Some of the previously mentioned KPIs may be further disaggregated down to UE, QoS Class (QCI) level, bearer type level and, more recently in 5G, to slice level statistics. The additional granularity dimension enables a first level of quality-of-service characterization, even if the statistical nature of the measurement collection does not allow yet a precise correlation with the real user experience.

Modelling PM counters as time series and adopting suitable Machine Learning techniques it is possible to predict traffic peaks or anticipate critical conditions likely to affect user performance. Moreover, the statistical characterization of the network behaviour can be used as baseline for anomaly detection and trigger a specific event analysis or corrective actions while the anomaly is taking place.

On the spatial dimension, clustering techniques may be used to identify similarity among groups of cells or cell types. Even with PM counter and KPI, by adopting spatial and temporal correlation, it is possible to enhance the cellular network traffic models and increase the accuracy of peak predictions and localization of hot spots.

As stated above, the nature of statistical counters/KPIs allows only a partial insight of user experience related network performance. Additional data types like cell/UE traces and events enable a complementary network behaviour modelling and a powerful functionality in support for service-oriented performance assurance.

Cell/UE traces and events are part of Performance Assurance functionalities since 2G/3G but from 5G introduction several architectural and functional improvements will boost the adoption of Trace Management.

Tracing management was originally introduced in support to fault management and trouble shooting and up to now was mainly intended to be activated on a measurement campaign basis, focusing on specific subsystems and interfaces under assessment. The detailed information made available from the network functions by logging interfaces messages and user data transfer has the drawback of huge amount of data produced. This has impacts both on network nodes computational resources to process and log the events and on connectivity resources to transfer the logs from the distributed RAN nodes towards the management systems. Moreover, this approach lacks support for near real time analysis because the activation-collection-transfer process is affected by intrinsic latency.

In the 5G framework, with several network functions implemented in centralized cloud platforms as virtual or containerized cloud functions, the raw data streams produced by the network functions may be consumed, filtered and processed by management functions located in the same platform. Moreover, trace activation can be automated and cloud compute resources may be dimensioned and allocated accordingly in a flexible way.

As previously mentioned, traces and events produced by network functions may be used to derive deeper knowledge about user-level performance. This requires a process of filtering and aggregation of data, with the aim of creating collection of information associated to a data session. Being the time granularity of such record in the order of seconds or more, it is possible to monitor (chunks of) data sessions and correlate the performance metrics described above (e.g., throughput, packet transfer delay) with network status (e.g., radio coverage and load). By means of ML techniques, it would be possible to identify the main causes of service degradation on a specific session/user/service.

Gathering so much information through the assurance process requires effective ways to understand and monitor data, to identify patterns, trends, outliers and similar sort of remarkable features.

Using proper data visualization tools is a good practice to obtain a clearer idea of what the information means, by setting it in visual contexts. This kind of tools also allows strong interactivity and dynamicity in data exploration, interpretation and correlation finding, through a comprehensive system of integrated filters. In Figure 38, an example of a filter panel is represented; it aims to easily explore mobile network cell faults, mixing various types of filtering: temporal (the day line, red-coloured according to the number of faults occurred in

each day), quantitative (fault duration ranges, number of fault occurrences) and qualitative (severity, different dispatching channels, other fault and trouble tickets details).

Figure 38. Example of filter panel.

Generally, many different visualization methods can be applied to create convenient representations: tables, charts (histograms, bar charts, pie charts, line charts, tree charts), maps, heatmaps, and their possible combinations into more integrated and complex info-graphical views, reports, panels, interfaces, dashboards.

Specifically, the multidimensional framework in which service assurance is to be performed offers a considerable variety of multi domain counters, KPIs, PM indicators. Since these data are typically processed and analysed with statistical and Machine Learning techniques, a visual approach can be useful to make algorithm outputs much more evident and explainable.

A typical case is the application of clustering techniques, often utilized to detect groups of cells which have similar behaviours. In this analytics method, a visual insight can be helpful for a smart navigation through clusters: thanks to a good system of filters and interactions, it is possible to compare counters of each group, simulate KPIs scenarios, compare statistics. The large number of cells can also be plotted as time series (e.g., traffic trends) and coloured according to the cluster in which every cell has been placed by the output of algorithms. for a better perception of how clusters are made (in Figure 39groups are depicted in different colours). It also enables to point out the centroids (i.e., the mean profile of each cell group; see Figure 40).

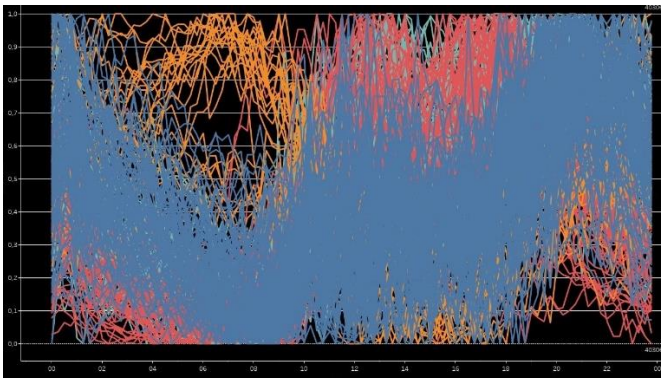


Figure 39. Cell clusters based on daily traffic.

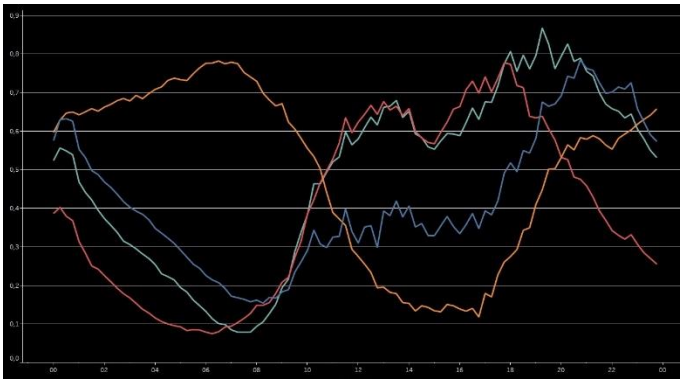


Figure 40. Cluster centroids.

Another common analysis consists of detecting samples and observations which differ significantly from the majority of the data; that is the anomaly detection. Once identified a baseline profile among the whole cells,

other profiles are modelled and compared with the baseline; those samples which exceed a certain threshold will be marked as abnormal.

The visual approach allows here to easily handle different ways of modelling data: while cell counters and KPIs are possibly elaborated without considering timestamp (in order to be optimally processed), it may be useful to see how outputs of the algorithm (normal and abnormal values) map into the actual time series (Figure 41).

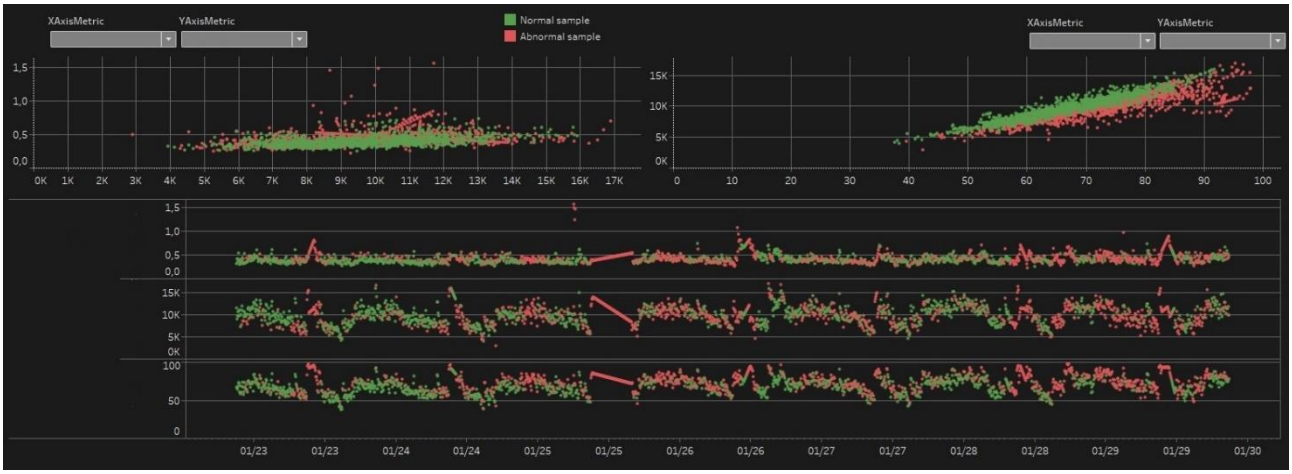


Figure 41. Cell counters modelled in two ways: statistical samples (scatterplots above) and time series.

On top of the analytics processes and visual tools described, from the observation of the current session (and based on the network and service characterization learned from previous observations) it would be possible to activate (near-) real time control actions (automatically or semi-automatically triggered) within the ongoing sessions.

Session level measurements may also be modelled and treated as multi-dimensional time series and some level of prediction can be obtained, in a shorter timescale.

5.3 AI Orchestration

5G-TOURS assumed a significant use of AI mechanisms in orchestration and this applies to both resource management and service management. The need to introduce AI in orchestration arose from the fact that, on the one hand, new virtualisation and slicing technologies open up the possibility of efficient delivery of vertical services with a given level of quality on a shared infrastructure, and, on the other hand, lead to high complexity of the entire system, which increases the requirements for resource management modules, especially the ability to predict changes in the behaviour of the entire system.

In such situations AI, with its feature to learn from the past behaviour of the system and its ability to predict future behaviour is an ideal tool to support decision-making processes. The use of AI modules, of course, raises questions of how to integrate these modules with the existing software infrastructure, how to select specific algorithms and evaluate their effectiveness, how to assess the resources needed for specific functions, and many others.

In the case of the infrastructure developed for 5G-TOURS, the issues of using AI in orchestration were outlined in the first deliverable D3.1 [2] , where chapter 5 described the plans of implementation including AI based data analytics. The topic was further elaborated in the second deliverable D3.2[3], where such technologies like VNF profiling, AI-based capacity forecasting, slicing empowered with Data Analytics, auto ML systems and AI enhanced MANO were identified. For some of them application examples were given and the methodology was described.

In the current document D3.3 the threads developed earlier are not set aside, but more emphasis is placed on standardisation issues, which brings order to the whole topic of using AI in orchestration. The other previously mentioned D3.1 [2] and D3.2 [3] threads will be further elaborated in the next versions of the document.

5.3.1 AI-based Autonomous Control, Management, and Orchestration in 5G: from Standards to Algorithms

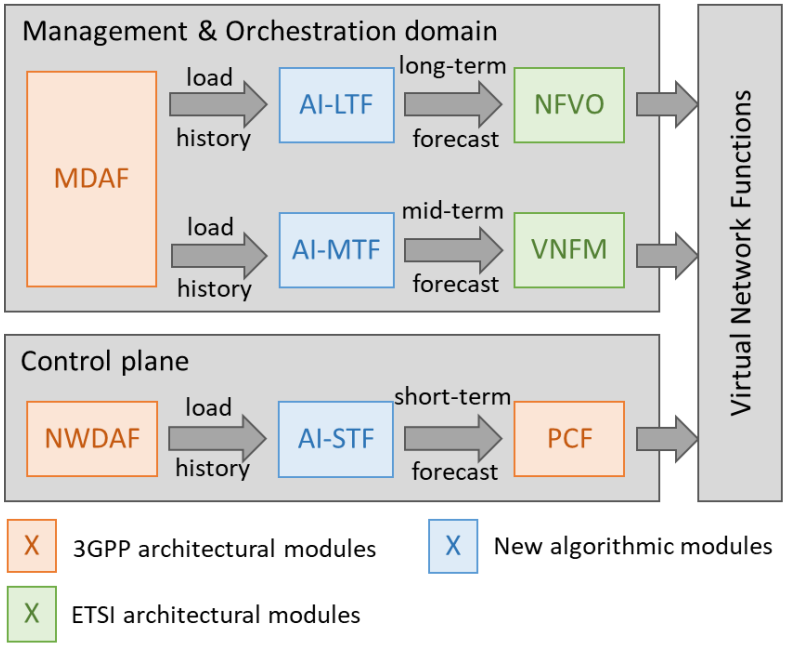


Figure 42. Architectural modules of an AI Framework.

Figure 42 above depicts the network data analytics framework as proposed by major architectural SDOs and also by 5G-TOURS. The framework design encompasses the Management and Orchestration domain as well as the Control domain functionalities, as AI can indeed improve the performance at all levels. Within each domain, we take the architecture proposed by 3GPP as reference, integrating it with the ETSI NFV MANO architecture and expanding it with other architectural proposals such as O-RAN (although this part could be independently managed by the network).

5.3.1.1 Management and Orchestration Domains

In the Management and Orchestration domain, the MDAF module is responsible for the so-called Management Data Analytics Service (MDAS) for all network slice instances, sub-instances and network functions hosted within the network infrastructure. This involves the centralized collection of network data for subsequent publishing to other network management and orchestration modules. In the proposed framework, we specifically employ this service to collect mobile data traffic loads generated in the radio access domain by the individual slices; in particular, the MDAS [45] comprises the load level at both Network Function (NF) and network slice levels, provided as a periodic notification and expressed either in absolute terms or relative to the provisioned capacity.

As a result, the MDAF allows building historical databases of the network demands for each base station and slice. These data are then exposed to the AI-based prediction algorithms for long-term forecasting (AI-LTF), and mid-term forecasting (AI-MTF).

The AI-LTF algorithm aims at assisting the VNF placement decisions taken by the orchestration system. To this end, AI-LTF leverages the network demand history to predict the future aggregate load across the different infrastructure locations. Then, the NFV Orchestrator (NFVO) compares such a prediction against the current available capacity in each infrastructure location and anticipates potential overload conditions. Subsequently, the NFVO can react, e.g., by moving VNFs out of the congested infrastructure (while meeting the requirements of the corresponding network slice). The AI-LTF algorithm operates on long timescales, typically in the order of hours: indeed, VNFs repositioning is quite a drastic action that involves substantial overhead, and consequently it is only performed occasionally and as an answer to substantial traffic fluctuations.

The second algorithm, AI-MTF, has a different purpose: it fuels the resource scaling decisions taken by the VNF Manager (VNFM). The VNFM has an interface with the Virtual Infrastructure Managers (VIMs) to monitor the resource usage of the VNFs of each slice, and it also leverages data collected and published by the MDAF to determine the level of unsatisfied demand and the number of unused resources.

Based on all this information, the AI-MTF algorithm assists the orchestration framework on the decision:

- to provide more resources to the VNFs of a slice when the predicted load exceeds the current resources, an operation typically referred to as upscaling, or
- to downscale resources to save cost when VNFs are leaving a significant fraction of the resources unused.

Such decisions must be taken over faster timescales than those affecting the VNF placement, and generally occur over intervals in the order of tens of minutes, which is the typical frequency for the execution of new VNF instances involving up- and downscaling.

It is worth noting that AI-LTF and AI-MTF only take as input the load history from MDAF and do not interact between themselves or with any other module. The forecasts of AI-LTF and AI-MTF are fed into the NFVO and VNFM engines, which may instead also leverage information obtained from other modules to take their decisions.

5.3.1.2 Network Functions control plane

On the control plane, the NWDAF module is responsible for collecting data on the load level of a NF or a network slice [46], playing a very similar role to that of the MDAF in the management domain. In our framework, these data are fed to the AI-based short-term forecasting algorithm (AI-STF), which predicts the future traffic load. The forecast is leveraged by the Policy Control Function (PCF) module, which provides a unified policy framework to govern the network behaviour. PCF can use the forecast provided by AI-STF to optimize its policies, such as:

- the QoS parameters (for those services that can be provided at different QoS levels),
- the access and mobility policies, or
- the UE Route Selection Policy (URSP).

In contrast to the previous modules, these updates are performed at rather fast timescales, down to hundreds of milliseconds.

While the NWDAF module has been designed for the network core, a similar approach can be applied to the radio access network (RAN). Although 3GPP has not yet proposed modules equivalent to NWDAF in the RAN, other initiatives such as the O-RAN alliance have taken this path. In the ORAN architecture [47], the Radio Network Information Base (RNIB) collects load information of flows or flow aggregates at the RAN level, the RAN Intelligent Controller (RIC) enables near real-time control of RAN elements/resources, and the RAN resource orchestrator handles the overall resources at the base station level. In this case, the AI-STF forecasts can be leveraged by the RIC to perform the optimization of the radio resources at a fine time granularity (in the order of hundreds of ms) and by the RAN resource orchestration to update the resource and bandwidth allocation at larger timescales (up to the order of minutes).

5.3.1.3 AI-based algorithms discussion and design

The above framework introduces three new AI-based algorithms: AI-LTF, AI-MTF and AI-STF. The three algorithms follow the same design guidelines, as all of them aim at providing network capacity forecasts. The main difference between them is that they work at different granularity in terms of traffic volume (at global, slice, or flow levels) and timescale (intervals of hours, tens of minutes, minutes or shorter). In the following, we present the unified design of these three algorithms.

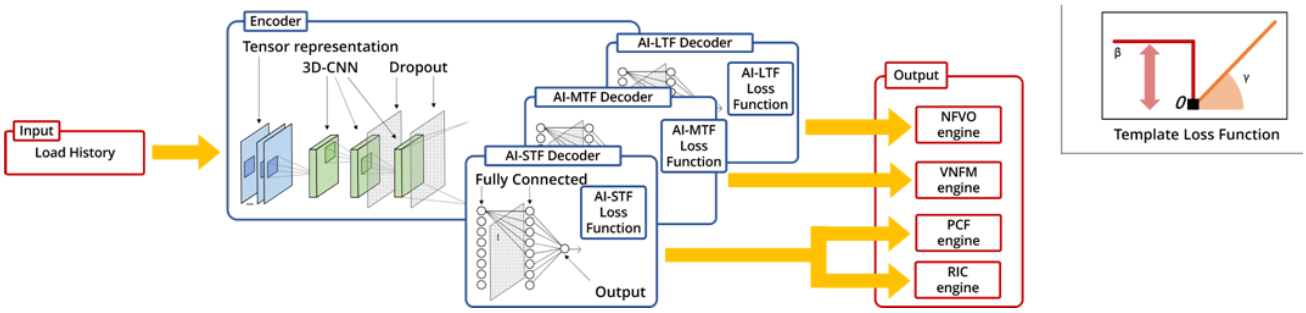


Figure 43. The proposed Deep Learning Structure.

The neural network architecture used by the proposed algorithms is summarized in Figure 43, and is composed of an encoder-decoder sequence. The encoder is composed of a stack of three dimensional Convolutional Neural Network (3D-CNN), while the decoder uses Multi-Layer Perceptrons (MLPs), a class of fully-connected neural layers where every neuron of one layer is connected to every neuron of the next layer. MLPs are able to learn global patterns in the input feature space, which allows forecasting the target capacity leveraging the local features extracted by the encoder. As already discussed in [3], general-purpose loss functions like Mean Square Error (MSE) or Mean Absolute Error (MAE) are clearly inappropriate to optimize the operator running cost, thus we propose the α-OMC [48] loss function, to properly optimize this factor.

We evaluated the system with three specific algorithms that populate the AI-LTF, AI-MTF and AI-STF, namely:

- AI-LTF: Long-term forecasting for VNF placement: this algorithm takes care of computing the exact placement of VNFs according to the available capacity at any point in time.
- AI-MTF: mid-term forecasting for NFVI scaling, which is the algorithm implemented in the ETSI ENI PoC (discussed in Section 0).
- AI-STF: short term forecasting for QoS policies, which reacts at faster timing to set QoS parameters for network flows.

While the interested reader can find the full evaluation in [49], we selected here the results related to AI-LTF. The long-term forecasting capabilities provided by the AI-LTF algorithm are useful to make decisions about the suitable placement of the VNFs serving one or more slices. To evaluate its performance, we consider a scenario where a datacentre with processing capacity C serves the seven slices and assume that the computational demand of a given slice is proportional to the number of transmitted bytes.

In this case study, we use a decision-taking interval equal to 8 hours to account for the fact that VNF placement decisions are typically taken with a coarse time granularity of hours due to the limitation of the underlying NFV technology. We focus on an edge network datacentre and employ AI-LTF to support the VNF placement decisions taken by the NFVO module by anticipating the overall traffic load at the target datacentre. Then, the NFVO can decide at every decision interval how many slices are served by the datacentre of capacity C, and which slices shall instead be placed elsewhere.

Figure 44 below depicts the result obtained with AI-LTF against that obtained with an Oracle algorithm that assists the NFVO with the knowledge of the real future demand (such an oracle algorithm is unfeasible in practice but provides an optimal benchmark to assess AI-LTF's performance). Figure 44 depicts the occupation ratio (top) and number of admitted slices (bottom) for each 8-hour orchestration period. The algorithm implemented by the AI-LTF module is compared against an optimal but unfeasible Oracle solution with perfect knowledge of the future traffic load. We observe that AI-LTF follows quite closely the oracle. The overall usage of the deployed infrastructure remains high at all times. The algorithm only moves more slices than needed away from the datacentre on very limited occasions. In rare cases, it places more slices than it should in the datacentre, leading to an overload situation that results in computational outages for the served slices; however, even when this happens, the actual overload levels are negligible. These results confirm that AI-LTF is a promising solution to assist effective VNF placement decisions.

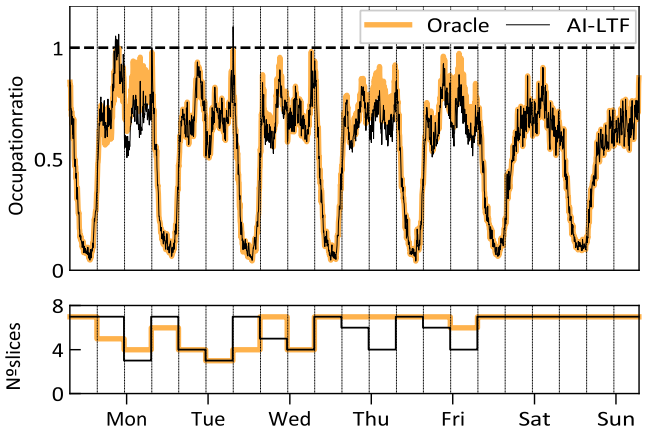


Figure 44. VNF placement of slices at one target datacentre.

5.3.1.4 PoC - AI approach in resource forecasting for 5GC

One example of the innovative AI use in predicting resource demand, being developed within 5G-TOURS activities, is the one preventing overloading of AMF server located on the Control Plane of the 5GC. This aspect is relevant when AMF server handles IoT traffic therefore necessary in 5G-TOURS, where a number of UCs (6,10) assume the use of IoT technology. IoT devices can generate large volumes of UE requests on the CP plane and AMF can be the bottleneck here.

The second innovation of this case is related to the use of CNF functions on the CP plane instead of VNF, which has become a new trend among telecommunication companies in the last few years. Usually User Plane and Data Plane, as more latency sensitive, are implemented as VM and CP is migrated as cloud native using containers and microservices. In 5G-TOURS, this trend is evident in the French site, where the WEF 2.2 5GC network is deployed as CNFs on a Kubernetes cluster in the BCOM data centre. At the same time, the data plane of the 5G network will be deployed as VNFs. The use of CNFs for the 5GC network does facilitate rapid scaling but the AMF/MME server can still be the bottleneck of the mobile network control plane, while it has to handle a huge number of UE requests.

The proposed Proof of Concept solution implements an AI model with various algorithms to predict the size of resources such as CPU and RAM of AMF server deployed as CNF. This gives a chance to optimally allocate resources in the network. According to the decision of the AI model, there will be scaling in/out of the CNF to handle future traffic flows. A schematic of this concept is shown below in the Figure 45.

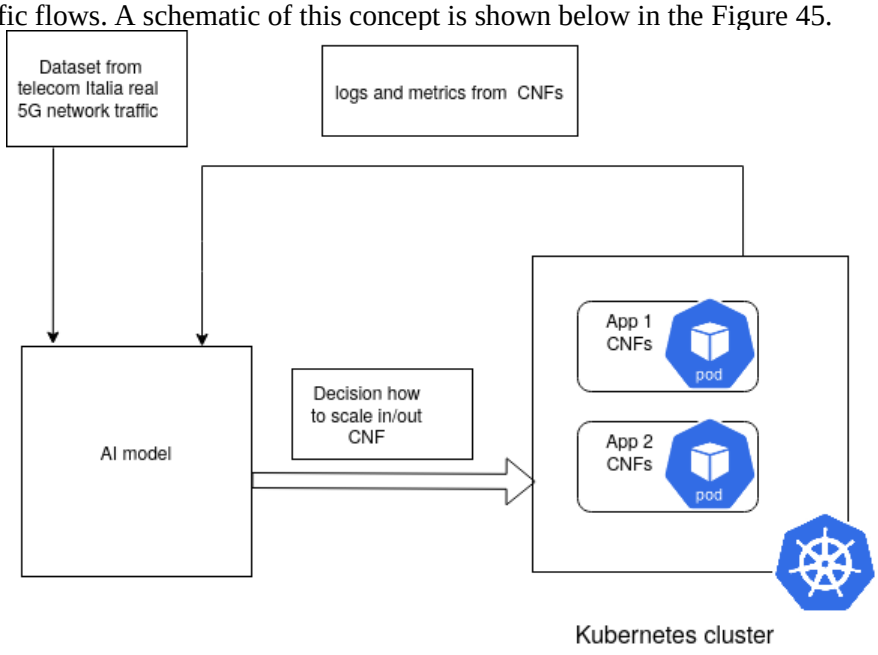


Figure 45. 5GC resource forecasting solution based on AI model.

AI model will be trained on mobile network traffic datasets. For the time being it is assumed to use the dataset collected during the Big Data Challenge organized by Telecom Italia [6]. The data set is Open Source and rich in information. It can be used to predict the evolution of the load in a core network.

Currently the data from 5G-TOURS is not available, as the use cases are still under deployments. Moreover, 5G-TOURS UCs will be used only by a few UEs, do not have KPIs connected with the number of UE requests and do not fit the occurrence when AMF server is overload.

The second input for the AI model will be CNFs logs including the actual number of requests handled by AMF server. The AI algorithm will return as output the decision on how CNF with AMF server should be scaled out/scaled in. There are developed scripts triggering rescaling CNFs (Pods) by Kubernetes cluster API.

Described solution shows the methodology of incorporating AI model in forecasting resources of 5GC network deployed in new manner as CNF. Potential integration with the ONAP orchestrator, which seems natural, must wait until ONAP platform supports CNF rescaling.

Hopefully the next version of the deliverable D3.4, due in December 2021, will showcase concrete results assessing the validity of the assumptions made and the effectiveness of the prediction methods used.

5.4 Broadcast Support

The ability to provide multicast/broadcast communications is viewed as a basic requirement in 5G systems. Enabling point to multipoint transmissions in the 5GC increases the efficiency in the network resources used, avoiding possible congestion occurring inside the transport network. In the scope of 5G-TOURS, Task 3.2 Broadcast support aims to devise technology that will be used to deliver high-quality multimedia content distribution, both using the LTE-based 5G Broadcast and the 5G-Native variants. The outcomes of this task will be performed in two trials inside Use Case 4: UC4.b and UC4.c [1]. The first trial uses state-of-the-art Rel-16 LTE equipment and the second one will use beyond state-of-the-art technology via software prototyping and simulated RAN environments, adapting the most recent Rel-17 3GPP standardization work. This section details the technologies used in both trials and its current state of implementation.

5.4.1 LTE-based 5G Broadcast

UC4.b focuses on the transmission of high-quality video distribution using broadcast delivery and it is further detailed in [1]. The content will be distributed using RAI's broadcasting network, with a HPHT (High Power High Tower) topology, to all users at once offering constant performance no matter the number of devices connected to the network. 5G-Broadcast LTE-Based is a DTT technology, similar to DVB-T or DVB-T2, that broadcasts TV content to an unlimited number of users. For this reason, the improvements proposed in Release 16 (from which 5G Broadcast LTE-Based arises) are based on some enhancements of the physical layer for an improvement in reception in high-speed mobility conditions and for receptions at greater distances (high coverage).

In a wide coverage situation (also known as Inter Site Distance), of up to 100km, a high-power infrastructure is needed, prompting the use of HPHT infrastructure, while the current infrastructure, such as the LPLT used in LTE and 5G mobile networks is unfeasible. In fact, 5G Broadcast LTE-Based can be considered as a complement to DVB, being less efficient than, for example, DVB-T2 in fixed reception conditions (for Rooftop reception) but reaching a greater number of users/devices (mobiles, tablets and in general devices with a certain mobility), further details about this comparison can be found in [15]. In short, the HPHT infrastructure is not only exclusive to DVB, but it can also be used by other technologies, such as 5G-Broadcast, both being complementary.

For that reason, the RAI HPHT infrastructure was proposed to be used in this sub-use case. This use case is split in two phases:

- Phase 1: live or pre-recorded video distribution, using Rel-14 FeMBMS, received in RAI production center premises or Palazzo Madama and transmitted via Wi-Fi multicast to all users using a broadcast tower.

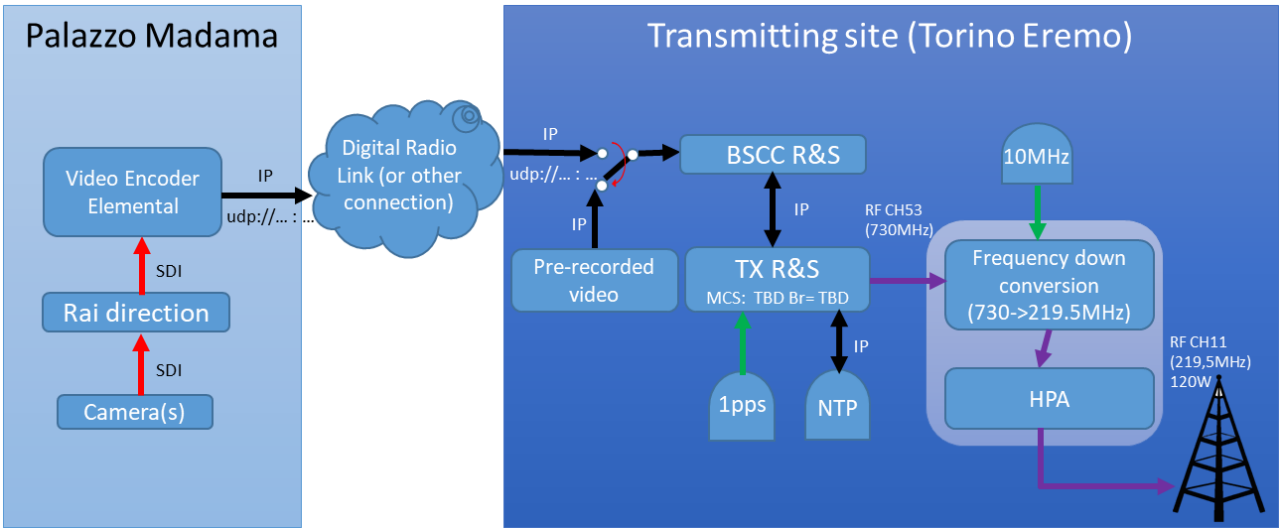


Figure 46. Phase 1 trials' configuration.

- Phase 2: live or pre-recorded video distribution using 5G Broadcast Rel-16 updated equipment in two different scenarios. The first scenario is located at Palazzo Madama to test larger coverages. The second one is an in-car scenario to test higher speed mobility.

The difficulties derived from the COVID-19 impact caused the adaptation of phase 1 trials which was validate through laboratory trials instead.

5.4.2 5GC Multicast

In contrast with UC4.b LTE-based 5G multicast implementation, UC4.c focus in the implementation of a 5G-Native Multicast Core, using 5G technology in both RAN and Core parts. The beyond state-of-the-art scope of this use case implied a lack of a 3GPP standardized mark to aligned to at the start of the implementation. The start of the development, as explained in previous deliverable [2], was scoping the implementation of NF out of 3GPP standardization mark. This innovative research will be tested in laboratory premises using professional Core testing tools, as there is no Rel-17 compliant commercial equipment. Software implementation will be explained later in this section.

3GPP standardisation

Currently, architectural enhancements for 5G multicast/broadcast are being standardized by 3GPP. 3GPP research is being approach by two different angles, one of them considering multicast/broadcast communication as a Resource optimization service over Unicast systems (Operator perspective) and the second one considering multicast/broadcast as a service itself (broadcaster perspective). The first approach is being studied inside a Study Item named “Study on architectural enhancements for 5G multicast-broadcast services” ended in December 2020. This Study Item aims to enable general MBS over 5G Systems specially targeting Transparent IPV4/IPV6 multicast delivery, V2X, Public Safety, IPTV and IoT use cases. The outcomes related to this item are reflected in TR 23.757 [3] and concluded with the need to include multicast functionalities inside the 5GC. For that reason, a new Work Item was approved named "Multicast-broadcast services in 5G”.

The architecture for 5G-multicast in the 5G-core is specified by SA2 in TR 23.757 (3GPP, TR 23.757: Study on architectural enhancements for 5G multicast-broadcast services) in annex A.3. This architecture illustrated in Figure 47 captures the 3GPP SA2 working group agreements.

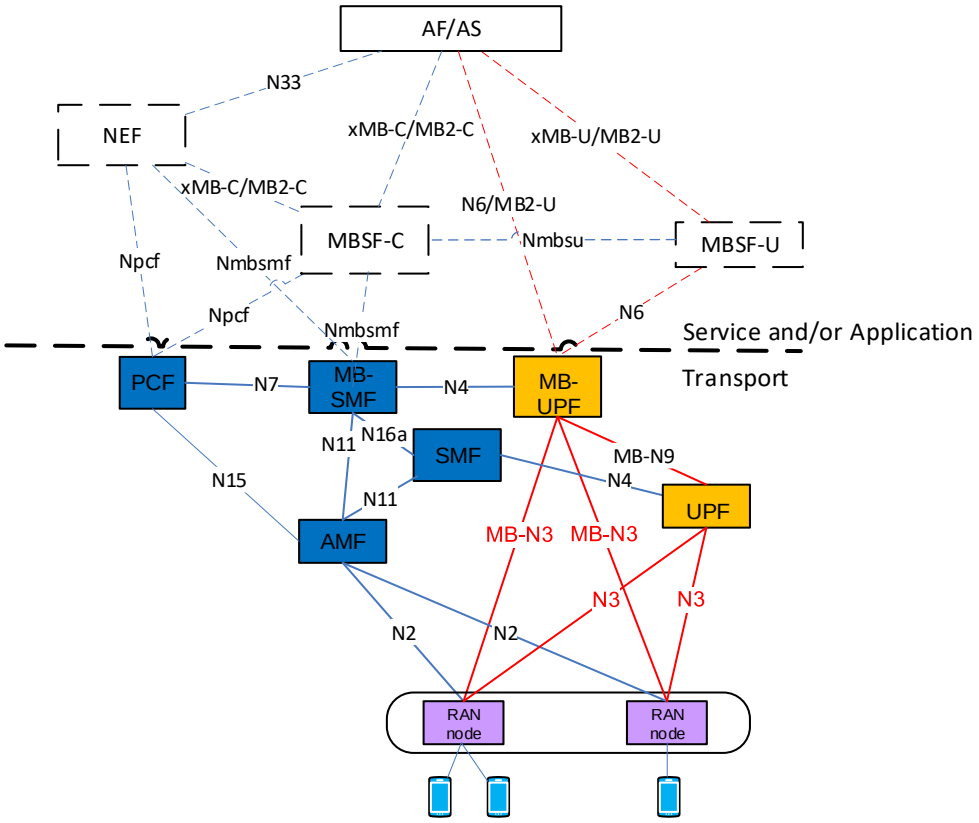


Figure 47. Reference architecture for 5G multicast/broadcast services.

2 new entities are defined in the 5GCore transport layer:

- **Multicast Broadcast SMF (MB-SMF)**: the MB-SMF is used for session management (including QoS control), and control of multicast transport, including configuring the MB-UPF and RAN (via AMF) for multicast/broadcast flows transport based on the policy rules for MBS services from PCF or local policy.

- **Multicast Broadcast UPF (MB-UPF)**: The MB-UPF is used for delivery of MBS flows to RAN (or UPF for individual delivery) and QoS enforcement for MBS services. The MB-UPF performs the following functions to support MBS:

- Packet filtering of incoming downlink packets for MBS flows.
- Distribution of MBS data packets to RAN nodes (or UPF nodes).
- QoS enforcement and counting/reporting based on existing means.

The 3GPP SA4 working group is currently studying the evolution of the 5G Media Streaming architecture (5GSM) within TR 26.802 (3GPP, TR 26.802: 5G Multimedia Streaming (5GMS); Multicast architecture), and is considering in particular the delivery stack (delivery protocol over UDP, signalling/service announcement, forward error correction, etc.). We can expect the same delivery stack as the one offered by the BM-SC for Multimedia Broadcast Multicast Service (MBMS) in LTE.

The 3GPP SA4 working group also considers integration and convergence with the multicast adaptive bitrate (ABR) capabilities lead by Digital Video Broadcasting (DVB).

Developing process

Software development is aligned with the ongoing 3GPP work, specifically with second approach and the Study Item Multicast Architecture Enhancements for 5G Media Streaming as, this project supports the idea of broadcast as a service and will aim for the architecture proposed by 3GPP. The architecture proposed in TR 26.802 can be seen in **Error! Reference source not found.**:

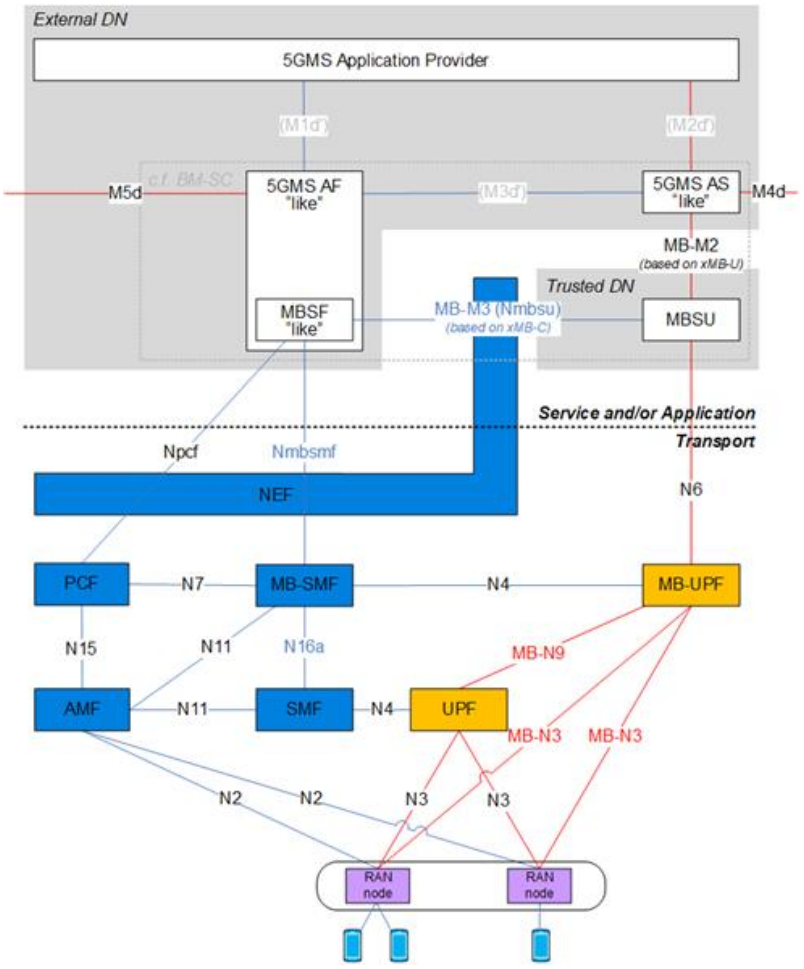


Figure 48. TR 26.802 proposed architecture [4].

The newly introduced Network Functions are:

- MBSF (Multicast Broadcast Service Function). The main functionalities include:
  - Interacting with AF and MB-SMF for MBS session operation and transport.
  - Perform UE authorization to join MB sessions.
  - Management of the MBSU.
- MBSU (Multicast Broadcast Service User Plane). This new component performs:
  - Modification of MBS data.
  - Media anchor in MBS data traffic.
- MB-SMF is an SMF enhanced to handle MB sessions. The main functionalities of this NF are:
  - Network resource management of the sessions.
  - QoS management of the sessions using PCF.
  - Delivery of MB Session information upon AMF requests.
- MB-UPF is an UPF enhanced to handle MB sessions. Its functionalities include:
  - Delivery and retrieval of content.
  - Sending of IP Multicast data to RAN using MB-N3 interface.

The capabilities of MB-SMF and MB-UPF include simultaneous handle of both PDU and MB sessions. However, it is possible to deploy and configure them to exclusively handle MB sessions

UPV is implementing an architecture using TR 26.802 [5] model adapted to 5G-TOURS. The adapted architecture's graph is shown in **Error! Reference source not found.**. In this figure, green blocks highlight the components UPV is currently developing.

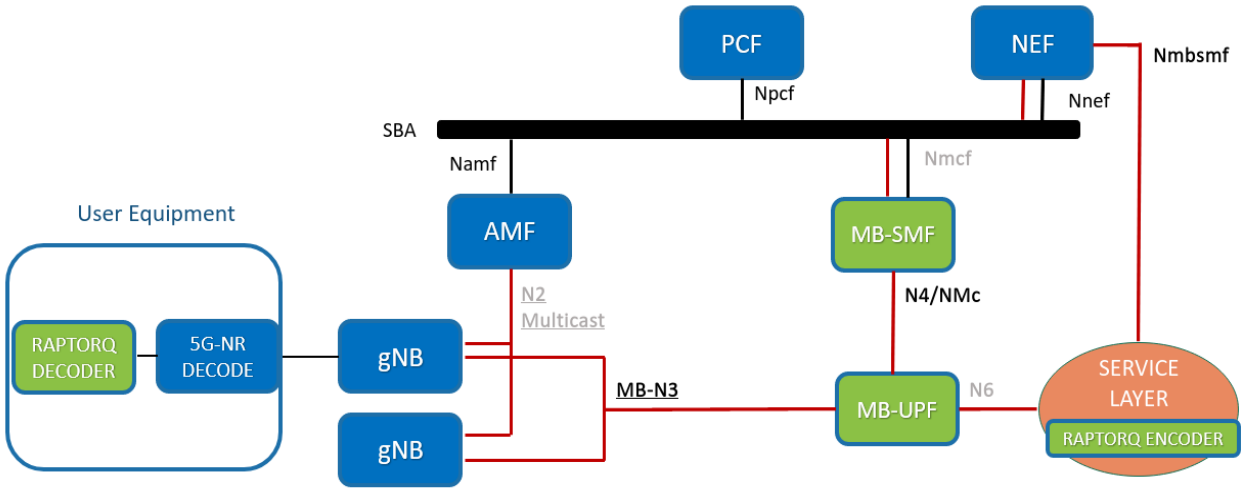


Figure 49. 5G multicast architecture developed in UPV premises.

The software implementation is ongoing and expected to be finished by the end of 2021. The system will be developed into UPV campus premises using the open-source software Open5GCore. Open5GCore is a practical implementation of 3GPP 5GC whose most recent version is Rel-16 compliant. The software-based 5GC will be enhanced with the proper network functions in order to allow multicast capabilities. Open5GCore will be connected and tested with simulated multicast RAN environment, if no commercial equipment is available by the end of 2021.

In addition, multicast 5GC system will include an application layer forward error correction (AL-FEC) technology to improve the transmission protection. The technology used will be Raptor codes latest version, RaptorQ. The combined solution enables scalable and efficient data delivery even in the most challenging environments. Further information about broadcast support and RaptorQ integration will be detailed in section 0

5.5 Service Layer

The service layer aims at providing easy-to-manage interfaces to establish and operate network slices for various industrial verticals such as multimedia, e-health, transportation and robotics by creating an easier interface that enables (when not already supported by 5G-EVE) network slice instantiation, monitoring, orchestration and operation. The service layer also handles new network functions dedicated to the support of broadcast/multicast in 5G.

An analysis of the expectations of verticals for the service layer was performed in D3.2 [3] by collaborating with WP 2 and by circulating a questionnaire to the target UC owners. The expressed requirements were broken out into three categories: Network Slice Management, Slice Monitoring, and Network Exposure. It appeared that these requirements induce a more complex management system than the Management and Orchestration infrastructure inherited from 5G EVE D1.3 [11], driving the innovations for enhanced MANO and AI orchestration presented in sections 5.2 and 5.3. The service layer design and implementation being highly dependent of the MANO technologies, several tailored implementations had to be launched, each associated to specific MANO innovations and specific use case requirements.

As discussed already in Section 3, the 5G-TOURS project partners are implementing an integration and insertion strategy into 5G-EVE of the novel technologies developed in the project. In this section, instead, we provide more details on the specific technologies. The functionalities of the service layers implemented in 5G-TOURS are the following:

- AI-enhanced MANO (section 5.5.1);
- Service layer for the vertical closed-loop integration (section 0);
- OSM AI-Agents (section 5.5.3);
- multicast/broadcast support (section 0).

Some of them will be released as open source, composing an SDK for the Service Layer concept devised by 5G-TOURS.

Table 7 details how the vertical requirements for the service layer, gathered in D3.2 [3], are addressed by these implementations.

Table 7. Supported service layer requirements per implementation

Require-ment	Description	AI-en-hanced MANO	ETSI ENI PoC	OSM AI-Agents	multicast/ broadcast support
NSM.1	SLA management. Vertical users can negotiate the QoS considering the cost of certain network KPIs, QoS settings/configurations. (UC5, UC7) Finally, the service interface could give suggestions to the application (and optionally do auto configuration itself) of the remaining network KPIs for a given target: e.g., to maximize performance, minimize costs, or maximize performance for a certain allowable cost.		Y		
NSM.2	Usage of AI Orchestration to find the best possible system orchestration (i.e., function allocation, network slices/devices/resource configuration, etc.) based on the vertical service requirements, and the anticipated network status (UC10).		Y	Y	
NSM.3	The vertical requirements can also be set by an underlying API (UC7).	Provided by 5G EVE			
NSM.4	Service layer and AI orchestration can make an opportunistic use of network re-sources (UC7) Note from ATOS. It applies to UC1 as well.		Y	Y (via AI Agents reading resources metrics)	
NSM.5	Service layer and AI orchestration can pilot the migration of VNFs to the edge or core, possibly dynamically (UC7)				
NSM.6	Vertical users can dynamically change the QoS requirements and the configuration	Y	Y		
NSMo.1	KPI validation/QoS validation: the service layer shall ensure that the service requirements (including the QoS) are continuously met. (All UCs)	Provided by 5G EVE			
NSMo.2	Performance diagnosis: in case of poor performance results, a diagnosis tool can be triggered responsible to collect more measurements and/or make a deeper analysis of the performance issues.	Y			

NSMo.3	The service layer shall fire events to the application in case network KPIs are getting below a threshold	Y			
NSMo.4	Output of the diagnosis tool is presented to the vertical over a GUI, proposing alternatives for improved performance.	Y			
NE.1	the service layer shall expose the device location to the vertical users (UC10).	Y			
NE.2	the service layer shall monitor and expose the device battery level and charging state (UC10).	Y			
NE.3	broadcast capability exposure. The service layer shall expose a northbound API to the vertical allowing the allocation of network resource for broadcast, and a corresponding API for ingestion (UC4).				Y

5.5.1 AI-enhanced MANO

The 5G-TOURS Greek site architecture utilises the tools designed initially by 5G EVE project, improving on them with new innovations that will be presented below. The Greek site architecture consists of a set of interconnected components as illustrated in Figure 50. The Inter-Working Layer (IWL) is connected to the site's orchestration infrastructure via the Multi-Site Network Orchestrator (MSNO). The Greek site's orchestration infrastructure introduces an AI-enhanced Management and Orchestration (MANO) component, specifically AI algorithms work together with the Open Source MANO (OSM) to provide intelligence to the system. In addition, a Virtual Infrastructure Manager (VIM) (i.e., Openstack) and an open-source stream-processing platform (i.e., Kafka Cluster) are also available for deploying VNFs and capturing data respectively. The Runtime Configurator at the IWL is responsible for any last-minute configuration of the VNFs that are deployed with the help of the AI and OSM at VIM. Also, a Central Kafka broker and a Monitoring module are included in the IWL to monitor the deployed VNFs and gather the required data for further analysis. These data can be used by the Performance Diagnosis tool, which has been developed based on data analytics processes to guarantee the full exploitation of the test results and enable the acquisition of in-depth insights per vertical category and use case. The AI-enhanced MANO component also acquires information from VNFs, applications, infrastructure and receive reports that the diagnostic component creates periodically. Finally, the main 5G EVE Portal is involved through the onboarding procedure, by stating various information in specific forms allows the design, creation, scheduling and deployment of vertical network services (NS).

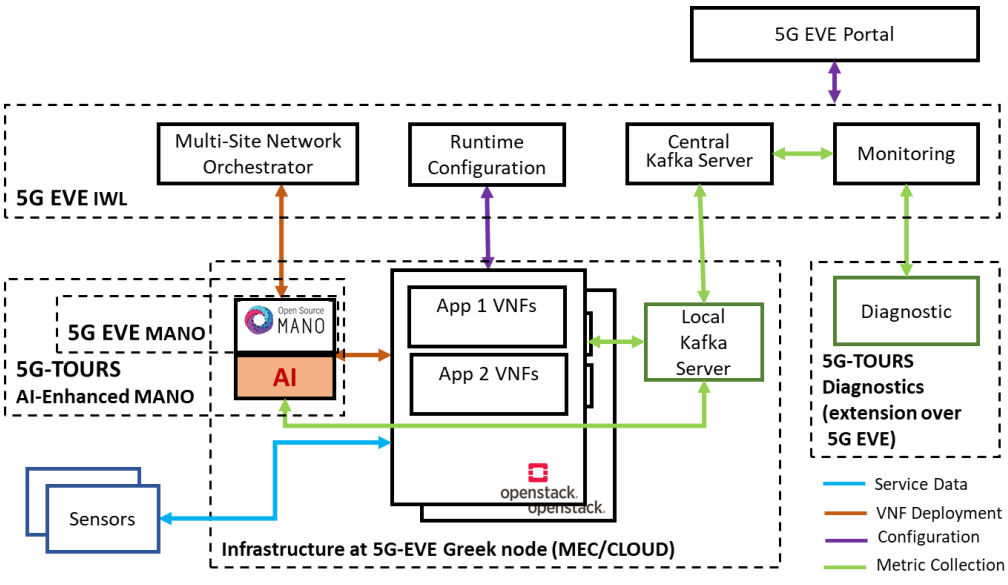


Figure 50. Greek site architecture.

For an experiment to be executed on the 5G-TOURS platform the necessary blueprints need to be onboarded on the service. The steps of the blueprint onboarding process are:

- Vertical Service Blueprint (VSB) and Network Service Descriptor (NSD) onboarding
- Test Case Blueprint onboarding
- Experiment Blueprint creation
- Experiment descriptor creation
- Experiment scheduling and execution

The AI-enhanced MANO will orchestrate and Manage the NS that will be deployed as VNFs at VIM. When all steps are completed, the experiment will run for the allocated time slot.

The 5G-TOURS Service Layer is designed to provide the tenants an intuitive & easy-to-manage interface to establish and operate their slices and services.

- Translate service requests into corresponding network orchestration primitives.
- Facilitate the provisioning of customized network slices and their monitoring.
- Re-use the APIs provided by the 5G EVE platform whenever possible.
- Aligned with 3GPP TS28.533 [12] Exposure Governance Management Function (EGMF).

Following the analysis of requirements from use case owners, a hybrid approach for Service Layer was adopted which consists of two streams. The first one is fully integrated with the 5G EVE infrastructure as already explained, where all information, actions and reports follow the basic functions of the infrastructure, enhancing it with new components. For the seconds stream, our service layer implements a direct connection, connecting directly the verticals to the designed for 5G-TOURS specific use cases components as Figure 51 illustrates.

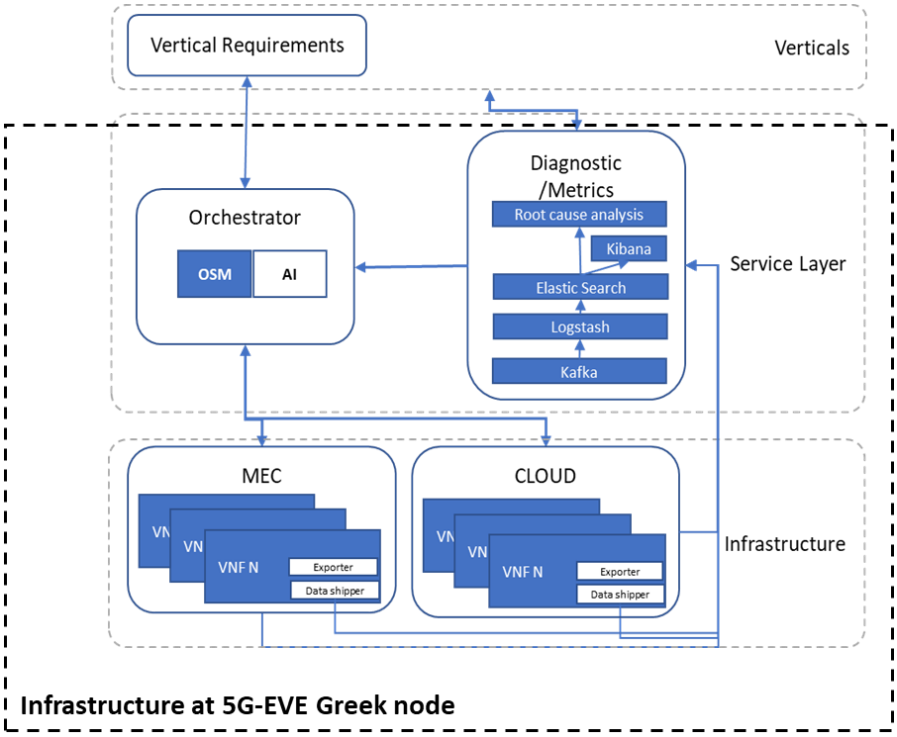
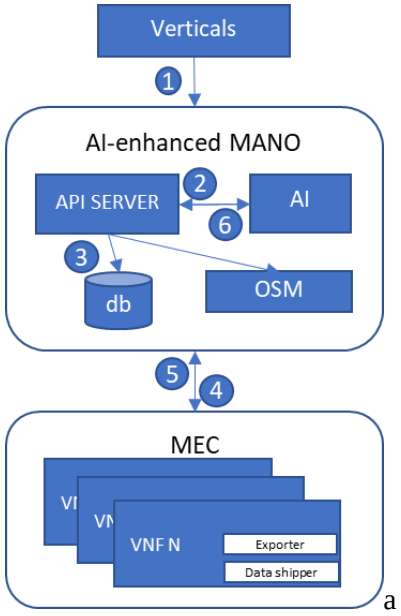


Figure 51. Service Layer at the Greek Node.

To enable the intelligence at the orchestration and management level, multiple interfaces with different components are used. Specifically, connection with Open-Source MANO has been enabled through the OSM's Northbound API featuring ETSI NFV SOL005. Communication with OpenStack has been done through its provided API with a RESTful web service endpoint to access, provision, allocate and automate the resources. Data and

logs are retrieved directly from Elasticsearch REST API, that is also used in cases of configuration and access to other Elasticsearch features. Additionally, custom APIs have been designed and developed for enabling the acquisition of information from VNFs, applications and infrastructure with the use of exporters that send specific metrics to the Elasticsearch and Diagnostic components. Finally, the connection with Verticals has been performed with a custom API that directly connects the AI-enhanced MANO with the Verticals retrieving in this way all requirements or specific SLAs as seen also in Figure 52. Figure 53 illustrates the flow chart of the AI-enhanced MANO. When designing the experiment using the 5G EVE Portal, metrics and Key Performance Indicators (KPIs) can also be inserted in the context blueprint that can be used for better monitoring of the system and the creation of reports at the end of each experiment, additionally, all information can be retrieved by 5G EVE APIs.



a

Figure 52 AI-enhanced MANO architecture

The architecture of the AI-component is shown at Figure 52 and its connections and functions are explained below.

1. Direct Verticals info (i.e., Metrics/KPIs) of the NS that will be deployed acquired from AI-enhanced MANO or by utilizing the 5G EVE APIs.
2. Internally an API SERVER gets the request and triggers the AI component to evaluate the optimal deployment of the VNFs based on Vertical requirements and KPIs.
3. The API SERVER gets the VNF's info from the OSM (OSM NB API) and historical data from db.
4. The API SERVER retrieves the metrics exported from the MEC platform and the MEC's VNF instances, to evaluate the metrics in real time.
5. The AI component acquires all information in real time (MEC's metrics, VNF IDs, and VNF requirements) from the API SERVER and uses old data and previous decisions. The AI finds the optimal deployments and migrations for the NSs in MECs.
6. The AI component triggers the OSM (OSM NB API), through API SERVER, for any migration or reallocation of resources.

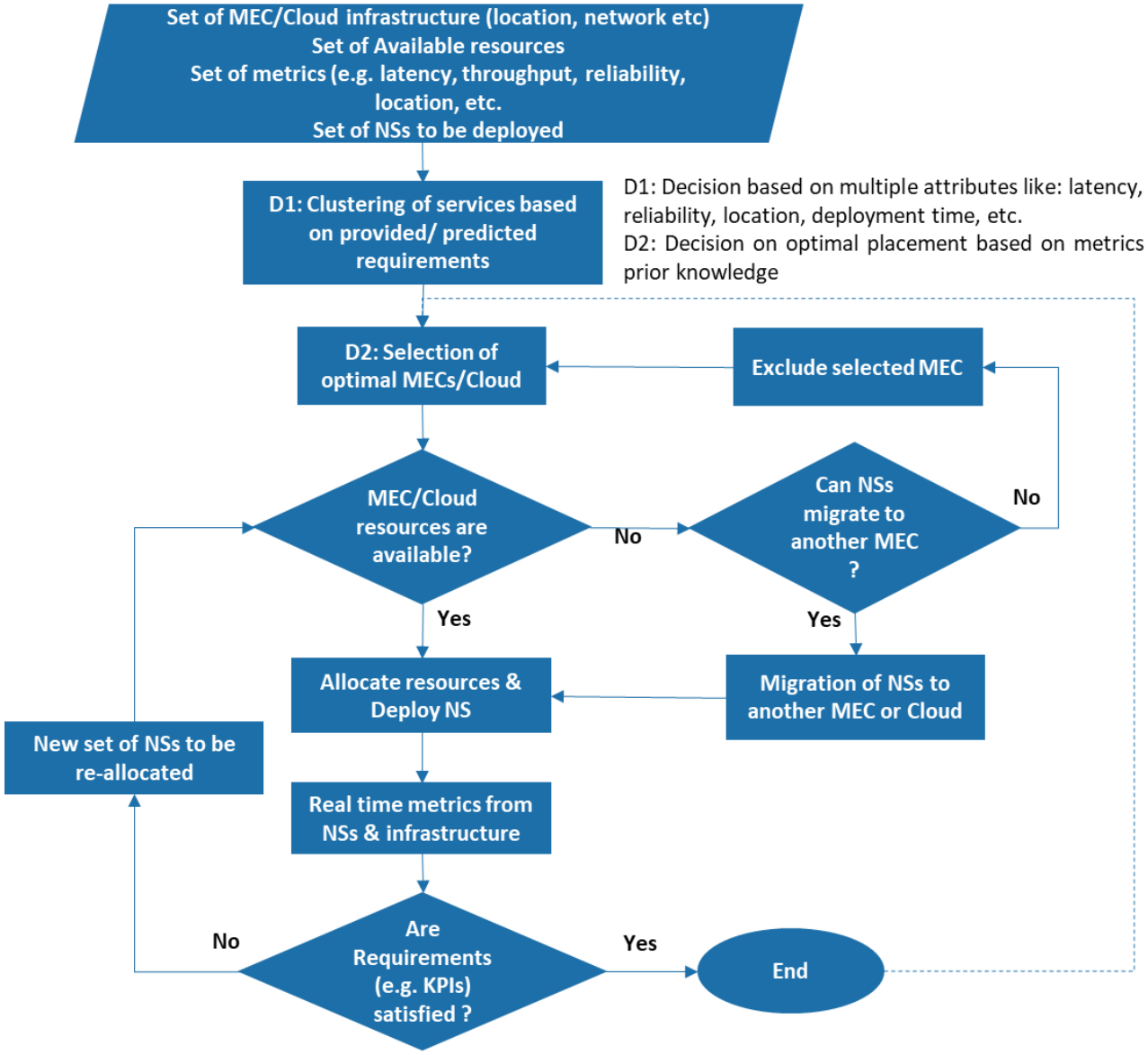


Figure 53 AI-enhanced MANO algorithm flow chart

As already discussed, OSM's embedded AI will analyse the various data and historical information to find out if changes to the deployment of VNFs need to be done. If a metric does not meet the vertical requirement for a specific VNF, OSM will make the necessary changes to fix it, either by relocation of the VNF, scaling or other appropriate actions. Some of these metrics are vertical requirements (i.e., Latency, Reliability, Availability, Throughput DL, Throughput UL, etc.), Realtime infrastructure metrics (e.g., Latency, Throughput, CPU, RAM, Disk utilization) and Application Metrics (Latency, CPU, RAM, Availability, etc.). User applications, each of which is composed of a set of Virtual Network Functions (VNFs) can be deployed at different Mobile Edge Computing (MEC) and Cloud infrastructures. In such a system, the end-users/IoT devices associated with a certain Network Service (NS) are randomly using the NS. Thus, the traffic demand of each NS is time varying according to the number of devices that needs to be serviced at each time. VNFs typically do not require constant resources but need resources based in the traffic load. Allocating a fixed number of resources often leads to either under- allocation or over-allocation of the system available resources. The machine learning (ML) algorithms (Support Vector Machines, Regression, Decision Tree, Multilayer Perceptron Neural Network) are used when appropriate by the AI-enhanced MANO component to predict the number of resources required to successfully manage a given traffic load from a VNF instance. The model's prediction can be used by the VNF placement algorithm. Integrating this algorithm into a VNF placement algorithm for joint dynamic resource allocation and placement improves the VNF placement quality while avoiding the over-allocation and under-

allocation. The contribution of AI-Enhanced MANO algorithmic approach can be summarized in the following steps:

- The learning model analyses and predicts the resource requirements of VNFs considering the traffic load at each time.
- The VNF placement algorithm taking into account the VNF resource requirements as predicted by the ML model, the infrastructure metrics/ available resources (e.g., latency, throughput, CPU, RAM, disk) and the service SLAs, minimizes the resulting end-to-end delay by deploying VNF instances close to their users.

The VNF placement algorithm is the Tabu Search which is a meta-heuristic approach that applies local search to provide solutions very close to optimal solutions. In the beginning, the Tabu Search algorithm constructs an initial solution (select the MEC/ Cloud infrastructure randomly or first sort available servers or select servers with highest capacity (CPU, RAM, disk)). Then it changes the current solution to a different one from a set of neighbour solutions and create a tabu list, which is a short-term memory of selected solutions. Finally, it updates the tabu list if a new solution led to lower cost. The algorithm is terminated after a maximum number of iterations is reached. This algorithm is a low complexity algorithm that can find an almost optimal solution for deploying VNFs in MEC / Cloud or in different distributed MECs. The cost function used by the tabu search algorithm can be formulated as:

$$cost\ function = \sum_{s_k^i \in S^i} \sum_{n \in N} delay(n) \cdot x_n^{s_k^i} + \sum_{s_k^i \in S^i} \sum_{n \in N} \frac{1}{resoures(n)} \cdot x_n^{s_k^i}$$

where N are the available MEC/ Cloud servers used to deploy the s_k^i VNFs of the network service S^i. The $x_n^{s_k^i}$ is a binary variable set to 1 if the VNFs$_k^i$ is allocated on the server $n \in N$.

Finally, to showcase our AI-enhanced MANO system, a web interface has been designed and developed as Figure 54 illustrates. This interface retrieves information from the AI-enhanced MANO component and presents in bar, pie and graph charts the different metrics from the VNFs and Infrastructure also presenting the migrations between different MECs or the optimal distribution of different VNFs in an automatic way when the AI deems necessary to do so.

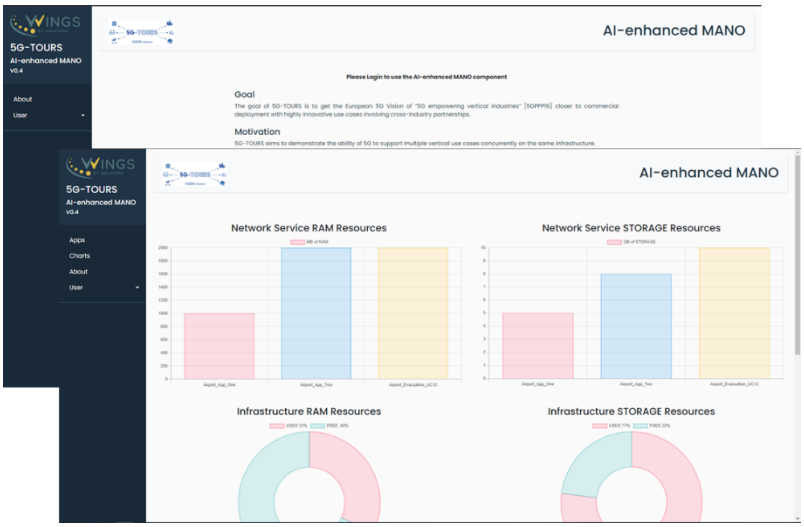


Figure 54. AI-enhanced MANO Graphical User Interface.

5.5.2 Service layer for the vertical closed-loop integration

In this section we discuss how the service layer can be used in 5G-TOURS to integrate the vertical into the operation of the network. While we are working for the integration of this functionality in the 5G-TOURS UCs, we initially developed this for the ETSI ENI PoC#9, that has been developed by many 5G-TOURS partners [54].

In this context, we focused on a test network service, composed by a Content Delivery Network VNF providing an eMBB-alike service, onboarded through a portal such as the 5G EVE one. Besides the orchestration system, the PoC is composed by several modules, as discussed next.

Network Service Monitoring

In order to take decisions such as the one envisioned by this instance of the service layer (e.g., scaling decisions when the target QoS not met) the monitoring framework is fundamental. For the purposes of our system, we utilize the CPU consumption metric as the main driver for the scaling decisions. Being the proof of concept not deployed in a production system, we emulate the system load by offering load according to well defined pattern. Specifically, we adopt the methodology provided in [55] to generate real world demands for a given service. In this case, we selected the YouTube service, given the similarities with the CDN VNFs running in the eMBB network slice used for this PoC. Hence, we created a simulated load that provides a proportional number of requests to the load recorded in the same work.

System functionality

The set of functionalities we developed for this system (we remark that we follow the ENI specifications for our implementation) are essentially two:

- **Intent translation from verticals**: this task helps the vertical service providers (a video content provider in this case) to operate the services without deep knowledge of the network internals. Besides the initial, high level, input towards the orchestration included in the templates files such as the ones included in the 5G EVE, verticals may want to be included in the loop to have a finer control level without going too much into the details of the specific implementation. As discussed also in the following, we provide a "knob" to the vertical that can be used to steer the autonomous operation in an understandable way.
- **Scaling automation:** by forecasting the future load of the network, the ENI -alike system can provide scaling suggestion to the assisted system (the orchestration one in this case). After the initial trial period, however, we detected that the timings related to the data forecasting were too fast to be coupled with the orchestration system. Thus, we designed two tiers monitoring and forecasting system that 1) performs a long term forecasting, and 2) continuously re-evaluate the decision at short term, to correct the long term one only if needed. By doing this we achieve a scalable operation of the system, by avoiding too frequent re-orchestrations of the resources, which have an inherent cost.

Network Automation Algorithm

For this implementation of the service, we employ the algorithm described in [56], which was initially discussed in D3.2 [3] along with its implication on the overall architecture. The overall structure of the machine learning algorithm is depicted in the Figure 55 below.

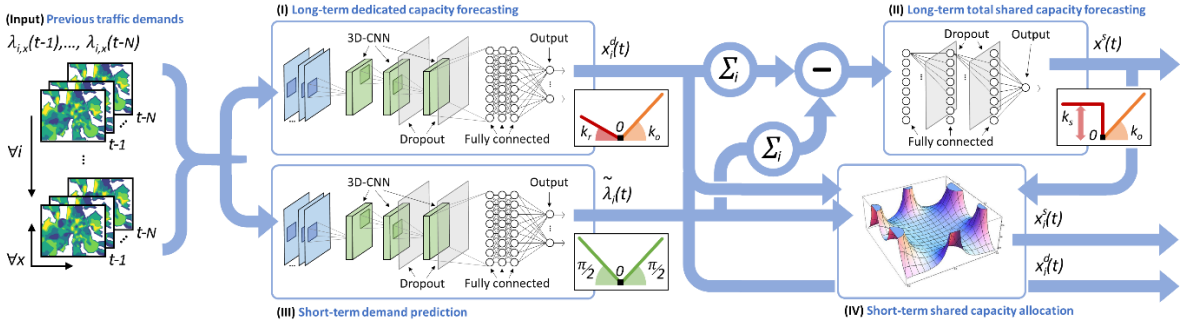


Figure 55. The architecture of the Machine Learning algorithm taken from [56].

The structure stems from the previous load of the system, gathered through the monitoring platform discussed before. Among the plethora of statistics that could be monitored with the Prometheus platform, we selected the CPU consumption, monitored at 5 seconds granularity. This is used as the input block that feeds the blocks I and III that provide the long-term and short-term capacity forecasting used afterwards. This information is merged into the blocks III and IV that perform the short-term allocation of the resources using a customized loss function, that is needed to interpret the vertical needs with respect to resource allocation. As exemplified in Figure 56 below, we employ a customized function that includes the cost incurred by the vertical in terms of resource consumption.

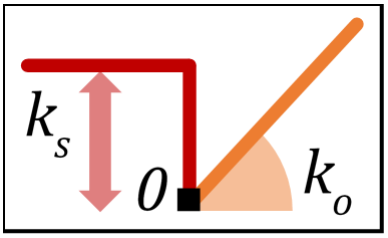


Figure 56. The cost function used by the PoC.

The cost function is modelled by two variables k_S and k_O, which embody the cost for the SLA violations of the system and the cost of overprovisioning, respectively. That is, if the prediction of the system is below the real value (the perfect prediction is marked as 0 in Figure 56), we incur into a cost equal to k_s, while positive errors means that more resources are allocated in the system, which have a proportional cost increase in the system. We summarize these two variables into one $\alpha = k_S / k_O$, that catches the relation between the "aggressiveness" of the system that can be configured by the vertical: higher values of α will push the operation of the system towards "more expensive" prediction and scaling decisions, as depicted in the next Figure 57.

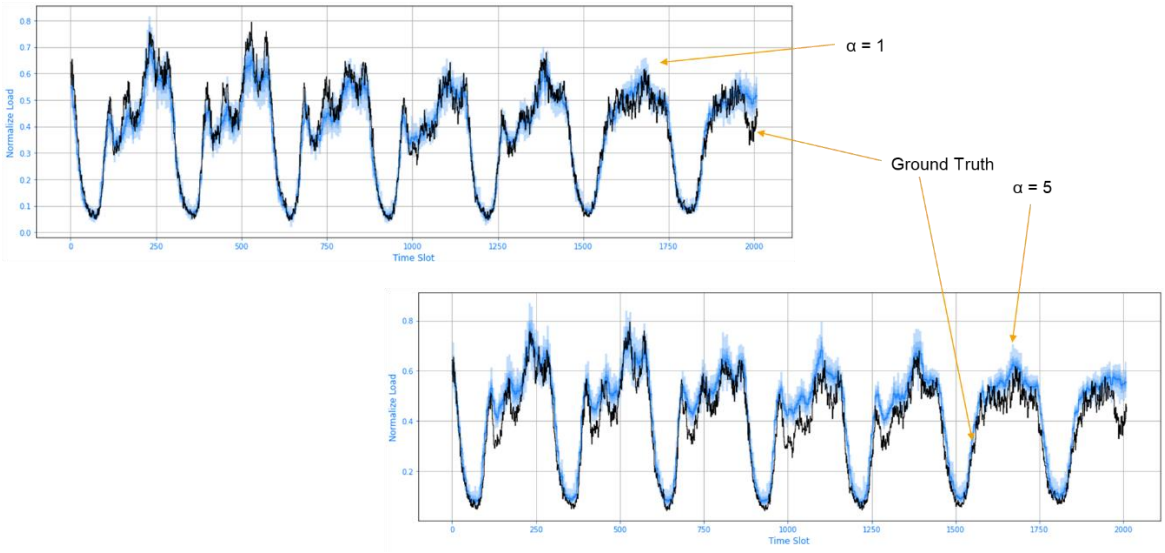


Figure 57. Load Prediction with different values of α.

From the figure, we can see that higher α values achieve a more resilient resource assignment (scaling is performed on thresholds) at the price of a higher resource provisioning.

In order to achieve a statistically significant variability for the short-term prediction reconsideration of previously taken actions we also employ a measure for uncertainty, depicted by the shades of blue in Figure 57. The overall neural network structure is available as Open Source at [57].

5.5.3 OSM AI-Agents

The general design approach of integrating the AI-Agents functionality as part of the OSM orchestration platform makes sure that the regular O&M activities associated to AI-Agents (lifecycle management, configuration,

etc.) are integrated in the orchestration platform itself, and therefore, remain within the scope of the operator in charge of operating the services that could make use of the OSM AI-Agents feature (5G EVE in our case).

However, it is also desirable to be able to delegate some of the functionality associated with AI-Agents to the verticals, namely, the access to the data for training the AI models (if necessary), the training process itself, and the deployment of the models once trained. In the end, the verticals have a precise understanding of their business domain and the metrics and data associated with it.

The design of the AI-Agents functionality on OSM in two separated blocks (the AI-Agents itself, and the AI-Models Server - see Section 5.2.1) facilitates this task, since, when the vertical requires it, the AI-Models Server can be deployed in the scope of the vertical itself, even if the AI-Agents are deployed in the VNFs composing the service (and are managed from the orchestrator); see Figure 58.

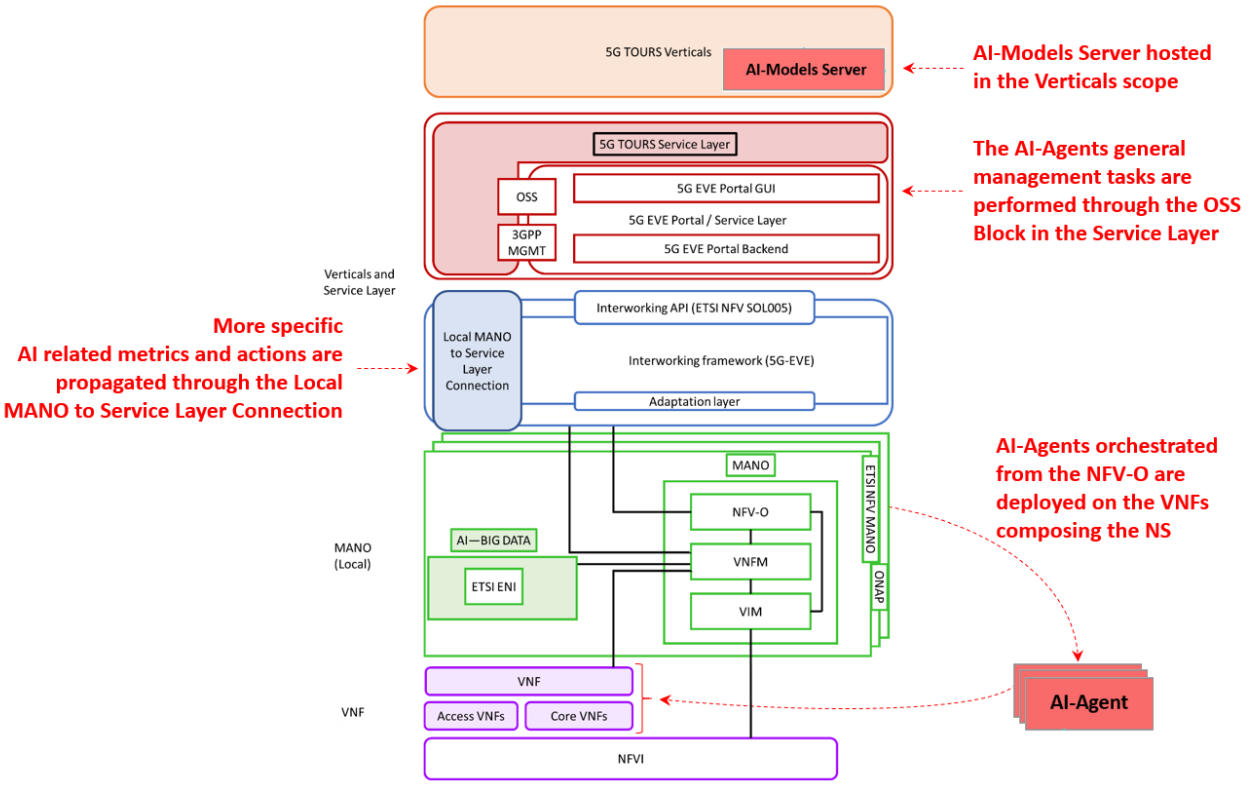


Figure 58. Deployment of the AI-Models Server in the Verticals scope.

With this approach, since AI-Agents are part of the Orchestration Platform, their basic management operations (e.g., configuration or lifecycle control) can be handled from the OSS block in the 5G-TOURS Service Layer, which is accessible through the 5G EVE Portal. However, the more specific actions related to the design and training of the artificial intelligence models are delegated to the vertical itself, which has a more precise knowledge of the metrics associated with its specific domain.

Naturally, in order to feed the service metrics to the AI-Models Server and for the AI-Agents to query the AI models deployed in the AI-Models Server, it is necessary to enable specific interfaces. Since these interfaces are an enhanced functionality specifically associated with the vertical, the access is done through the local MANO to Service Layer connection enabled in the Service Layer. Given that the design of the AI-Agents functionality for OSM is based on an agnostic approach (the vertical can freely choose which technology its model's server is based on), the specific implementation of the interface depends on the framework used by the vertical to implement the AI-Models Server. In the models developed in 5G-TOURS based on TensorFlow Serving (see section 5.2.1), a specific REST API has been used. Other interfaces could also be chosen if the vertical selects a different implementation for the AI-Models Server.

5.5.4 Multicast/broadcast functionality

The work on the service layer for multicast/broadcast support follows two axes: the development of the new network functions being specified at 3GPP (5.4.4.1), and the optimization of the transport delivery stack (5.5.4.2).

5.5.4.1 5G architecture for the multicast/broadcast service layer

The devised architecture will be compliant to 3GPP architecture as described in section 5.4. As the 3GPP architecture is not yet finalised, the architecture will certainly evolve during the 5G-TOURS project. According to 3GPP work in TS 26.802 [4], 2 new entities are defined in the 5GCore Service Layer:

- **Multicast Broadcast Service Function – User plane (MBSF-U)**: Generic packet transport functionalities available to any IP multicast enabled application such as framing, multiple flows, packet FEC (encoding) and ROHC, Multicast/broadcast delivery of input files as objects or object flows, sourcing of IP Multicast if needed, Media anchor for multicast/broadcast data traffic if needed.
- **Multicast Broadcast Service Function – Control plane (MBSF-C):** Service level functionality to support MBS and interworking with LTE MBMS, it interacts with AF and MB-SMF for MBS session operations and with MB-SMF to determine the MBS session transport parameters.

The multicast broadcast service functions provide the transport functionalities offered in 4G by the BM-SC. The inclusion of these service functions has been discussed a lot within the 3GPP SA2 working group, as only a transparent transport of IP flows was initially considered.

The 5G-TOURS service layer for multicast/broadcast service layer implementation will be based on the current virtualised BM-SC, nevertheless a continuous look at 3GPP activities is needed to verify the progress on the new interfaces specified by the 3GPP SA2 working group.

5.5.4.2 Low latency MABR delivery

A 3GPP multicast/broadcast Service is usually based on DASH/HLS. DASH and HLS are also widely used for Over The Top (OTT) services in order to offer high quality streaming of media content over the Internet, delivered from conventional HTTP web servers.

DASH and HLS works by breaking the content into a sequence of small segments containing typically few seconds of data. ABR technique suffers from a significant latency (up to 30 seconds for HLS) compared to broadcast delivery, mainly due to the need of buffering segments at the point of reception, which can jeopardize the viewing experience, especially for live event, such as sport [59].

The solution designed by the industry relies on the Common Media Application Format (CMAF) which allows to split the media segments into chunks which can be consumed before the end of the segment generation, as illustrated in Figure 59.

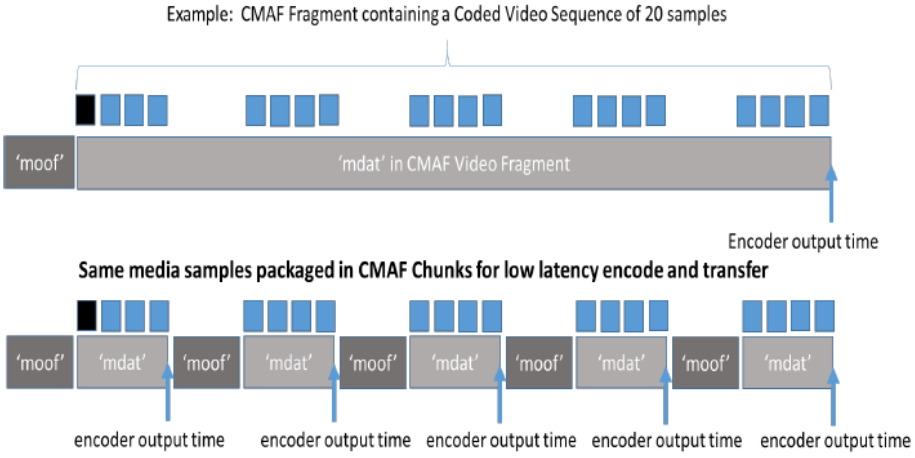


Figure 59. CMAF segments split into chunk for low latency.

Delivering CMAF chunks over multicast has several impacts on the service layer:

- The northbound API shall be able to ingest CMAF chunks while the media segment is still being generated, which can be achieved by using HTTP chunked transfer encoding (IETF).
- The multicast delivery protocol shall be adapted to transport chunks of media segment instead of media segments.
- The local cache at the point of multicast reception shall be able to serve CMAF chunks without waiting the full reception of the media segments.

Hence, the delivery stack of the current BM-SC was enhanced to deliver chunks through the FLUTE protocol [60] with the following rules illustrated in Figure 60:

- one transport Object Id (TOI) per chunk announcement by a specific File Delivery Table (FDT),
- chunks are identified with the following URI: [media segment URI]#[chunk byte_offset][&last=true].

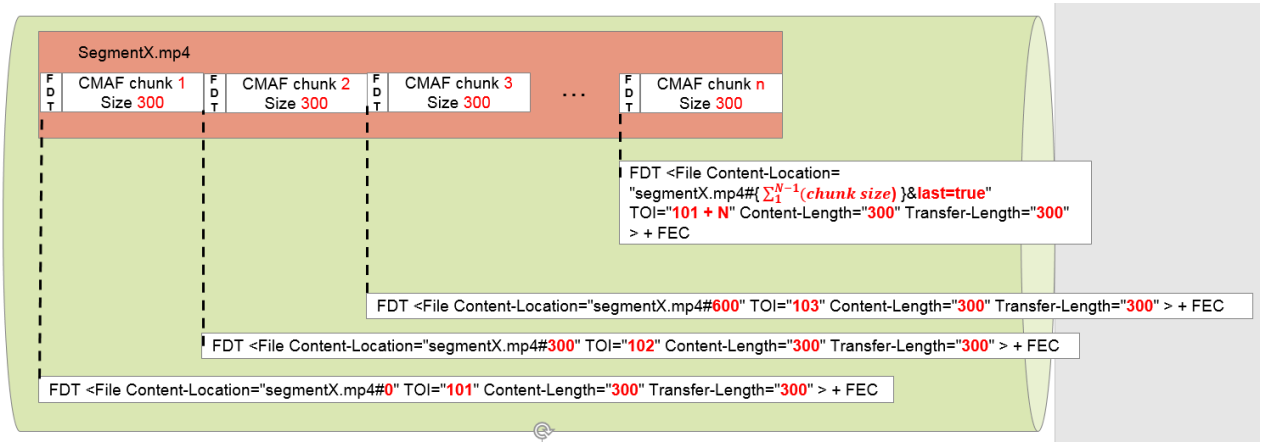


Figure 60. Delivery of CMAF chunks with the FLUTE protocol.

The support of low latency has also been specified by DVB for M-ABR DVB using this solution.

5.5.4.3 Service Layer error correction with RaptorQ

Raptor codes are the first known class of fountain codes with linear time encoding and decoding. Raptor codes perform software-implemented Forward Error Correction (FEC), thus, there is no need to integrate any additional dedicated hardware in the system to perform AL-FEC. The increment of reliability in data transmission derived from Raptor Codes determined its standardization for MBMS AL-FEC.

However, it exists a new generation of fountain codes named RaptorQ [18]. RaptorQ, developed by Qualcomm, is the most recent and improved version of Raptor codes. This code effectively reduces the redundant FEC information outperforming conventional Raptor codes and offering near-optimal properties with minimum processing overhead and excellent flexibility in packet lost recovery.

In the encoding process, RaptorQ generates the recovery symbols from the original data or source symbols. Then, an identifier is created and included in each symbol before the data transmission. A graph of the process can be seen in Figure 61.

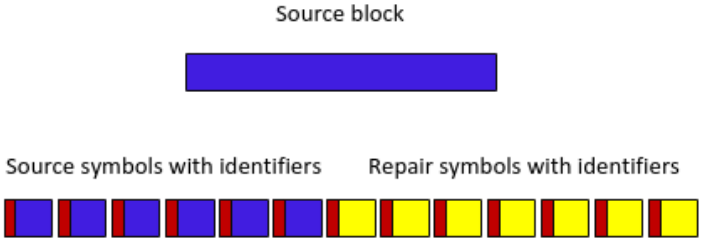


Figure 61. RaptorQ encoding process.

In reception, RaptorQ decoder only needs slightly more information than the original to recover from high Bit Error Rate and even packet loss. In practice, RaptorQ only needs 2 extra symbols comparing to the original quantity to reduce Bir Error Rate to less than one in a million. In addition, its excellent flexibility enables the same decoding performance no matter if the received symbols are source or repairing symbols.

RaptorQ will be included inside the service layer to perform the application layer forward error correction process. High quality video will be encoded with RaptorQ technology before being injected into the MB-UPF. The encoded data will be decoded inside the SDR-based UE.

Multicast sessions enhanced with AL-FEC RaptorQ technology will optimize both efficiency and reliability allowing high quality video transmission even in the most challenging environments.

6 Conclusions

This document describes the progress on the architecture and physical deployment in 5G-TOURS. The architecture has been designed to accommodate all the novel technologies developed during the project that will support and enhance the proposed use cases. Architectural 5G-TOURS design is considered as foundational basis to host and nurture these innovations. Baseline 5G-TOURS architecture, while encompassing 5G-EVE functional layers, fulfils the requirements posed by the use cases. Architectural "instantiations" are also made, describing the functional assets that are common to all use cases per trial site, and how different WP3 innovations are linked over them.

The physical implementation progress on the three trial sites - Turin, Rennes and Athens - is explained. A description of the current capabilities, including use case on-boarding and 5G EVE integration, alongside the plans for future enhancement. The different sites are now moving towards the second phase of the implementation, requiring an intermediate reality check for the network instantiations in trial sites. This includes the technology and insertion strategies of genuine 5G-TOURS components – in addition to what 5G EVE already offers – also discussed in the document. In particular, continuous check of consistency on alignment between network infrastructure and use case descriptions is done, as well as work on the architecture to onboard the core network and 5G RAN, physical and logical connectivity, etc.

The novel mechanisms brought by the project are presented. In particular, an in-depth discussion and status report is given of Enhanced MANO, AI based orchestration, broadcast/multicast and a new service interface for verticals to deploy services, manage slices and monitor network KPIs. Several implementations of 5G-TOURS technical innovations as part of the expanded infrastructure are described in detail, namely:

- Service layer to provide an interface to vertical customers: 4 different implementations,
- Enhanced MANO: 2 implementations,
- AI-based data analytics and orchestration: 2 implementations,
- Broadcast support: 2 tracks.

The next deliverable D3.4 "Final Architecture and Deployment Results" is planned at M31. It will present the achieved results in the final architecture, technological innovations and infrastructure deployment in the trial sites. In particular, it will provide the refined architectural instantiations per site from the functional point of view, forward-thinking "security-by-design" considerations, lessons learnt from technology integration into 5G EVE and advancements on research innovations per identified domain: Enhanced MANO, AI, broadcast, service layer (while possible exploitation and market impact of these technological innovations will be tackled in conjunction with WP8). D3.4 deliverable will finalise the WP3 activity report.

Acknowledgment

This project has received funding from the EU H2020 research and innovation programme under Grant Agreement No. 856950.

References

[1] L. Vignaroli, “5G-TOURS D4.2: First Touristic City use case results,” 2020. [online]. Available: http://5gtours.eu/documents/deliverables/D4.2.pdf.

[2] M.Gramaglia, “5G-TOURS D3.1, Baseline Architecture and deployment objectives,” 2019. [online]. Available: https://5gtours.eu/documents/deliverables/D3.1.df

[3] C. Barjau, "5G-TOURS D3.2, Technologies, architecture and deployment initial progress," 2020. [online]. Available: http://5gtours.eu/documents/deliverables/D3.2.pdf.

[4] 3GPP TR 23.757: “Study on architectural enhancements for 5G multicast-broadcast services," 2020. [Online].

[5] 3GPP TR 26.802: “5G Multimedia Streaming (5GMS); Multicast architecture," 2020. [Online].

[6] T. Italia, “A multi-source dataset of urban life in the city of Milan and the province of trentino dataverse.” [Online]. Available:https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EGZHFV

[7] F. de Lange, “5G-TOURS D5.2, First safe city use cases implementation results”. [online]. Available: http://5gtours.eu/documents/deliverables/D5.2.pdf

[8] E. Giannopoulou, “5G-TOURS D6.2, First mobility efficient city use cases implementation results” [online]. Available: http://5gtours.eu/documents/deliverables/D6.2.pdf

[9] GSMA Generic Network Slice Template v3.0 https://www.gsma.com/newsroom/wp-content/uploads//NG.116-v3.0.pdf

[10] Marc Mollà Roselló, "5G EVE D3.4: Second implementation of the interworking reference model", 2018. Available: https://www.5G EVE.eu/wp-content/uploads/2020/07/5geve-deliverabled3.4-final.pdf

[11] Román, Ricardo, "5G EVE D1.3: 5G EVE end to end facility reference architecture for vertical industries and core application", 2019

[12] 3GPP TS 28.533 v16.0.0, “Management and Orchestration of Networks and Network Slicing; Management and Orchestration Architecture (Re-lease 16),” Jun. 2019.

[13] 3GPP TS 23.288 v16.1.0, “Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (Release 16),” Jun. 2019.

[14] O-RAN Alliance White Paper, “O-RAN: Towards an Open and Smart RAN,” Oct. 2018.

[15] D. Bega, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, " α -OMC: Cost-Aware Deep Learning for Mobile Network Resource Orchestration," IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Paris, France, 2019, pp. 423-428, doi: 10.1109/INFCOMW.2019.8845178.

[16] D. Bega, M. Gramaglia, R. Perez, M. Fiore, A. Banchs and X. Costa-Perez, "AI-Based Autonomous Control, Management, and Orchestration in 5G: From Standards to Algorithms," in IEEE Network, vol. 34, no. 6, pp. 14-20, November/December 2020, doi: 10.1109/MNET.001.2000047.

[17] D. Gomez-Barquero, J. J. Gimenez, R. Beutler, "3GPP Enhancements for Television Services: LTE-based 5G Terrestrial Broadcast ", Jan. 2020

[18] IETF, “RaptorQ Forward Error Correction Scheme for Object Delivery” 2011. [Online]

[19] Emmanuel Cordonnier, “5G-TOURS D5.1: 5G-enabled solutions for safe cities”, 2019. [online]. Available: http://5gtours.eu/documents/deliverables/D5.1.pdf

[20] Internal report IR5.1, 5G-TOURS_IR5.1 Internal report_v1.0.pdf

[21] Wireless Edge Factory, https://b-com.com/fr/bcom-wireless-edge-factory

[22] Non-Standalone, https://www.gearbest.com/blog/how-to/nsa-or-sa-who-is-the-real-5g-network-mode-for-smartphones-and-whats-the-difference-8585

[23] Flexible Netlab platform, https://b-com.com/fr/bcom-flexible-netlab

[24] OpenStack, https://www.openstack.org/

[25] Amarisoft Callbox, https://www.amarisoft.com/products/test-measurements/amari-lte-callbox/

[26] OpenVSwitch, https://www.openvswitch.org/

[27] OpenDaylight, https://www.opendaylight.org/

[28] KVM Hypervisor, https://www.linux-kvm.org/page/Main_Page

[29] Huawei 5G Pro Customer Premise Equipment (CPE), https://consumer.huawei.com/en/routers/5g-cpe-pro/

[30] Models for vertical descriptor adaptation, 5G European, 5G EVE: 5G EVE - D4.3 Models for vertical descriptor adaptation (5G EVE.eu)
[31] Vivid E95 Cardiac Ultrasound, https://www.gehealthcare.com/courses/vivid-e95-cardiac-ultrasound
[32] First version of the experimental portal and service handbook, 5G EVE: 5G EVE - D4.2 (5G EVE.eu)
[33] KVM hypervisor, https://www.linux-kvm.org/page/Main_Page
[34] Vicens F., Subhankar P., "AI Agents for OSM - Feature description," 11 06 2020. [Online]. Available: https://osm.etsi.org/gerrit/#/c/osm/Features/+/9063/1/AI+Agents+for+OSM.md. [Accessed 08 03 2021].
[35] Lavado G., "OSM Service Assurance," 2019. [Online]. Available: https://osm-download.etsi.org/ftp/osm-6.0-six/8th-hackfest/presentations/8th%20OSM%20Hackfest%20-%20Session%208%20-%20OSM%20Service%20Assurance.pptx.pdf. [Accessed 08 03 2021].
[36] https://osm.etsi.org
[37] https://osm.etsi.org/wikipub/index.php/OSM_PoC_11_Deployment_of_AI-Agents_in_OSM
[38] Olston C. et al., "TensorFlow-Serving: Flexible, High-Performance ML Serving", 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. [Online]. Available: http://learningsys.org/nips17/assets/papers/paper_1.pdf
[39] https://www.acumos.org/
[40] https://flask.palletsprojects.com/en/1.1.x/
[41] https://www.tensorflow.org/
[42] https://keras.io/
[43] https://colab.research.google.com/
[44] https://helm.sh/docs/topics/charts/
[45] 3GPP TS 28.533 v16.0.0, "Management and Orchestration of Networks and Network Slicing; Management and Orchestration Architecture (Re-lease 16)," Jun. 2019.
[46] 3GPP TS 23.288 v16.1.0, "Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (Release 16)," Jun. 2019.
[47] O-RAN Alliance White Paper, "O-RAN: Towards an Open and Smart RAN," Oct. 2018.
[48] D. Bega, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, " α -OMC: Cost-Aware Deep Learning for Mobile Network Resource Orchestration," IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Paris, France, 2019, pp. 423-428, doi: 10.1109/INFCOMW.2019.8845178.
[49] D. Bega, M. Gramaglia, R. Perez, M. Fiore, A. Banchs and X. Costa-Perez, "AI-Based Autonomous Control, Manage-ment, and Orchestration in 5G: From Standards to Algorithms," in IEEE Network, vol. 34, no. 6, pp. 14-20, Novem-ber/December 2020, doi: 10.1109/MNET.001.2000047.
[50] M. Gramaglia, "5G-TOURS D3.1: Baseline architecture and deployment objectives" 2019. [online]. Available: http://5gtours.eu/documents/deliverables/D3.1.pdf
[51] GSA, "LTE to 5G: March 2021 – Global update", March 2021. [Online]
[52] GSA, "5G Devices ", March 2021. [Online]
[53] 5G EVE, "Collaboration with vertical use case projects – 5G EVE presentation at 5G PPP Technology Board," June 2020. [Online].
[54] Poc 09: Autonomous Network Slice Management for 5G Vertical Services - wiki (etsi.org)
[55] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs and Z. Smoreda, "Identifying Common Periodicities in Mobile Service Demands with Spectral Analysis," 2020 Mediterranean Communication and Computer Networking Conference (Med-ComNet), Arona, Italy, 2020, pp. 1-8, doi: 10.1109/MedComNet49392.2020.9191477.
[56] Ref to D. Bega, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, "AZTEC: Anticipatory Capacity Allocation for Zero-Touch Network Slicing," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, Toronto, ON, Canada, 2020, pp. 794-803, doi: 10.1109/INFOCOM41043.2020.9155299
[57] https://github.com/wnluc3m/AZTEC
[58] Tuan Tran, Ed. "5G-Xcast Deliverable D4.1: Mobile Core Network", 2018.
[59] DASH-IF/DVB Report on Low-LatencyLive Service with DASH. Available : https://dashif.org/docs/Report%20on%20Low%20Latency%20DASH.pdf
[60] RFC 6726 "FLUTE - File Delivery over Unidirectional Transport" Available: https://tools.ietf.org/html/rfc6726

[61] Y. Wang et al., "Network Management and Orchestration Using Artificial Intelligence: Overview of ETSI ENI," in IEEE Communications Standards Magazine, vol. 2, no. 4, pp. 58-65, December 2018, doi: 10.1109/MCOMSTD.2018.1800033.
[62] https://www.amarisoft.com/technology/enodeb/
[63] 3GPP TS 33.501, "Security architecture and procedures for 5G System", available at http://www.3gpp.org/ftp//Specs/archive/33_series/33.501/33501-g30.zip.
[64] Standard - ETSI ISG MEC and 3GPP specifications - Harmonizing standards for edge computing.
[65] "Advanced Research Report: Multi-Access Edge Computing." Del'Orro, 2020.
[66] P. Donegan, "Security Requirements for Deploying MEC at Scale," Heavy Read., 2017.
[67] M. Liyanage, I. Ahmad, A. B. Abro, A. Gurtov, and M. Ylianttila, A Comprehensive Guide to 5G Security, Com édition. Hoboken, NJ: Wiley-Blackwell, 2018.
[68] INSPIRE-5Gplus Consortium "D2.1: 5G Security: Current Status and Future Trends," p. 101, 2019.
[69] Peter Schneider and Josef Urban, "Security in 5G Networks," Breitbandversorgung Dtschl., Mar. 2020.
[70] J. Hodges, "5G Security Strategy Considerations," Heavy Read., 2019.
[71] Standard - ETSI GS MEC 003 - Framework and Reference Architecture.
[72] P. Ranaweera, A. D. Jurcut, and M. Liyanage, "Survey on Multi-Access Edge Computing Security and Privacy," IEEE Commun. Surv. Tutor., pp. 1–1, 2021, doi: 10.1109/COMST.2021.3062546.
[73] Q.-V. Pham et al., "A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art," ArXiv190608452 Cs Math, Jan. 2020, Accessed: Dec. 09, 2020. [Online]. Available: http://arxiv.org/abs/1906.08452.
[74] R. Tourani, A. Bos, S. Misra, and F. Esposito, "Towards security-as-a-service in multi-access edge," in Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, Arlington Virginia, Nov. 2019, pp. 358–363, doi: 10.1145/3318216.3363335.
[75] N. Neshenko, E. Bou-Harb, J. Crichigno, G. Kaddoum, and N. Ghani, "Demystifying IoT Security: An Exhaustive Survey on IoT Vulnerabilities and a First Empirical Look on Internet-Scale IoT Exploitations," IEEE Commun. Surv. Tutor., vol. 21, no. 3, pp. 2702–2733, 2019, doi: 10.1109/COMST.2019.2910750.
[76] I. Stellios, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, "A Survey of IoT-Enabled Cyberattacks: Assessing Attack Paths to Critical Infrastructures and Services," IEEE Commun. Surv. Tutor., vol. 20, no. 4, pp. 3453–3495, 2018, doi: 10.1109/COMST.2018.2855563.
[77] Guardtime company, "Enabling Multi Party Trust in the Era of 5G and Multi-Access Edge Computing." 2020.
[78] R. Borgaonkar, L. Hirschi, S. Park, and A. Shaik, "New Privacy Threat on 3G, 4G, and Upcoming 5G AKA Protocols," Proc. Priv. Enhancing Technol., vol. 2019, no. 3, pp. 108–127, Jul. 2019, doi: 10.2478/popets-2019-0039.
[79] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data Security and Privacy-Preserving in Edge Computing Paradigm: Survey and Open Issues," IEEE Access, vol. 6, pp. 18209–18237, 2018, doi: 10.1109/ACCESS.2018.2820162.
[80] A. Masood, D. S. Lakew, and S. Cho, "Security and Privacy Challenges in Connected Vehicular Cloud Computing," IEEE Commun. Surv. Tutor., vol. 22, no. 4, pp. 2725–2764, 2020, doi: 10.1109/COMST.2020.3012961.
[81] G. Choudhary, J. Kim, and V. Sharma, "Security of 5G-Mobile Backhaul Networks: A Survey," J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl., vol. 9, no. 4, pp. 41–70, Dec. 2018, doi: 10.22667/JOWUA.2018.12.31.041.