A secure and reusable Artificial Intelligence platform for Edge computing in beyond 5G Networks

D2.1 Use cases, requirements, and preliminary system architecture

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 10101592

D2.1 Use cases, requirements, and preliminary system architecture	
WP	WP2 – Use cases, requirements analysis, and system design
Responsible partner	Conservatoire national des arts et métiers (CNAM)
Version	1.0
Editor	Stefano Secci (CNAM)
Authors	Antonino Albanese (ITL), Salah Bin Ruba (CNAM), Flávio Brito (EAB), Miguel Catalan-Cid (I2CAT), George Chatzinkonstantis (8BELLS), Estefania Coronado Calero (I2CAT), Cristina Costa (FBK), Jérôme Francois (INRIA), Leonardo Goratti (SPI), Wei Jiang (DFKI), George Lentaris (ICCS), Neiva Linder (EAB), Marco Marchetti (CRF), Giampiero Mastinu (POLIMI), Brendan McAuliffe (SRS), Daniele Munaretto (ATH), Daniel Reti (DFKI), Roberto Riggio (RISE), Gonzalo Sanchez (AERO), Stefano Secci (CNAM), Nimish Sorathiya (DFKI), Sotirios Xydis (ICCS), Bruno Lepri (FBK)
Reviewers	Françoise Sailhan (CNAM) Javier Melian Hernandez (ATOS) Jovanka Adzic (TIM) Nicola Di Pietro (ATH) Roberto Riggio (RISE)
Deliverable Type	R
Dissemination Level	PU
Due date of delivery	30/06/2021
Submission date	24/06/2021

Version History				
Version	**Date**	**Authors**	**Partners**	**Description**
0.1	16/04/2021	Antonino Albanese Neiva Linder Stefano Secci	CNAM, EAB, ITL	First internal draft
0.2	10/05/2021	Antonino Albanese Salah Bin Ruba Estefania Coronado Calero Cristina Costa Leonardo Goratti George Lentaris Neiva Linder Marco Marchetti Giampiero Mastinu Daniele Munaretto Daniel Reti Roberto Riggio Gonzalo Sanchez Stefano Secci Sotirios Xydis Wei Jiang	ITL CNAM I2CAT FBK SPI ICCS EAB CRF POLIMI ATH DFKI RISE AERO CNAM ICCS DFKI	Preliminary contributions to all sections added by the authors
0.3	12/05/2021	Cristina Costa Marco Marchetti Brendan McAuliffe Daniele Munaretto Roberto Riggio Stefano Secci	FBK CRF SRS ATH RISE CNAM	Refined draft
0.4	26/05/2021	Miguel Catalan-Cid Estefania Coronado Calero George Chatzikonstantis Cristina Costa Leonardo Goratti Marco Marchetti Brendan McAuliffe Daniel Reti Roberto Riggio Miguel Rosa Gonzalo Sanchez	I2CAT I2CAT 8BELLS FBK SPI CRF SRS DFKI RISE AERO AERO	Addition of sections 7, 8.4, conclusions, and general improvement of the text. Complete version for internal review

		Stefano Secci Nimish Soratiya	CNAM DFKI	
0.5	03/06/2021	Antonino Albanese Jérôme Francois Leonardo Goratti Neiva Linder Marco Marchetti Daniel Reti Miguel Rosa Stefano Secci	ITL INRIA SPI EAB CRF DFKI AERO CNAM	Update of use-case KPI description and addition of content to section 8. Restructuring of Section 8.1. Complete version for external review (by partners not included in WP2)
0.6	15/06/2021	Jovanka Adzic Nicola Di Pietro Marco Marchetti Javier Melian Hernandez Daniel Reti Françoise Sailhan Stefano Secci	TIM ATH CRF ATOS DFKI CNAM CNAM	Reviewed version with enhancements. Complete version for review by the Technical Manager
0.7	18/06/2021	Antonino Albanese Leonardo Goratti Bruno Lepri Marco Marchetti Roberto Riggio Stefano Secci	ITL SPI FBK CRF RISE CNAM	Inclusion of preliminary AIF requirements. Complete version for final editor review
1.0	24/06/2021	Salah Bin Ruba Stefano Secci	CNAM	Final Version

Table of Contents

List of Tables

List of Figures

Glossary	
5GAA	5G Automotive Association
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AIF	Artificial Intelligence Function
AMF	Access and Mobility management Function
APN	Access Point Name
AR	Augmented Reality
BVLOS	Beyond Visual Line of Sight
C-V2X	Cellular Vehicular communication
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPU	Central Processing Unit
CU	Centralized Unit
DE	Deliverable Editor
DL	Downlink
DNN	Data Network Name
DSP	Digital Signal Processing
DU	Distributed Unit
EDA	Electronic Design Automation
FaaS	Function as a Service

FL	Federated Learning
FPGA	Field-Programmable Gate Array
FPV	First Person View
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GPU	Graphic Processing Unit
HIL	Hardware In the Loop
HITL	Human-in-the-loop
ICT	Information and Communication Technology
IFE	In-Flight Entertainment
IIoT	Industrial Internet of Things
IoT	Internet of Things
IoU	Intersection over Union
I-UPF	Intermediate User Plane Function
KPI	Key Performance Indicator
LUT	Look Up Table
mAP	Mean Average Precision
MEC	Multi-access Edge Computing
ML	Machine Learning
mMTC	massive Machine-Type Communication
MNO	Mobile Network Operator

MQTT	Message Queuing Telemetry Transport
MTC	Machine Type Communications
NFV	Network Function Virtualization
NFVI	Network Function Virtualization Infrastructure
NFVO	Network Function Virtualization Orchestrator
NSSAI	Network Slice Selection Assistance Information
OWL	Web Ontology Language
PCIe	Peripheral Component Interconnect express
PDCP	Packet Data Convergence Protocol
QoS	Quality of Service
RAN	Radio Access Network
REST	REpresentational State Transfer
RIC	RAN Intelligent Controller
ROS	Robot Operating System
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSU	Road Side Units
RTL	Return To Launch
RU	Radio Unit
S-NSSAI	Single – Network Slice Selection Assistance Information
SBA	Service-Based Architecture
SDN	Software-Defined Networking
SMF	Session Management Function

TRL	Technology Readiness Level
UC	Use Case
UL	Uplink
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-reliable Low Latency Communications
VR	Virtual Reality
V2I	Vehicle to Infrastructure
V2N	Vehicle to Network
V2V	Vehicle to Vehicle
4G, 5G, 6G	Fourth, Fifth, Sixth Generation of cellular networks

Executive Summary

This deliverable details the AI@EDGE use cases and preliminary functional and system requirements. Use-cases are described in terms of context, involved stakeholders, technical framework, requirements, and key performance indicators. Preliminary plans for the experimentation platform are also provided. Then the requirements are summarized and linked to justify the preliminary functional and system specifications and future research directions. This deliverable also draws a preliminary functional view on the 5G AI@EDGE system architecture that will be further developed in D2.2, D3.1 and D4.1.

1 Introduction

This deliverable describes the use cases and scenarios addressed by AI@EDGE and is related to the activity of Work Package 2, Task 2.1.

The deliverable documents the activity performed to characterize and specify the project use cases and respective overall system-level and functional requirements, which will be considered as part of WP3, WP4, and WP5 for the development and overall assessment of the AI@EDGE network and service automation platform and of the AI@EDGE connect-compute fabric. The approach used to collaboratively work on the description and requirements of the project relied on the writing of synthetic use case templates in the beginning of WP2, aiming at harmonizing the analysis methodology and the presentation structure of the four use cases.

Use case leaders were responsible for gathering all the information and filling in the different fields of the template as, for example, the related actors and scenario, the description and objectives, the technical requirements and Key Performance Indicators (KPIs), as well as the features of interest coming from the adoption of AI@EDGE's technical solutions, requested to fulfill the use case needs.

In the characterization of the use cases, the analysis considers the various types of users to be served by the AI@EDGE platform and the types of services and applications that are to be supported by it. This includes, for example, smart-mobility services in UC1, security of industrial systems in UC2, ground surveillance services in UC3 and airplane communication services in UC4.

The deliverable also caters the architectural and system requirements for AI@EDGE; in particular, the type of users, services, and applications defined by the use cases, in order to derive application-driven network requirements. Furthermore, system-level and functional requirements are defined, considering the potentials of the serverless paradigm combined with AI-enabled network operations.

Task 2.1 effort is meant to contribute to the definition of a set of architectural recommendations for the 5G-PPP program with particular focus on the realization of sustainable beyond-5G and 6G architectures capable of supporting the innovative services and use cases envisioned by AI@EDGE.

The document is organized as follows. Section 2 below gives a general introduction to the AI@EDGE use cases. Sections 3 to 6 detail each use case, UC1, UC2, UC3, UC4, respectively. Section 7 summarizes the use case requirements. Section 8 draws preliminary system-level and functional specifications of the AI@EDGE architecture. Section 9 concludes the deliverable and draws the next steps.

2 Use cases general description

The project addresses the following use-cases:

UC1: Virtual Validation of Vehicle Cooperative Perception. Vehicles exchange in real-time their trajectories and use artificial intelligence models to understand the surrounding environment and predict possible dangers.

UC2: Secure and Resilient Orchestration of Large Industrial IoT Networks. Smart factory communication and computing infrastructures, involving a large and heterogeneous set of industrial actuators, sensors, specialized application servers and network fabric, are designed to be secure and resilient against faults, attacks, bugs and load variations.

UC3: Edge AI Assisted Monitoring of Linear Infrastructures in Beyond Visual Line of Sight Operations. Monitoring drones exchange data with ground computing facilities to detect anomalies, by using 3D environment reconstruction and data fusion to guide drone mobility and operations along large distances.

UC4: Smart Content and Data Curation for In-Flight Entertainment Services. High definition multimedia content is offered to passengers by dynamically computing the content of interest and aggregating 3GPP and non-3GPP network technologies to reach high throughput and reliability.

Table 1 summarizes which of the AI@EDGE technological enablers are exploited by each use case.

Table 1 Technological enablers exploited by each use case

Technological Enabler	**UC1**	**UC2**	**UC3**	**UC4**
Distributed and decentralized serverless connect-compute platform	**Y**	**Y**	**Y**	**Y**
AI-enabled application provisioning	**Y**	**Y**	**Y**	**Y**
Cross-layer, multi-connectivity radio access				**Y**
Hardware accelerated serverless platform for AI/ML	**Y**	**Y**	**Y**	**Y**
Network and service automation platform	**Y**	**Y**	**Y**	
Secure, reusable, & resilient machine learning for multi-stakeholder environments		**Y**		

Starting from the use case framework and requirements described in the next sections, we summarize the key technical requirements in Section 7 and elaborate how the adopted technological enablers will allow us to meet these requirements.

The AI@EDGE technologies are evaluated, specified and developed in the framework of WP2, WP3 and WP4, and experimentally assessed within WP5. Section 8 describes the project preliminary view on these technologies. These technologies will also be designed in relation with relevant 5G-PPP working groups and the 3GPP standardization roadmaps toward their inclusion in future releases. Moreover, these technologies will be positioned as well in perspective of forthcoming 6G architectures' development, as detailed in Section 8.

For each use case, a specific validation methodology is defined to reflect close-to-real scenarios, with conditions similar to production. Moreover, for each use case, the minimum set of expectations from the

5G and beyond-5G cellular technologies, in terms of KPIs, are also provided, and will be further elaborated during WP5 activities. The Technology Readiness Level (TRL) expected for the AI@EDGE use cases and related platforms goes up to 6, i.e., "technology demonstrated in a relevant environment" according to the TRL classification provided by the Annex G of the H2020 Work Programme.

The main expected use case outcomes of the AI@EDGE project, along with their TRL levels, are summarized in Table 2.

Table 2 Expected TRL elevation for AI@EDGE use cases

AI@EDGE use case platforms	**Expected TRL Final (Initial)**
UC1: Platform for virtual validation of vehicle cooperative perception	TRL 6 (4)
UC2: Platform for secure and resilient orchestration of large (I)IoT networks	TRL 5 (4)
UC3: Platform for edge AI assisted drones in beyond-visual-light-of-sight operations	TRL 6 (4)
UC4: Platform of smart content & data curation for in-flight entertainment services	TRL 5 (3)

The detailed description of each use case in the following four sections follows the same structure to ease the comparison in terms of requirements and functionalities among use cases.

3 Use case 1: Virtual validation of vehicle cooperative perception

The vehicle cooperative perception use case (UC1) is based on a reference setting where several vehicles exchange data related to their trajectories. Data is gathered at the network edge and is used to build a view of the surrounding environment that will be used by Artificial Intelligence Functions, named AIFs in the project, to predict potential collisions and dangers. The end-to-end system development to demonstrate this use case is challenging, given the complexity and costs in stake. For these reasons, UC1 adopts an emulation environment able to scale with such a complexity and to perform exhaustive and reproducible tests.

AI@EDGE provides a set of technologies in its connect-compute fabric that relies on 5G and AI, which allow to make the road safe and the vehicular traffic fluid. Particularly challenging is the roundabout situation, where fluidity and safety are of paramount importance, and UC1 will focus therefore on this challenging situation. Safer and more fluid (thus, less pollutant) traffic are goals stated by the European Commission to be reached by 2030.

3.1 Reference scenario

Today, the validation of vehicles' cooperative perception is a challenge because it deals with numerous vehicles that have to: detect in real-time the surrounding traffic scenario; exchange their sensed data; and share their intended manoeuvres with other vehicles. Large tests are needed even to address one single traffic scenario. In particular, within roundabouts, the problem of cooperative perception and exchange of data on the intended path of the vehicles is crucial for allowing safe and fluid traffic.

Cooperative perception tests become even more complex when dealing with mixed real and virtual traffic scenarios. To overcome the problem of simulating human behaviour by means of a mathematical model, we plan to interconnect a dynamic driving emulator operated by a real human driver with a traffic simulator to design, implement, and test the digital twinning of a mix of real and emulated vehicles. In such a context, a major factor is the commonly called Human In The Loop (HITL) factor; it can be exploited to check the behaviour of a human driver among highly automated and connected vehicles. Actually, the information on the intended vehicle path at a roundabout shall be given to the driver. This requires a cognitive workload that has to be experimentally assessed. Additionally, the HITL factor can be exploited to mimic a car with a perfect automated vehicle.

The technical challenge of UC1 is to create the network data exchange required to build a cooperative perception between emulated vehicles and a virtual human-driven vehicle. The main advantages of adopting the cooperative perception simulation-emulation environment are:

- Realistic and insightful simulation of corner cases and scenarios with a high number of vehicles that are difficult and expensive to execute and reproduce with real vehicles on the road.
- Emulation of network behaviors and configurations (handover, network congestion, latency issues, etc.) that cannot be perfectly reproduced and controlled in an out-of-lab environment using a 5G network testbed.
- Emulation of mobile channel impairments (multipath, Doppler frequency, fading, shadowing, etc.) that cannot be controlled in a real environment.

The AI-based digital twinning process will leverage the features of the AI@EDGE connect-compute platform. In particular, the network and service automation features will allow the digital twinning to cope

with mutating radio network environments, dynamically learning changing network states and taking reconfiguration action. In terms of 5G stack, the AI@EDGE platform 5G core will be interfaced with a 5G network emulator to allow testing a broader range of scenarios and network configurations, while measuring KPIs. Besides supporting the performance evaluation of vehicular applications depending on the cooperative perception, the AI@EDGE platform will also contribute to demonstrate the advantages and the applicability to the use case of the deployment of virtualized core network components at the network edge. This kind of testbed will be innovative also because it will showcase the possibility of integrating two different simulators of connected vehicles and "smart" driving.

3.1.1 Actors, roles, stakeholders

The main actors and stakeholders involved in the use case scenario are the:

- Driver.
- Automobile Manufacturer.
- Telecommunication Operator.
- Testing Facility.
- Network Equipment Vendor.
- Municipality/Road Operator.
- AI models/applications developer/provider.

The use case Partner roles are:

- ATH: will provide the 5G core network functions at the edge, collocated with the AI-based traffic controller, to enable low latency communications.
- CRF (Stellantis): will participate as Automotive Manufacturer developing the Testing Facility for Automotive Telematic Boxes, which are vehicle on-board embedded units to connect Vehicles to the network.
- POLIMI: will host the testing facility for driving emulation and simulation.
- FBK: will define, develop and test AI (machine learning-based) models for the Cooperative Perception application.

3.1.2 Use case context

The Cooperative Perception system can be useful in many vehicular scenarios such as Lane Change, Intersection Movement Assist, and Roundabouts intersections. In this use case, we concentrate on a Roundabout scenario, depicted in Figure 1. Typically, a roundabout smooths traffic only when this does not exceed a certain number of vehicles, and depends on incoming and outcoming traffic in the surrounding area. A distributed intelligent management of the local traffic can manage the roundabout traffic suggesting how to approach the roundabout to drivers or sending information to self-driving vehicles. To provide this kind of relevant information it is also necessary to create a **digital twin of the roundabout and of the surrounding** collecting real time information from vehicles or from telematic units on the road, as for instance Multi-access Edge Computing (MEC), Road Side Units (RSUs) or Virtual RSU run at the MEC facility, Fog computing nodes, cameras, as depicted in Figure 1.

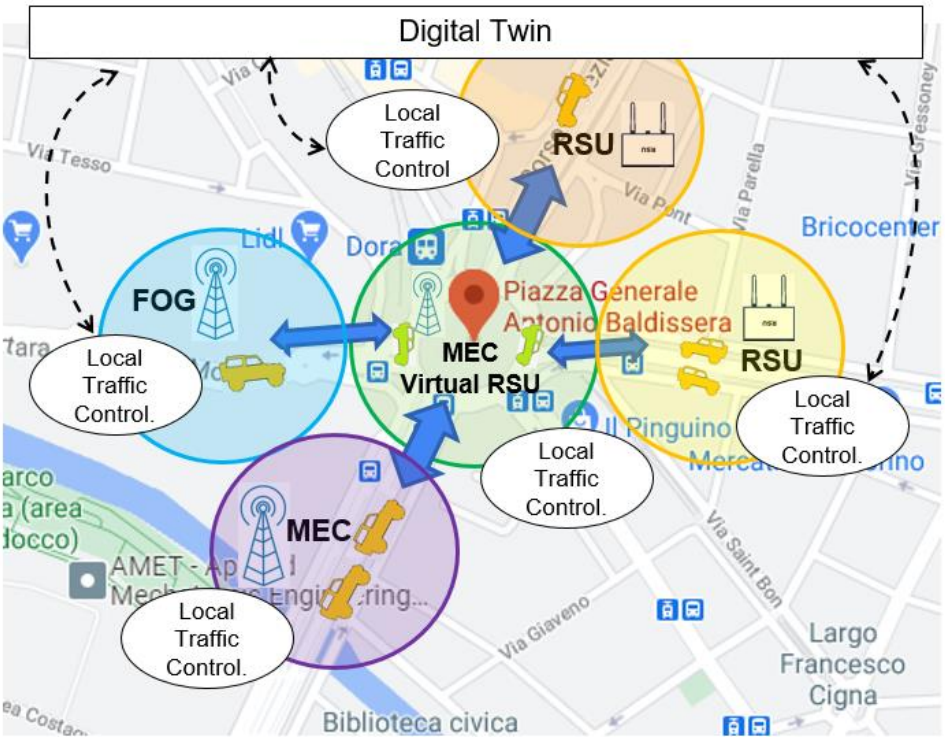


Figure 1 Use case 1 roundabout scenario

3.2 Main objectives

The main objective of the use case is to make use of the AI@EDGE fabric to create a geographically distributed Virtual Validation testbed and support the cooperative manoeuvres between vehicles; in details:

- Create a geographically distributed Virtual Validation system connected to a 5G network, using the 3GPP 5G system Uu and PC5 interfaces to validate the use case.
  In particular, the PC5 is the short-range network interface of the cellular vehicular communication (C-V2X) subsystem, which operates on the 5.9 GHz band; it is defined by 3GPP Release 14 and is essential for implementing self-driving vehicles and the connected car, and in Releases 15 and 16, [16], 3GPP continued its standardization for 5G systems.
- Support cooperative manoeuvres between vehicles:
  - Collecting and sharing information from/to vehicles.
  - Building a mixed approach between centralization and decentralization where vehicles will learn, using multi-agent reinforcement approaches, cooperative approaches and a centralized AI-based digital twinning system will aggregate and elaborate the information collected from Vehicles and Road Side Units.
- AI Traffic Control, using Digital Twin, will send information messages to the vehicles to solve inefficiencies in learned cooperative policies.

WP5 activities will allow linking to these objectives a dedicated scientific and technical roadmap to reach them.

3.3 Use case architecture

The use case architecture, depicted in Figure 2, is made of the following main parts:

- Simulation Environments
  - The roundabout simulation environment.
  - The Outcoming-Incoming surrounding-traffic simulation environment.
- The AI Traffic Controller.
- The Digital Twin.
- The 5G Network Infrastructure.
- The telematic Box (On board unit that provides connectivity to the vehicle).

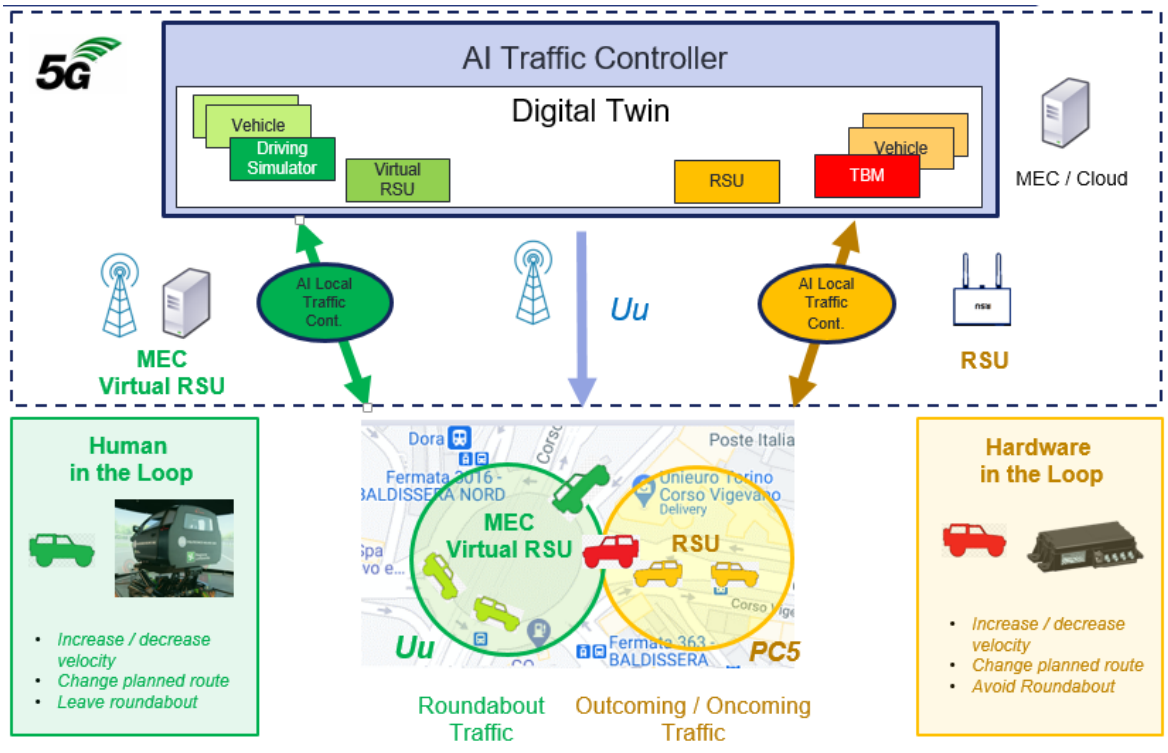


Figure 2 Use case 1 functional architecture

Simulation Environments

The Use Case will have 2 simulation environments, one for the roundabout and the other one for the Surrounding, the "reason why" two simulation systems are deployed is for possible future business scenario requirements such as sharing testing/simulating facilities.

Roundabout simulation environment

The main roundabout traffic emulation/simulation system, with its driving simulator is a HITL (Human-in-the-loop) that is a human based driving based system; its photo is in Figure 3.

Studying a roundabout scenario with a driving simulator is a challenging task. Driving simulators are professional devices that can be extensively used by professional drivers only. Everyday driving's simple tasks are not easy to reproduce within driving simulators if untrained drivers are involved.

We are planning to perform extensive tests with the POLIMI driving simulator with:

- the extensive use of the driving simulator with a panel of non-professional drivers to evaluate the AI@EDGE connect-compute fabric impact on achievable application KPIs.
- the use of the driving simulator with a professional driver mimicking a highly automated vehicle, assessing the AI@EDGE connect-compute fabric impact on KPIs again.

The tests, as mentioned above will be made in a challenging environment where a traffic generator will be used together with the actual driving simulator.

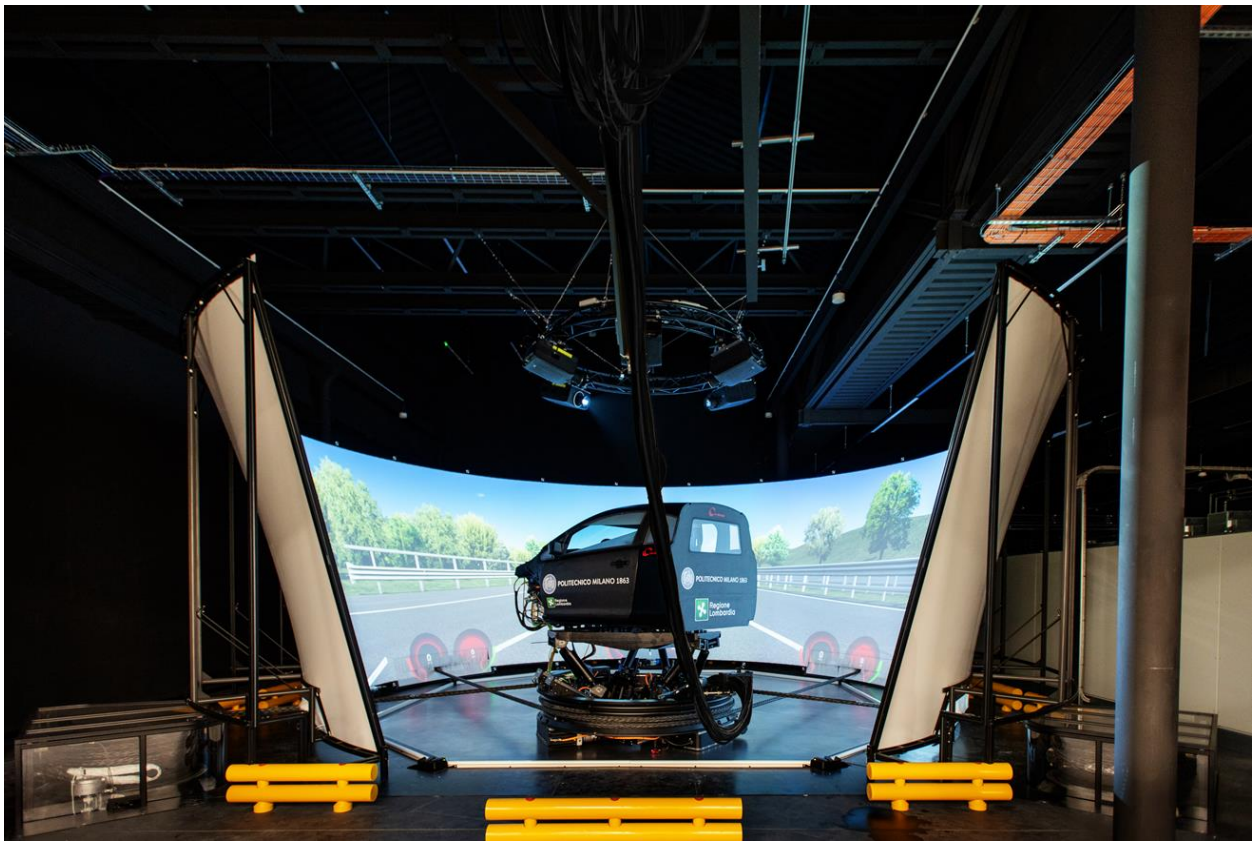


Figure 3 Use case 1 experimental emulation facility at POLIMI

Surrounding-traffic simulation environment

The main surrounding-traffic simulation system - with its Vehicle to Vehicle (V2V), Vehicle to Infrastructure (V2I) and Vehicle to Network (V2N) - is a HIL (Hardware-in-the-Loop) that is an embeddedsystems driving based system not Human controlled. Two types of equipment will be used to create a 5G network access environment for V2V/V2I and V2N communications.

The simulation of V2V/V2I communications is based on the 3GPPP C-V2X stack using the PC5 radio channel as follows:

- The user configures the traffic scenario using the simulation environment editor, assigning parameters like vehicle velocity and path.
- Once defined, the traffic scenario is deployed into the radio environment that supports the scenario execution.
- The scenario is at last executed, and the communication layer simulates data transmission over PC5 interfaces and forwards received information to the Telematic Box part, enabling the verification of transmitter and receiver communications.
- The scenario and the Telematic Box are synchronized and the GNSS (Global navigation satellite system) signal generator is used to provide the positioning to the Telematic Box.

The simulation of V2N communications is based on the Uu interface usage as follows:

- The network emulator is connected with the Telematic Box using the 5G RAN.
- The emulator offers an IP connection to the internet so the Telematic Box can reach MEC or Cloud services.

Digital Twin

The digital twin is a virtual representation of the roundabout and its surroundings and it is created collecting information from various sources; its main information layers are:

- Static data: road and intersection information obtained from static digital maps.
- Semi-static data: road signs, landmarks.
- Semi-dynamic data: information for temporary changes, such as weather, traffic jams.
- Dynamic data: dynamic rapidly changing information, such as vehicle information (GPS position, speed, heading, etc.).

AI Traffic Controller

The traffic control system is AI Function (**UC1 Traffic Controller AIF**) that relies on the digital twin of the city traffic environment. The AI traffic controller is based on the digital twin information and on data learned, it recommends how to approach or how to leave the roundabout sending alerts to drivers and to self-driving vehicles. Depending on the traffic conditions and if the vehicle is inside or outside the roundabout the AI application identifies the situation and can send alerts like: Increase/decrease velocity, Change planned route, Leave/Avoid roundabout, etc.

A number of vehicle mobility and trajectory features will need to be estimated in real-time by the AIF, such as for instance ingress and egress vehicle speeds and angles at round-abouts. We will first focus on ingress speed estimations. Given the real-time constraints, we target to adopt a reinforcement-learning approach and to compare it to a baseline reactive decision taken using real-time measurements. At the current stage of the research in this area, we target a comparison in terms of ability to avoid accidents. The estimation precision will be numerically studied and evaluated in Task 2.3 and WP5 activities. At the time being, we guessed that an acceptable accuracy target could be 99% precision in accident avoidance. Likely, this would be a lower bound. Further activity in Task 2.3 and WP5 will allow us to refine the AIF performance requirements also based on simulation and emulation data.

5G Network Infrastructure

The mobile core network component designed as a MEC-based solution will be collocated with the AI Traffic Controller function. The Digital Twin will benefit from this private network solution "at the edge" as it allows the vehicular traffic generated locally by the V2X devices to be kept local and be processed in real-time for a quick control loop decision.

From the AI@EDGE Platform support and integration point of view, there are three main technical enablers:

- Distributed and decentralized serverless connect-compute platform.
- AI-enabled application provisioning.
- Network and service automation platform.

In details, the required components and functionalities are:

- AIF Availability: methods to ensure very high availability of the Traffic Controller AIF.
- Data-driven service-lifecycle management approaches for AIFs.
- Resource allocation and deployment of distributed AIFs.
- Softwarized VNFs for the mobile core network.

The detailed list of VNFs/core components and their features will be further agreed and specified, during the activities of WP5.

Telematic Box

The 5G Telematic Box On board unit that provides connectivity to the vehicle:

- Uu and PC5 radio interfaces
- GPS
- MQTT/AMQP and C-V2X Client

3.4 Requirements

We describe the use case requirements distinguishing among requirements in terms of necessary technology, security and privacy features and KPIs.

3.4.1 Technical requirements

Concerning required functional technical components, given the emulation and simulation framework, the UC1 architecture requires the following specific elements:

- Near MEC/Edge server.
- Far Edge with capable GPU for learning and inference of AI applications.
- Radio network capability for 5G and PC5 interfacing.
- Computing system able to monitor psychological human signals.

For the latter, the involvement of humans is performed according to the most strict ethical rules. The recorded signals and the subjective answers to proper questionnaires/forms will be made within an anonymous framework. Researchers will never be able to identify the subject associated with the recorded signals. The certification of the ethics procedure is issued by the Quality Office of Politecnico di Milano.

3.4.2 Security and privacy requirements

The main security and privacy UC1 requirement is privacy preserving in the exchange of messages, because sensitive and personal information may be exchanged. Another important aspect regarding privacy is the data storage and how the Digital Twin persistence is managed.

Moreover, the robustness of the Traffic Controller AIF against attacks such as radio jamming has also to be ensured and assessed.

During its development and testing, the use case shall rely on anonymised data sets coming from data collected on public datasets. Such data will be used to train the AI Traffic Control system to identify traffic situations. Given the consideration above it is not expected to have any critical challenge regarding privacy of data.

3.4.3 Key Performance Indicators (KPIs)

There are a lot of reference use cases and standards related to "traffic scenarios". 5GAA (5G Automotive Association) defines in [34] some use cases and related KPI. The 5GAA use cases more similar to vehicle cooperative perception UC1 are:

- Cross-Traffic Left-Turn Assist.
- Intersection Movement Assist.
- Traffic Jam Warning.
- Real-Time Situational Awareness & High Definition Map.
- Cooperative Lane Change (CLC) of Automated Vehicles: Lane Change Warning

The CLC and Traffic Jam Warning scenarios appear particularly relevant for UC1 in terms of vehicle dynamics and movement, and of traffic management. Aligning to these two scenarios' KPIs in [34] and this commonality, we can draw the following KPIs for UC1:

- *Latency KPIs*:
  - 160 ms for vehicle dynamics and movement (V2V communication).
  - 2000 ms, from Jam detection to received Traffic Alerts messages to the driver (V2N communication).
- *Vehicle density KPI*: 12000 vehicle/km^2 as expected number of vehicles per a given area.
- *Positioning KPI*: 1.5 m in order to deal with vehicle dynamics and movement.

Further KPI analysis will be done in T2.3 and WP5 to possibly refine and precise the boundary conditions to apply the KPIs.

3.5 Use case testbed

The UC1 testbed is being built as a geographically distributed testbed consisting of 2 sites connected with a 5G Network and its distributed computational resources (see Figure 4):

- Turin (Italy) Site:
  - Emulation of an On board Vehicle setup with Telematic Box (using Uu and PC5 interfaces), CAN Bus, Head unit with a HMI (Human Machine Interface).
  - PC5 channel emulator.

- V2X simulation environment (Vehicles/Road Side Units).
- 5G RAN emulation.

- Milan (Italy) Site:
  - Driving Simulator.
  - Traffic simulation environment.
  - Sensors for detecting psychological workload of the driver and his/her acceptance of the driving experience (namely, measuring the skin potential resistance at the hands, heart rate variability, if needed, eye tracking).
- 5G network:
  - 5G Core Network as a MEC-based network solution to allow local traffic break-out.
  - MEC and Cloud Server.

AI Traffic Controllers and the Digital Twin are deployed on the 5G network and managed through the AI@EDGE Platform.

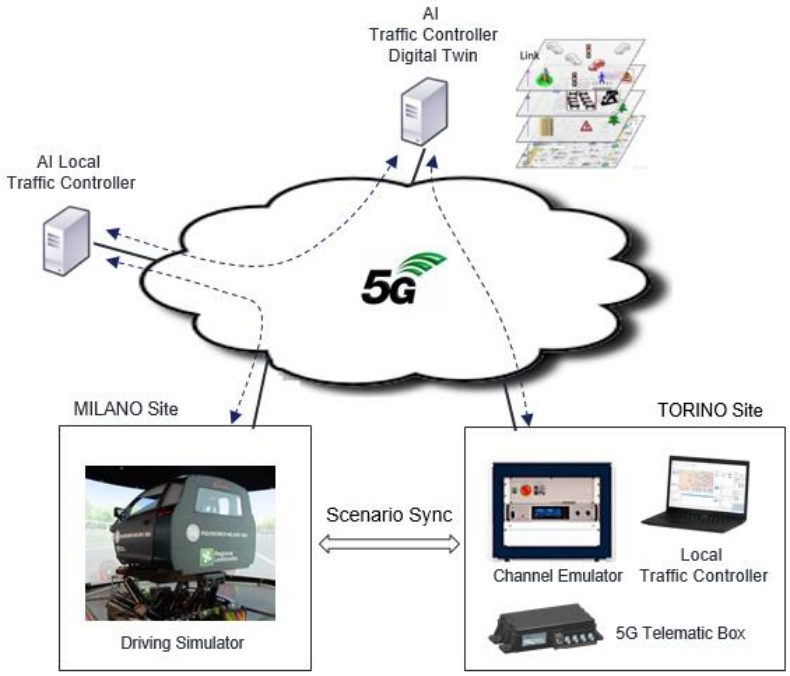


Figure 4 Envisioned use case 1 testbed environment

3.5.1 Platforms features requirements

To ensure the use case KPI requirements are met, the reliability of the platform and the communication latency have to be controllable and adjustable by the network automation loop.
Other important platform requirements are related to the orchestration functionalities to provide the deployment at the edge of the network of the core components and to deploy services ensuring the AIFs Life Cycle Management. As a consequence of these requirements the hardware hosting, the core network, and the MEC have to be dimensioned based on the use case requirements (e.g., support of HW acceleration).

3.5.2 Hardware equipment

The hardware equipment required for the UC1 testbed is:

- 5G network fabric for V2X communications.
- Driving Simulator.
- MEC server and Cloud Server, e.g. COTS HW server (Intel x86).
- Telematic Box.

3.5.3 Application and software components

The required key software components are:

- Traffic Simulator.
- Softwarized 5G CN, deployed at the edge.
- Traffic Controller AIF running at the MEC facility co-located with the 5G CN.

3.5.4 Preliminary testbed deployment and access time-line

As of the current activities on UC1 preparation, we envision the following main steps in UC1 experimental activities:

- At first the testbed will be composed of a traffic generator and the driving simulator of POLIMI in static form (i.e. not moving). Preliminary milestone: M21; final M30.
- After expected integration issues of the co-simulation of vehicle motion with the traffic generator are solved, the POLIMI driving simulator will be used at its full capability. At this step, fault and disturbances will be artificially injected to check the performance of the drivers running into a roundabout. Period: from M21 onward.
- Professional drivers will be involved in the assessment when the first group of test drivers will have completed their experimental activities. Period: from M21 onward.
- The first Turin test site will be ready with a local traffic simulator integrated with the radio access emulator to create a setup with a first 5G telematic box workbench with Uu and Pc5 interfaces. Preliminary milestone: M21; final M30.

The precise timeline will be elaborated within WP5.

3.6 Evaluation criteria

From the Cooperative Perception use case point of view the main evaluation criteria is the reduction of traffic jams in the roundabout avoiding collisions and hazards.

From the Virtual Validation point of view the main evaluation criteria is the number of non-human driven vehicles managed and the vehicle density supported in a distributed validation environment maintaining low latencies. In the distributed validation system it is crucial that a driver is able to drive in situations comparable to real situations on the road.

3.7 Risks and potential issues

Risks and issues are related mainly to technical integrations and human factors. The set of identified risks and related mitigation actions are given in Table 3.

Table 3 Use case 1 identified risks and envisaged mitigation actions

Use case 1 identified risks	**Mitigation Actions**
From the simulation point of view, one issue is related to the "Real time" scenarios integration between the V2X simulation environment and the driving simulator.	To guarantee a real scenario to the driver it is necessary to align the running scenarios between the two sites (Turin and Milan); the vehicles simulated between the two test sites must therefore move in the scenario with acceptable latencies for a simulation with human presence.
For the Network emulation it is important to consider eventual issues in integrating the emulated RAN with the core components (especially under the performance constraints imposed by the targeted KPIs), also possible difficulties in integrating applications in the MEC server with the 3GPP standard-based 5G network components it is to be considered.	We plan to minimize these risks by selecting for the use case an emulated RAN solution whose interfaces are compliant with 3GPP standards: Athonet's core network, indeed, has full compatibility with such standards and its integration with 3GPP-compatible RAN components has already been successfully carried out in many other occasions. Concerning the compatibility of the core network components deployed at the edge and the applications running on the MEC server, the risk of inefficient or ineffective integration will be mitigated by deploying the core components and the applications over separate servers. The interaction between these two servers will be limited to an IP-based data packet exchange governed by the user plane function.
From the human factors point of view, expected issues are related to possible difficulties in using the driving simulator with normal drivers but countermeasures are being developed. Unexpected interaction of the professional driver with external information on the intentions of other cars is also possible.	Issues related with normal drivers using the driver simulator are currently under study at POLIMI by a multidisciplinary team including psychologists. We expect to receive sound information enabling us to solve or minimize such issues. In any case, the capability of adapting to a driving simulator is very personal, so we intend to recruit more people than necessary, ensuring replacements for those who show negative responses to the simulator. Professional drivers can be instructed on how to react to unexpected situations. They also can be trained to interact with the developed algorithms for traffic information.
Another important threat to a smooth UC1 experimentation with the AI@EDGE connect-compute fabric is related to the availability of data for training the ML model for the Traffic Controller AIF.	Use datasets with similar scenarios from other European projects (ex: L3Pilot) or from open data.

4 Use case 2: Secure and resilient orchestration of large (I)IoT networks

In this section we present the second use case context and the evaluation approach.

4.1 Reference scenario

Smart and connected factories have the goal of integrating devices so that industry production processes can respond to evolving factory floor conditions more rapidly and intelligently than in standard factories. As depicted in Figure 5, a smart interconnected factory has a greater number of attack surfaces and is thus more vulnerable to cyber threats. In use cases where near-instantaneous data transmission is needed, latency is a concern. To overcome the latency challenge, edge computing is a promising direction in the redesign of industrial factory ICT environments. At the same time, novel edge and IoT devices introduce a new attack surface, making the Industrial Internet of Things (IIoT) network environment vulnerable. This special attention should be dedicated when evaluating novel connect-compute fabrics powered by 5G technologies. Use case 2 focuses on 5G-enabled smart factory environments that use 5G massive-machine-type-communication (mMTC) slices to link IIoT and MEC facilities. AI/ML capabilities in the network and the production modules are used for detecting anomalies including security threats, and hence supporting mitigation and reorchestration of the connect-comput fabric.

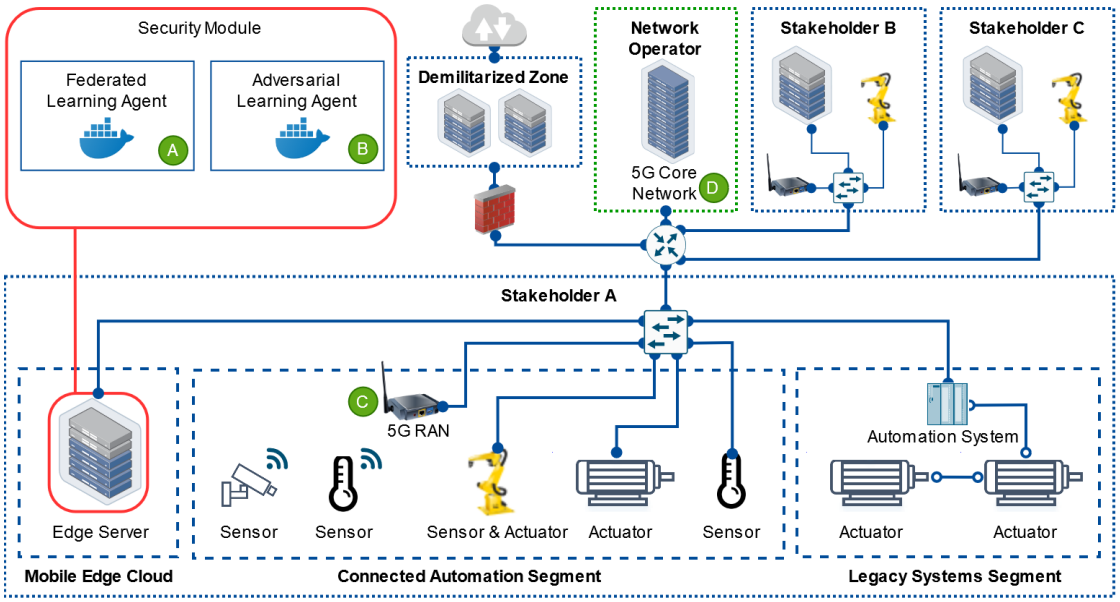


Figure 5 Use case 2 reference scenario

4.1.1 Actors, roles, stakeholders

When an industrial automation system suffers from attacks, different stakeholders require a precise diagnosis of the failures that led to the malfunctioning. A hypothetical network attack would involve 5G and local area network operators, an operator of the target industrial system, malicious actors within the premises and remote, as well as the stakeholders of the attacked system such as customers or vendors.

In such an environment, the security operations should encompass a number of distributed components, able to monitor, learn, and detect anomalies, in order to apply appropriate countermeasures. In the framework of AI@EDGE, we assume the security module of the edge application (server) owner is connected to the network operator's 5G core network, trained against anomalies as well, and operated on the near edge. Scalable distributed learning can be done through federated learning, thanks to its capacity to guarantee confidentiality and scalability; in federated learning, models trained on data from different stakeholders are merged to form a distributed federated learning model instrumental for accurate and comprehensive anomaly detection.

The main actors and stakeholders involved in the use case scenario are the:

- Operators of the target industrial systems.
- Malicious Actor/Attacker.
- Edge cloud / application server operator.
- Network equipment vendor.
- Municipality operator.
- AI models/applications developer/provider.

Under this general context, the use case partner roles are as follows:

- ATH will provide network functionalities such as an edge component for traffic breakout and will give support for the serverless architecture.
- CNAM will contribute to the definition of anomaly detection algorithms, their design using federated learning and smart network interface cards (P4-NetFPGA Smart-NICs [23]), and testbed integration.
- DFKI is coordinating the use case, providing the network testbed and IIoT components, and helping with the data collection, federated learning, anomaly detection, and testbed experimentation activities.
- INRIA will contribute to the federated learning, anomaly detection, and data collection with a particular expertise on the creation of attack scenarios.

Additional partners could join the UC2 efforts bringing their particular expertise where needed.

4.1.2 Use case context

As shown in Figure 5, the IIoT-powered production plant utilizes 5G to connect various industrial automation systems. A 5G campus network can be shared among distinct smart factory operators as stakeholders, with possibly an edge cloud in each factory hall, as well as a 5G RAN for sensors and actuator interconnection to the computing fabric. Pervasive monitoring systems are in place to allow data collection useful for distributed anomaly detection at edge nodes using a federated learning approach. In the mobile edge cloud, an AI security function (**UC2 Security AIF**) is scaled and orchestrated so as to be able to support federated learning and anomaly detection operations; it will, in particular, encompass an adversarial learning agent to protect the edge cloud from adversarial machine learning on the machine learning logic (e.g., data-poisoning [19], model poisoning [20], and free riding [21] attacks). While ensuring a level of privacy adopting a distributed federated learning approach, the security AIF will run an anomaly detection against a heterogeneous set of processing and network data.

The edge cloud servers are further connected to the main cloud platform. The network automation framework will conduct workload management on a unified connect-compute fabric. Flexible, intelligent, and secure service management solutions will be developed with focus on designing a multi-tier and multi-stakeholder federated AI@EDGE infrastructure.

4.2 Main objectives

The smart factory scenario recognizes the presence of various segments, each of which is possibly managed by a distinct stakeholder and is responsible for controlling an industrial automation system while maintaining confidentiality. On top of the AI@EDGE architecture, the UC2 aim is manyfold:

- To develop, deploy, and test AI/ML frameworks for secure orchestration of large-scale IIoT applications. We consider IIoT and production scenarios to showcase the solutions developed in this use case. In these scenarios, the network automation framework will conduct workload management on a unified connect-compute fabric.
- To create intelligent industrial service automation solutions, with a focus on smart building AI-enabled secure and reusable multi-tier infrastructures able to cope with a large set of technically and strategically distinct stakeholders.
- To conceive how to use distributed network security AIFs for intrusion detection at the device-level and 5G-component levels with the goal to detect attacks (e.g., zero-day attacks, federated learning attacks) and production-threatening anomalies.
- To design data-driven control loop components to function in conjunction with security monitoring and mitigation operations, resulting in intelligent, safe, and reliable network services that serve advanced 5G applications.

WP5 activities will allow linking to these objectives a dedicated scientific and technical roadmap to reach them.

4.3 Use case architecture

In this use case, the AI@EDGE connect-compute fabric components will be integrated into a smart factory demonstration room in the facilities of DFKI Kaiserslautern (Germany). Targeting the previously described scenario, at least three mobile edge servers will be set up in the showroom to run the federated learning AIF solution, able to collect and analyze data from the actuators, sensors, servers and 5G system components toward anomaly detection. Simultaneously, the adversarial learning agent will protect the edge cloud from feeding adversarial input data into the machine learning applications.

Each edge server represents a separate party to simulate the multi-stakeholder environment, and all share a private 5G campus network. The showroom has a 5G network on the N-78 Band (3.7-3.8 GHz) with OpenRAN from the AirSpan vendor [35] and a Core Network from DRUID [36]. Edge servers and other network nodes will be equipped with P4-NetFPGA smart-NICs able to compute features at line-rate. In addition, traffic offload solutions among edge computing nodes will be integrated to improve endpoints connect-compute experience, namely in terms of latency.

As an IIoT device running on the infrastructure, a cloud controlled automated guided vehicle (AGV) is planned to be integrated as well, in order to create concurrent sensing and actuation related communications, and to create spatial anomalies as well.

4.4 Requirements

In order to meet UC2 milestones, certain requirements must be met. The system needs to be able to handle the complexities of real-life problems. As for UC1, we describe the requirements in terms of technical components, security and privacy features and KPIs.

4.4.1 Technical requirements

One of the most difficult problems that service and storage providers face in the cloud-to-thing spectrum era is effectively managing diverse and heterogeneous computing environments, such as the IIoT smart factory one.

In this technology space, we expect to leverage on the serverless capabilities of the AI@EDGE fabric to scale with changing computing demands; to help the serverless resource management, the current state of the art based on Kubernetes/Docker can be the capable container-based orchestration for this technology, by offering the easy and stateless deployment required by a serverless architecture.

Moreover, the high computing load deriving from the continuous machine learning load, of the security AIF to constantly monitoring the running fabric state, hardware acceleration solutions are expected to be integrated in the use case environment; not only NetFPGA systems for computing traffic features at line-rate, but also possibly GPU and standard FPGA systems to offload some the learning computational effort, by means of parallelization of atomic operations.

Finally, to communicate with the edge server at the ground level, radio network capacity for the 5G network will be the essential requirement.

4.4.2 Security and privacy requirements

In this use case, the 5G network is shared between the various stakeholders.

Adopting a federated learning approach in data processing allows sharing trained models while preserving the data confidentiality among different stakeholders, for network operators to industrial actuators and sensors belonging to different entities. The UC2 environment is indeed particularly challenging due to its multi-stakeholder heterogeneous environment where many security and privacy factors surface, such as responsibility ambiguity/data ownership, bylaw conflict/location of legal disputes, shared environment, different objectives for trust, loss of governance/ loss of control, service provider lock-in, visibility, trans-border data flow, information transfer to third party [25].

4.4.3 Key Performance Indicators (KPIs)

In order to measure the success of the technology applied in this use case, the following KPIs have been identified:

- *Known-attack detection*: we target an attack detection accuracy of at least 97% against known attacks. This number could be revised depending on the complexity of the attacks. WP5 will elaborate on the possible attacks and the mitigation plan and expected mitigation deadline for each attack.
- *Zero-day detection*: we target an attack detection accuracy of at least 97% against unknown (zero-day) attacks. We will build a set of zero-day attacks, such that the attack detection framework is not trained taking them in consideration, starting from those identified in [22]. A known attack should be detected and blocked within a few minutes, which is a common acceptable lag to cover both slow denial of service attacks and command and control communications often happening before zero-day attacks.
- *False Alarm Rate*: to keep the risk of alarm fatigue low, it is not only vital to detect as many anomalies as possible, but also to keep incorrectly detected anomalies as low as possible to prevent alarm fatigue. Following the state of the art, the false alarm rate should be below 0.1% [26].
- *Adversarial federated learning attack detection*: lower than the federated learning epoch duration, that will be determined during Task 3.3 and WP5 activities.

Task 2.3 and WP5 will further elaborate on (i) specific smart industry detection targets taking into account reference sensing and actuation systems, and (ii) the mitigation plan and targets for known and zero-day attacks, possibly classifying possible mitigation plans as a function of the attack scope when this can be determined. The result of these forthcoming activities may lead to a refinement of these KPIs.

4.4.4 Use case testbed

In order to test the anomaly detection performance, a set of attacks such as man-in-the-middle, denial of service, data and model poisoning attacks, will be injected to test if the anomaly is detected, therefore also including adversarial agent attacks against the federated learning logic. The infrastructure on top of which the experiments will be conducted will be composed of:

- 5G Core: all components are connected over the core. On the core network the traffic will be monitored, and additional data will also be analyzed for anomaly detection, including CPU, RAM, and storage metrics from both physical and containerized servers.
- Edge platform: composed of edge physical and containerized servers, equipped with P4-NetFPGA Smart-NICS. Traffic offloading modules will also be used to reduce forwarding latency. On these edge nodes, network traffic will be measured in real time using Smart-NICs as well as physical and containerized servers as for the 5G Core functions.
- Attacker Simulation: to evaluate the use case and showcase its functionality, several attacks on need to be implemented. This includes network attacks as well as attacks on the AI using adversarial ML methods. Traffic generator nodes may be integrated to the testbed for this purpose.
- Hardware acceleration: besides NetFPGA for real-time traffic monitoring, GPU and standard FPGA cards may be included to the testbed in order to accelerate the learning effort hence to decrease the learning time while ensuring high accuracy.

4.4.5 Platforms features requirements

The platform should support federated learning, anomaly detection, network automation, data collection and protection from adversarial machine learning. Also, hardware acceleration will be used.

4.4.6 Hardware equipment

For federated learning applications, multiple Nvidia Jetson AGX are used as mini PCs with GPUs. On the operating side, devices operated over a 5G network like AGVs (Automated Guided Vehicles) are placed. The AGV is available at DFKI with a ROS (Robot Operating System) interface. A cloud control for the AGV will be developed by DFKI. All devices are connected over the 5G network infrastructure in the DFKI show room. ATH will provide the necessary traffic offload components for MEC local breakout operations, and CNAM will provide P4-NetFPGA Smart-NICs for traffic features computation.

4.4.7 Application and software components

The use case will leverage various software components and applications at different stages and for different purposes.

Namely, for the automated guided vehicle ROS instructions will be used for communication. For traffic offloading and Smart-NIC operations, the usage of an SDN controller for P4 [32] data structure collection used by the security AIF will be explored, in particular using P4 Runtime South-Bound-Interface. Additional software tools for traffic injection and capture, such as dpdk-packet generator, tcpdump, TCP-replay [33] are softwares expected to be used for the experimentations.

Moreover, federated learning open libraries, such as scikit learn [44], TensorFlow [45] or Pytorch [46] will be evaluated for the integration of the federated learning logic; this will be further used by neural networks for anomaly detection, namely using long-short-term-memory (LSTM) deep neural networks.

4.4.8 Preliminary testbed deployment and access time-line

The whole use case can be broken down following the different workload components. As mentioned, the AGVs and 5G network will be the heart of this project, both are accessible at DFKI. The first step should be the setup of the 5G network and edge cloud to provide the environment for the IIoT. The computing and network equipment is currently being assembled. The following step would be the collection and pipelining of the data from the UC2 testbed elements and the set-up of the anomaly detection framework. For this a suitable set of features and data types has to be identified. Proper training data is the most important aspect of an AI model, these steps with data pre-processing would represent a time-consuming step. In parallel a federated learning infrastructure will be set up through multiple edge servers in the DFKI showroom. By then, the cloud-based control interface for AGVs should be finished and would continue towards the integration and the injection of attacks. An in-depth work plan and time-line for UC2 will be worked out within WP5.

In terms of testbed access timeline, a preliminary plan includes the following incremental steps.

- Begin of Integration of Hardware into test site possible from M8.
- Collection and pipelining of the data from the UC2 beginning from M12.

- Setting up Federated Learning Infrastructure from until M18.
- Setting up test cases for attacks until M20.
- Integrated use case 2 rehearsal of demonstrations from M30.

4.5 Evaluation criteria

The evaluation criteria to be used to assess the proposed technologies essentially reside on the capability to detect attacks and other anomalies within the time-frame esteemed to be appropriate for UC2 operations. The success in the reach of KPI targets in terms of attack and anomaly detection accuracy and lag is accompanied by a sensibility analysis against a number of changing parameters, corresponding to reference UC2 IIoT applications that will be specified in the frame of WP5 activities.

4.6 Risks and potential issues

Table 4 reports on the identified risks and envisaged mitigation actions.

Table 4 Use case 2 identified risks and envisaged mitigation actions

Use case 2 Identified Risks	**Mitigation Actions**
Scarcity of high-quality data from UC2 testbed components.	Artificially added data built from an emulated environment at UC partner facilities.
Remote access to the testbed limiting the operations that can be made remotely.	Visiting weeks for researchers from Paris, Nancy and Trento to Kaiserslautern can be planned.
Impossibility to install additional hardware in the UC2 testbed.	Virtual circuits between academic partners networks in France and Germany can be installed thanks to National Research and Educational Networks (NRENs) infrastructure to add virtual nodes with the required hardware capabilities.
Instantiation of attacks is a necessary step before trying to detect them. It may require an effort and can face practical issues (e.g. lack of available attack scripts).	A first set of attacks will be rather simple to implement such as flooding (DDoS) or scanning with common tools.

5 Use case 3: Edge AI assisted monitoring of linear infrastructures using drones in BVLOS operation

The AI@EDGE connect-compute fabric, through the use of AI and Edge Computing combined with 5G, will contribute to the so-called '4.0 Industry' revolution in the industrial sector. In this use case, the use of drones in an industrial environment is investigated as a solution to this digitization process to open new doors towards more efficient solutions for surveying and monitoring of large surface areas.

5.1 Reference scenario

The monitoring of large areas (plots, farms, roads network) through the use of drones is a highly demanded service today, which however suffers from both practical and technical limitations that currently prevent a widespread application. One of the most relevant limitations are inefficient communications, both command and control (C2) communications and for remote transmission of images, data or information to be processed in a head-end computing infrastructure.

This UC3 is aiming to expand the AI@EDGE connect-compute fabric border to the drone embedded system, in order to use 5G capabilities to take care of the above-mentioned problems. The drone will be controlled in a BVLOS (Beyond Visual Line of Sight) mode through 5G, to scan industrial infrastructures, to make corresponding 3D modeling, then to identify the different incidents that could exist and to send notifications to the drone operator alerting that an incident has been found. Meanwhile, the information (images, telemetry) is sent in a continuous manner to the central office in order to improve the drone's operator decision-making process.

5.1.1 Actors, roles, stakeholders

The different actors involved in the use case scenario are as follows:

- Drone operator
- Infrastructure concessionnaires
- Technical staff
- Infrastructure users
- Drone manufacturer and maintainer
- 5G vendor for the communication infrastructure stack
- IT integrators for the management of the computing infrastructure including specialized hardware and software

Within UC3, the roles of the involved partners are as follows:

- AERO: AI@EDGE platform integration, UC3 coordination and development of drone automated monitoring functionalities based on AI and edge computing.
- ATOS: in charge of transferring WP3 and WP4 technologies to the development of drone automated monitoring functionalities based on AI and edge computing.

- EAB: will contribute to the transfer of WP3 and WP4 technologies; in coordination with its Ericsson Spain branch (ERI-ES), EAB will provide pilot testing facilities through the 5TONIC experimental facility located in Madrid (Spain) [28].
- ITL: in charge of integrating HW acceleration features in UC3 experimentations, in coordination with WP5 activities.

5.1.2 Use case Context

The general context of use case 3 is depicted in Figure 6.

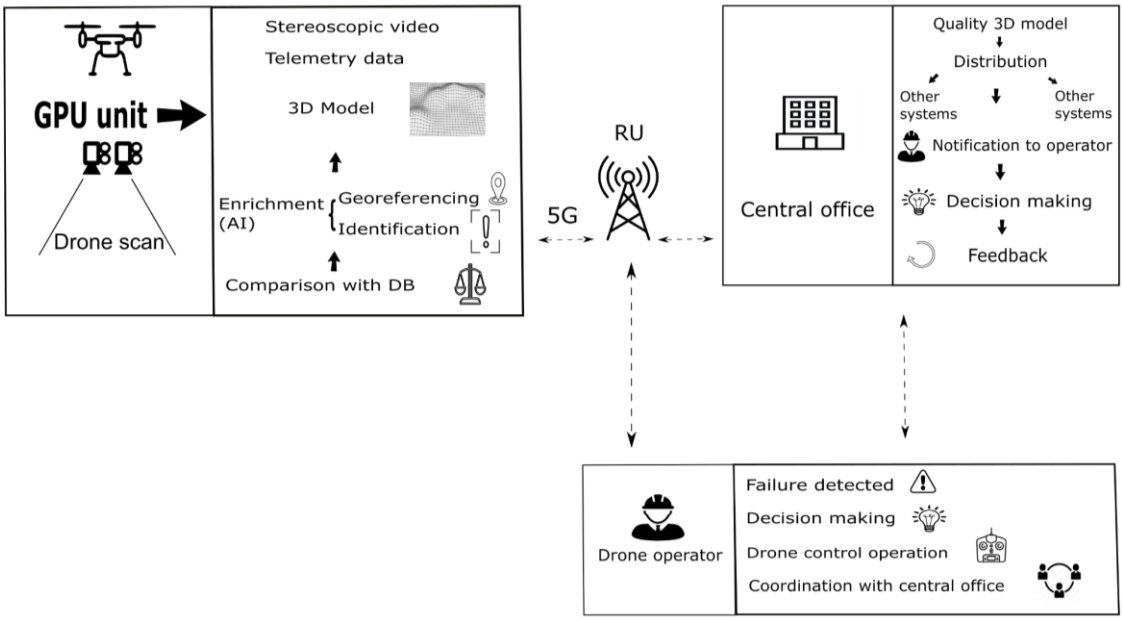


Figure 6 Use case 3 context

As explained before, this UC3 is aiming to embed part of the AI@EDGE platform on a drone in order to use 5G capabilities (e.g., near-zero latency, wider bandwidth and higher velocity) within an industry environment making use of AI and Edge Computing. The use case requires a composite monitoring Artificial Intelligence Function (**UC3 monitoring AIF**) to carry out infrastructure monitoring in real time. Our approach is to have the AIF distributed in order to support low AIF response time, redundancy and increase its availability guarantees, also taking into consideration that distributing the surveillance AI allows to locate the video streams to few different receptors. For this purpose, two different types of cameras are employed by the drone. The first one is stereoscopic cameras used to create 3D models and the second type is a FPV (First Person View) camera utilized to have a more direct control of what the drone sees.

This continuous view of the environment raises the level of security on the generated data. Following this reasoning, the FPV video will be continuously displayed on the drone operator's screen for security reasons. The operation to be performed is described in Figure 6: the drone scans the infrastructure elements, compares them with the reference models (previously stored in a database) and if any incident

is detected, it is identified and georeferenced. Then, a notification to the operator is sent and the decision-making process is done in a coordinated manner with the central office that is continuously receiving the information to help make the adequate decisions. While the drone waits for a response from the operator, it can perform several actions, i.e., it can keep orbiting with the camera aimed at the point of interest. At last, a feedback to the drone with new actions to take could be sent by the operator, and so on and so forth during the drone operations.

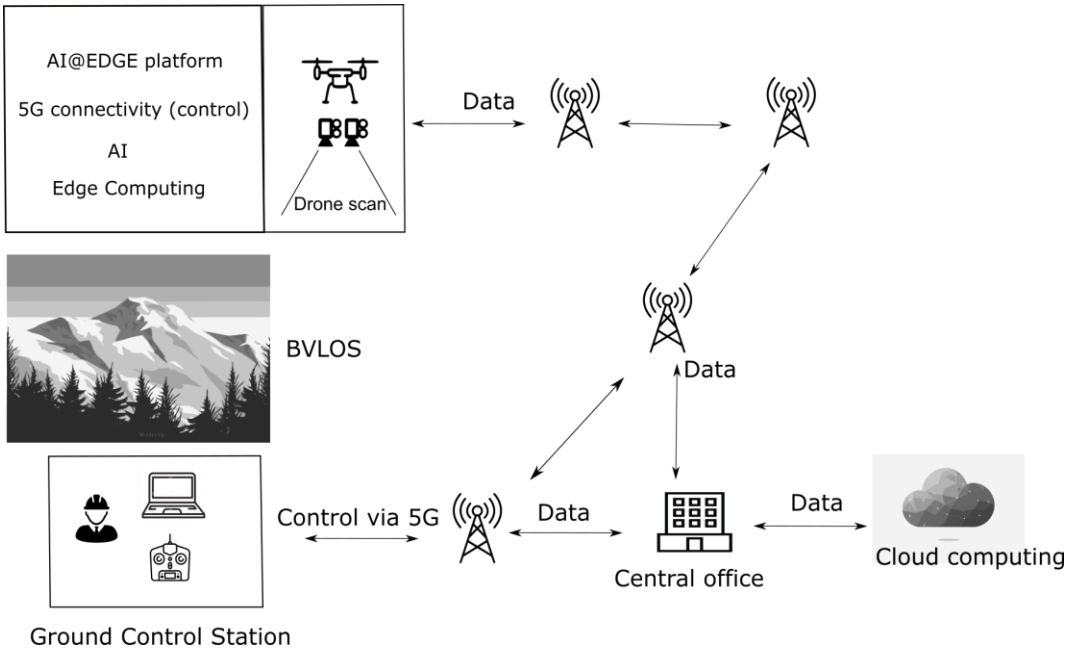


Figure 7 Example of use case 3 scenario

In a reference scenario, depicted in Figure 7, the drone is controlled via 5G in BVLOS mode by the drone operator. The figure describes the path followed by the data highlighting the continuous communication between the drone operator, the drone itself and the control center. This communication is bidirectional between these three elements in order to coordinate all the actions and take the best decision.

5.2 Main objectives

The following UC3 objectives are defined:

- Design solutions to use the AI@EDGE fabric to make linear or superficial infrastructures BVLOS monitoring seamless. With this objective we will be able to do surveillance, events detection (using AI), notification, tasks execution and computation, hence improving the overall decision-making and user-machine interaction process through Edge Computing.
- Conceive methods to use on-board or shore-based GPUs to create the environment 3D modelling, depending on the most optimal configuration to be found, and leveraging on the additional computational functionalities made available through Edge Computing.
- Determine how to integrate 5G systems for video distribution traffic. The traffic load, within video distribution, consists of enriched video images sent to the drone operator and the central office in real time to contribute to the decision-making process.
- Determine how to integrate 5G systems for drone control traffic. The traffic consists of drone primitive signaling and of telemetry data and FPV images for the operator.
- Specify how to leverage on 5G network slicing to allow an independent and isolated communication with the drone throughout the operation.

WP5 activities will allow linking to this objective a dedicated scientific and technical roadmap to reach them.

5.3 Use case architecture

The UC3 architecture, already represented and described in Figure 6 and 7, will leverage on the AI@EDGE fabric features. In particular, the following technological enablers forming the AI@EDGE fabric will be used in this use case:

- Distributed and decentralized serverless connect-compute platform.
- AI-enabled application provisioning.
- Network and service automation platform.
- Hardware accelerated serverless platform for AI/ML.

5.4 Requirements

In this section the UC3 requirements are presented, as grouped in terms of technical features, security and privacy aspects and KPIs.

5.4.1 Technical requirements

Network Bandwidth and Slicing: The required **5G radio bandwidth** will be proportional to the video definition and the number of users observing it through the 5G network, in this case the central office and the drone operator. With respect to **slicing**, a secure and isolated environment is required to prevent interferences with external operators.
Computing: An edge computing/AI device or system of devices that allows to make an onboard 3D monitoring in real time for the use case application.

Video data beamer bit-rate: the video stream bitrate needs a speed of at least 5 Mbps (HD) and it would be great to achieve Full HD or 25 Mbps (Ultra HD).

5.4.2 Security and privacy requirements

Two different **security** levels can be distinguished:

- High: The drone control channel must have the highest security level possible in order to limit radio interferences as much as possible.
- Low: The data transmitted (video, datalink) also needs to be protected but it has less importance in terms of security.

These levels could correspond to different slices, or even be within a given slice.

Privacy: The main restriction founded on this aspect are video-images recorded subject to data protection. These images should be only accessible by the drone operator and central office, so that no external agent can view or use them.

5.4.3 Key Performance Indicators (KPIs)

Four main KPIs are of particular interest for the use case and are detailed in the following.

- *Environment KPI***:** Range: geographical reach of at least 20 km (according to the state of 5G technology and deployment at the trials).
- *Drone operation KPIs:*
    - The *latency* KPIs sets as 100ms the maximum end-to-end latency budget. It is composed of two components:
        - Control Signal latency: it should be the lowest possible, and lower than 50 ms based on current awareness on the general system.
        - Video processing latency: it should be the lowest possible, so that the total end-to-end latency budget stays below 100 ms.

    A more precise assessment on the acceptable latency budget is needed to possibly update these preliminary figures, which will be done in WP5.
    - The *reliability* KPI (tentative metric) is in terms of control signal packet loss which should be lower or equal than 1%. This value is set because control signal must be ensured at all times.

- *AIF KPI:* Mean Average AI Precision in object detection: in the integration of AI-assisted drone framework on the 5G network, detecting incidents through AI analysis processed on-board and at edge-node to generate response action from centralized control station. The metric commonly employed to evaluate the performance of the model for automated detection of incidents in the scenario is the Mean Average Precision (mAP), with an Intersection over Union (IoU) equal to 0.5. This target KPI for the AI@EDGE project, according to the dataset used for the project, will be mAP@.5 >= 0.6 (defining classes as identifiable items such as "persons" or "vehicles") - mAP@.5 refers to the mean average precision at an intersection over union value of 0.5.

5.5 Use case testbed

The UC3 testbed will be built on top of the 5TONIC laboratory [28]; situated in Madrid (Spain), it is an open research and innovation laboratory focusing on 5G technologies. The testbed specific hardware allows configuring different network topologies of variable size and capacity that will be used to emulate a 5G network. It can provide a NFV infrastructure, 30 mini-PC computers supporting the experimentation with VNFs at smaller scale, as well as the management and orchestration of virtual machines. This setup allows the deployment and/or testing of different NFV/SDN domains, multi-layer control & orchestration, multi-tenancy NFV/SDN and multi-vendor NFV/SDN.

In this section it is explained how the 5TONIC laboratory will be extended for the UC3 testbed and employed to emulate the real scenario.

5.5.1 Platforms features requirements

In order to demonstrate the use case functionalities, the following requirements for the testbed platform have been deducted:

- Capability to ensure low latency.
- Support of data transfer monitoring.
- Reliability: Always prioritize the traffic of the control signal above the video signal reducing packet loss.
- Decision-making related to network performance and resource availability: to administrate video quality with respect to latency or any other network parameter that can significantly affect the use case.
- Support of HW acceleration: manage the different video signals and compute a basic 3D model on-board.
- Handover: In order to ensure the high-mobility condition of a drone that would require frequent change of node.

5.5.2 Hardware equipment

In order to demonstrate the use case functionalities, the following requirements for hardware equipment have been identified:

- UAS: Drone(s) and Ground Control Station (Radio-control command, PC, screen)
- Stereo camera(s) to make a 3D infrastructure monitoring.

- FPV camera.
- 5G connection device.
- Computational units: microcontroller, GPUs i.e., NVidia Xavier NX [37] or AGX Xavier [38].
- 5G network infrastructure, RU drone connection, data monitoring to track network behaviour, scenario emulation.

5.5.3 Application and software components

The following main software components are identified:

- AERO software or mission planification, drone control, image recognition, transformation of radiofrequency control command to 5G signal, 3D model generation.
- AI@EDGE platform software components.
- 5TONIC data monitoring software to track network behaviour, scenario emulation.

5.5.4 Preliminary testbed deployment and access time-line

Operations to be performed in the UC3 testbed include:

- Monitor and ensure the continuity of the control signal. For this purpose, different 5G network configurations will be tested and analysed, monitoring especially the 5G reliability KPI for different latency values of interest in the control cycle and its usual time-outs. Conditions of different test scenarios relevant to the application cases (distance, line of sight, etc) will also be configured and emulated.
- Monitor and ensure the quality-of-service level of concurrent communications of control and video signals. For this purpose, both the QoS perceived by the operator and in the 5G network the 5G latency KPIs, jitter, throughputs UL & DL will be monitored. Conditions of different test scenarios relevant to the application cases (distance, line of sight, interferences, concurrency with other services, etc) will be configured and emulated.
- Evaluate actions that, in certain events and circumstances to be defined relevant to the use case, can be taken by the network itself or by the application itself, determining their impact on the monitoring of the evolution before, during and after the event under study, of the application KPIs and of the 5G network.

The tentative testbed time-line is planned according to the project development structure:

- Check initial approach, state of the art of 5G deployment. Period: Fall 2021.
- First tests with the AI@EDGE platform, actualization with state-of-the-art technology updates. Period: during 2022.
- Final tests. Period: during 2023.

A more precise plan will be developed in the framework of WP5.

5.6 Evaluation criteria

To track the development of the project the three following criteria have been defined:

1. Being able to operate drone and onboard systems emulating the real scenario.
2. Being able to successfully identify incidents.
3. Being able to successfully send video images.

5.7 Risks and potential issues

The risks and related mitigations are reported in Table 5.

Table 5 Use case 4 identified risks and envisaged mitigation actions

Use case 3 Identified Risks	**Mitigation Actions**
Hardware: Weight and size of the physical device associated with the AI@EDGE platform to be integrated onboard allowing an optimal operation. Physical devices associated with the AI@EDGE platform have high power consumption, preventing acceptable flight time.	If the physical device associated with the AI@EDGE platform cannot be integrated onboard (either by weight and size or power consumption) its functions could be, partially or totally, offloaded to a ground station.
Scenario: Stability of radio access connections linked to the status of 5G deployment in the selected area for the use case.	Focus on AI@EDGE platform features that better suit the use case and its technical requirements.
Access to full range of functionalities: Related to network deployment in the selected area: 5G connexion is NSA instead of SA.	If the 5G status on the use case area is not optimal to efficiently perform the operation, repeaters to extend 5G coverage will be deployed.
Equipment on board: Interferences with other drone equipment (controller, GPS satellites, etc).	Pre-checks and damping measures in case of detected interferences or different positioning of the equipment inside the drone.
Connection: Data corruption between the operator and the drone prevents proper control of the last one because of the state of 5G connection.	Automated procedures will be applied until the drone regains connection i.e., intelligent RTL (Return To Launch) function will be fully available.

6 Use case 4: Smart content & data curation for in-flight entertainment services

6.1 Reference scenario

With the rapid market evolution of Inflight Entertainment and Connectivity (IFEC), the demand of passengers, airlines and airplane OEMs for data services is continuously growing, thus accelerating the digital transformation of the aircraft cabin. In this context, the mission of a next generation IFEC system is to aggregate, deliver and manage a curated (i.e. personalized) entertainment content and data experience to the aviation stakeholders (mainly airlines and passengers).

Currently, there is an expectation to provide dynamically curated content to the airline passengers, based on available data about, for instance, flight routes, passenger demographics known from personalized engagement channels (e.g. frequent flyer subscriptions) and other specific origin/destination information. These content sources include traditional IFE content (movies) and new sources of content such as live and near-live TV streams including news and sports. Additionally, sponsored content of all types form part of the overall content strategy. Such content involves eBook platforms, user developed content (the so-called passion economy), content from news aggregators, and other social media-oriented content (Reddit, Twitter, Facebook, etc.). To complete this picture, ground systems must provide the ability to ingest, process, and deliver such a myriad of real-time content options to the aircraft via satellite uplinks or mobile network connectivity when the airplane is grounded. The ingested content can be massive and stored in the on-board infrastructure, inside large libraries of continuously changing locally stored content for on-demand viewing. The content is stored and/or distributed throughout the aircraft to be consumed by passengers via their seatback system or off-the-shelf personal devices relying on the wireless connectivity on-board. Since aircraft backhaul technology mainly relies on costly satellite bandwidth, the possibility to perform as many tasks as possible at the aircraft edge, making use of AI and distributing content by means of 5G, stands for a clear path forward in the IFEC industry.

6.1.1 Actors, roles, stakeholders

AI@EDGE's use case 4 targets different types of actors and stakeholders, as follows.

- Equipment vendors (e.g., SPI).
- Airlines (e.g., Lufthansa).
- Aircraft OEMs (e.g., Airbus).
- On-board aircraft Content Service Provider (CSP) (e.g., Gogo, Sitaonair).
- Content producers (e.g., Hollywood, Netflix).
- Passengers (content consumers).

It is worth remarking that, besides SPI that leads use case 4, other already identified contributing partners include ATH, SRS, ITL, CNAM.

- ATH will in particular operate and integrate the 5GC.
- SRS will integrate the software radio access network elements.
- ITL will contribute with hardware acceleration.

- CNAM will work on the multi-connectivity aspects.

6.1.2 Use case context

At the onset, In Flight Entertainment (IFE) was introduced by airlines to deliver media content to passengers, but the system evolved ever since to focus on connectivity and broadband services, including Wi-Fi on-board and satellite backhaul toward the ground network. The combination of content and connectivity has indeed spurred the concept of IFEC, nowadays a major trend in the market of ancillary aviation services. With the consolidation of 5G, network softwarization and Artificial Intelligence, the cabin infrastructure, composed of servers, wireless termination points (e.g. 5G small cell) and seatback screens, has the potential to transform into a full scale smart edge-cloud. It is worth emphasizing that the on-board content curation is a major topic tackled by AI@EDGE use case 4.

As shown in Figure 8, an evolved IFEC system can offer to passengers different personalized bundles of content produced on ground by different over-the-top content producers. Nowadays, IFE mostly relies on mainstream movie makers. An ultimate IFE system may allow full customization in such a way that a passenger can bring its own content on-board without any physical media. Simply the content is curated for the passenger that can thus enjoy it on board for a full end to end experience. An example of such an end to end experience can be provided by any media content (movie, TV series, newspapers, etc.) that a passenger may be able to access on board in the same way it can be done through a ground telecommunications infrastructure.

If on the one hand such a level of content personalization is far from trivial, an advanced concept exemplifies the possibility to make a rough content selection from a ground pool of contents. The selection shall be made by a system as automated as possible that relies on an AI offline training against historical data of different airlines, source/destination information and passengers' subscriptions. The refined online curation of the content instead takes place on board in the edge-cloud infrastructure relying on light-weight AI content selection until a personalized bundle of media contents is offered to a passenger to the extent possible. In the context of airplane infrastructure with limited resources, computationally light AI approaches are crucial to do a wise use of the network-compute fabric and reach fast convergence time in content curation. In this regard, use case 4 targets to also include hardware acceleration to validate enhancements to the cabin edge-cloud infrastructure to carry out possibly computation intensive AI on-board.

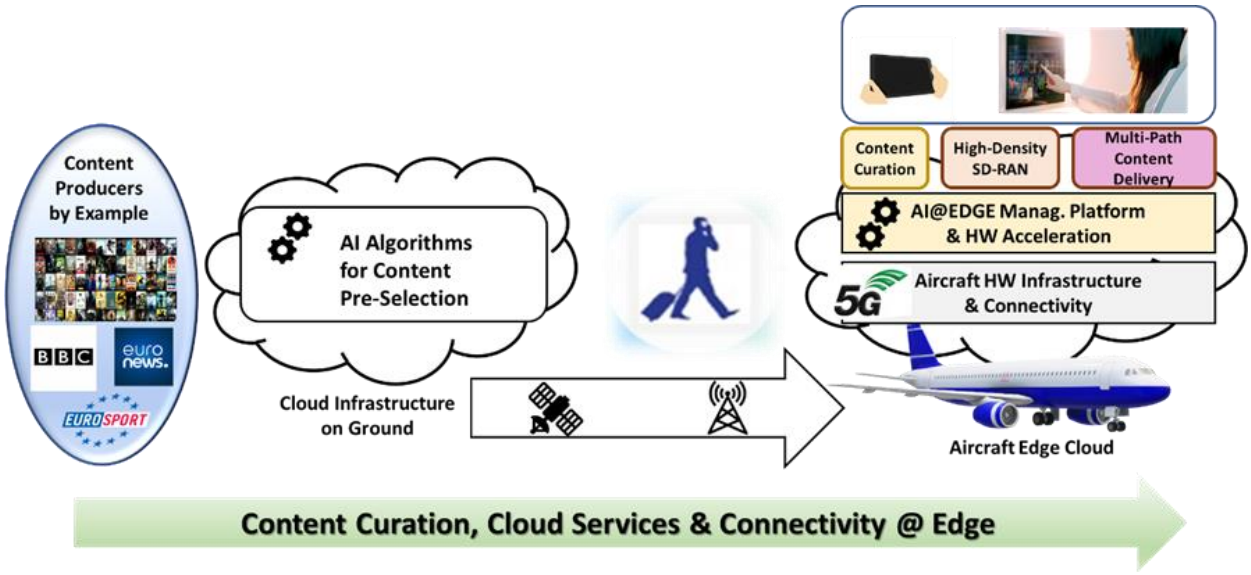


Figure 8 Use case 4 context - End-to-end view leveraging on the AI@EDGE concepts and approach

6.2 Main objectives

The AI@EDGE use case 4 targets to achieve the following main objectives during the development and consolidation of the use case infrastructure:

- Demonstrate content curation to airline passengers for personalized content consumption.
- Demonstrate the feasibility of MEC and microservice-based serverless computing to enable a disruptive approach for the next generation IFEC system that will accelerate the digital transformation of the cabin.
- Develop an IFEC system that can harness content curation, content loading and streaming.
- Demonstrate the AI@EDGE concept with both COTS and aero-certified hardware at the servers side and clients side, respectively aero-certified servers and seatback screens and personal devices such as tablets and smartphones.
- Develop and trial an autonomous edge-cloud computing platform that avails AI with reduced human intervention.
- Demonstrate the possibility of fast service deployment as an indication of the customizability of the new platform as compared to the legacy system.
- Prove the feasibility of storage-as-a-service on-board, which opens the opportunity to a variety of existing and new entrant stakeholders to share and use the IFEC infrastructure.
- Demonstrate 5G connectivity on-board, including mobile core network and radio access, with the end goal to demonstrate 5G Stand Alone mode (5G SA).
- Demonstrate superior availability of the communication path between the content and an end-user relying on multi-path communication capability provided by MP-TCP (including wired and different types of wireless technologies), in possible conjunction with layer-2 aggregation techniques.

6.3 *Use case architecture*

The vision developed by use case 4 for smart content and data curation is shown in Figure 9. The figure shows the different hardware segments and software tools in their complex relations and interplay. On the edge-cloud side, the bare metal substrate is assumed to be made of hardware manufactured and certified by SPI according to the stringent regulations of the aviation sector and off-the-shelf hardware. The testbed thus includes one or more aero-certified IFE servers, as well as several aero-certified IFE seatback screens. At the same time, COTS servers shall be also included in the cloud infrastructure, thus yielding a heterogeneous hardware infrastructure. Pre-existing radio access is made of off-the-shelf Wi-Fi access points and SPI manufactured and certified access points. But such a radio access configuration will be enriched with 5G connectivity going through NSA and finally SA versions. Personal devices such as tablets and smartphones shall be also included for demonstration.

The overall substrate constitutes the NFVI wherein the edge-cloud platform is deployed to host both the cloud management & orchestration and value-added services, with content curation of utmost importance. The cloud system developed within AI@EDGE is based on Kubernetes[1], which defines the necessary level of abstraction and internal relations within the edge-cloud. On the hardware substrate the necessary abstractions are developed to enable different levels of operation in a containerized manner. A content curation AIF (**UC4 content curation AIF**) is used to select the content on a per-user or per-group of users based on popularity of contents and availability of network and radio resources. Further, it is envisaged that the content curation AIF will be deployed as a distributed set of AIF instances and possibly making use of hardware acceleration. At infrastructure level, while a containerized deployment of the non-real-time RAN intelligent controller is deployed to provide joint network self-optimization features, some functionalities that sense and manage the infrastructure should be running as xAPP at the RIC level.

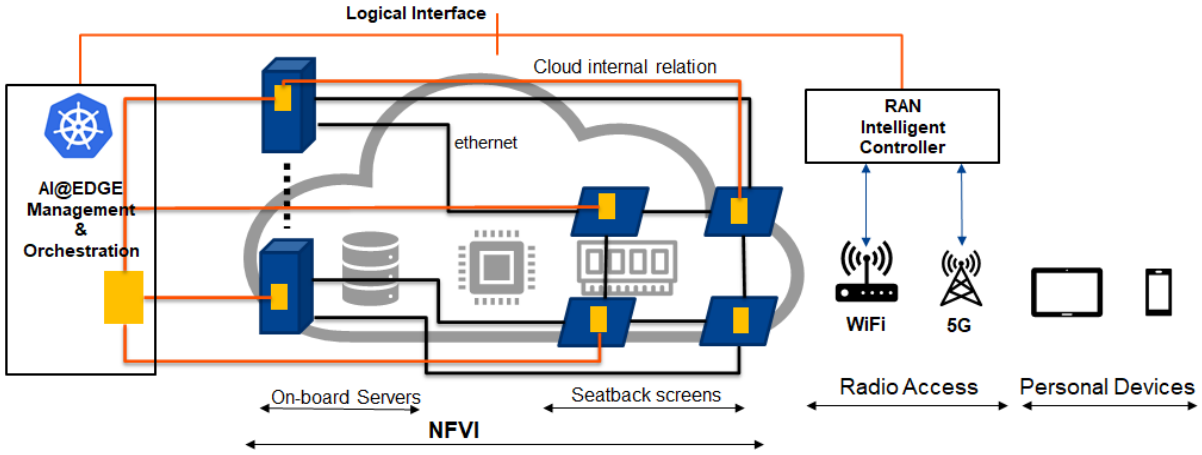


Figure 9 use case 4 high-level system architecture

1 Please see:https://kubernetes.io/

6.4 Requirements

6.4.1 Technical requirements

As part of the AI@EDGE system, use case 4 per se creates the demand for the following requirements in order to show improvements over the state-of-the art IFEC platform.

- Bandwidth requirements of wired (i.e. Ethernet) and wireless links in the order of gigabit per second for content streaming and fast content upload (i.e. 5G RAN).
- Slices creation to accommodate optionally different content providers (see Section 6.1.1).
- AI@EDGE platform based on a container orchestration such as Kubernetes to make efficient use of limited resources of the edge connect-compute infrastructure on-board.
- Infrastructure and service monitoring.
- Enable artificial intelligence at infrastructure level and at overall service level for system management.
- Connect fabric mainly based on 5G, but integrating also Wi-Fi and Ethernet to deliver customised content (e.g. video content) to passengers through the curation system.
- MPTCP [15] capability and proxy network functions on-board to harness multiple communication paths and technologies (wired and different types of wireless connections).

6.4.2 Security and privacy requirements

During its development and testing, use case 4 shall rely on anonymised datasets coming from data collected on real flights in collaboration with different airlines that are currently customers of SPI. Such data will be used to train the recommendation system and to identify passengers' preferences for content selection and distribution based on AI. In a more elaborated setting that involves the content curation service, datasets to be processed by the recommendation system shall be considered both of offline and online types. The offline type of data shall be used with the main purpose to train the system to thus avoid a 'cold start' on-board. The online type of data can be used for finer adaptations of the content curation to produce a true personalized package of media contents for each passenger. In the planned implementation of use case 4, the possibility to use the infrastructure to generate online data implies a non-trivial effort that requires automation through a testing suite. Such an effort can be therefore re-evaluated only during the test bed development.

Given the consideration above, SPI, as use case leader, does not expect any critical challenge regarding privacy of data. In a more realistic situation of a flight operated by an airline, the content curation should be envisaged as an ancillary service that the airline can offer to the passengers. This could happen during the ticket reservation phase or rely on the personal data of frequent flyers. In both cases, it is reasonable to assume that a passenger can choose whether to use the service or not and provide explicit consent to the processing of personal data in full respect of the European privacy regulations (i.e. GDPR). Even nowadays, the possibility to navigate contents through the IFE system by means of an airline App can be done, and by agreeing to use the App an explicit consent shall be provided by the passenger. In terms of security, until now the IFEC infrastructure has shown to be very secure and no hacking of the system has been detected so far for SPI products.

6.4.3 Key Performance Indicators (KPIs)

At the current stage of developing the use case, SPI has identified the following KPIs that target passengers, considering that such KPIs are typically set for the existing legacy IFE system that is hinged around Gigabit Ethernet connectivity. Performance indicators for the wireless system are usually relaxed and understood as a function of the number of users accessing content through the radio access network on board. Since wireless connectivity typically relies on different Wi-Fi generations, the amount of connected users can be a critical factor.

Preliminary quantification of key performance indicators for use case 4 is provided below.

- *Data rate/client for streaming*: > 15 Mbps (4K video).
- *Data rate/client for content loading*: ≥ 200 Mbps.
- *Average aggregate throughput*: ≥ 20 Mbps/sqm.
- *Service deployment time*: few minutes.
- *Curated content delivery time*: ≤ 180s for a consistent quality of experience of a passenger.
- *Service recovery time*: ≤ 180s to avoid service disruption with consequent dissatisfaction of passengers, airline customers and OEMs.
- *Number of served passengers*: ≥ 12 for demonstration.
- *Content curation precision of recommendation:* >= 80%, based on an initial estimate. A dedicated numerical analysis will be conducted on this KPI.

6.5 Use case testbed

The testbed that is targeted by the end of use case 4 is shown in Figure 10 from the infrastructure standpoint. The management and orchestration system developed within AI@EDGE, as well as other network services are not shown at this stage since already discussed in other sections (e.g. Section 6.3). Hardware specifications instead are provided in Section 6.5.2. The idea is to build the testbed combining the SPI cabin mock showroom with an evolved test set up starting from the initial testbed that is shortly described in Section 6.5.4. The software components for developing the edge-cloud, the content curation AIFs for content personalization and recommendation and for edge platform management and the other features mentioned in Section 6.4.1 pose significant challenges that will be tackled throughout AI@EDGE evolution.

As it will be later clarified in Section 6.5.4, the use case 4 is already assumed split in a ground and aircraft part, with the focus of the use case on the aircraft infrastructure. While remote access to testbed partners can be granted through a VPN client such as Cisco anyconnect, a test rack made of several SPI manufactured seatback screens, one SPI manufactured IFE server and COTS servers will be built to integrate AI@EDGE edge-cloud tools and to demonstrate the overall use case. A major step forward in this regard is expected with the deployment of a container-based management system based on Kubernetes, which constitutes the basis for the container orchestration, networking and performance monitoring for the AI@EDGE framework. Radio access currently made of Wi-Fi access points and 4G connectivity shall advance toward 5G that shall be demonstrated in the airplane cabin mock up.

Already in this early stage of the use case 4 elaboration, a clear challenge was identified for the cabin edge-cloud since it will be developed on top of a heterogeneous hardware substrate. Since the ambition is to

include also aero-certified seatback screens besides aero-certified and COTS servers, two important aspects ought to be taken into due consideration: the limitations that arise from the embedded operating system installed in aero-certified equipment and that the seatback screens per se create a challenge for the cloud development. Typically, seatback screens are assumed as client devices not included in the cloud infrastructure, whereas in UC4 development the target is to bring them inside the edge-cloud platform. Such a peculiar set up, specific to an evolved airplane cabin network, is motivated by the fact that a large number of seatback screens (in the order of two hundreds) is typically available on board. Each seatback screen is a relatively powerful device, constantly powered after take-off and the collection of all screens can participate to carry on with different workloads.

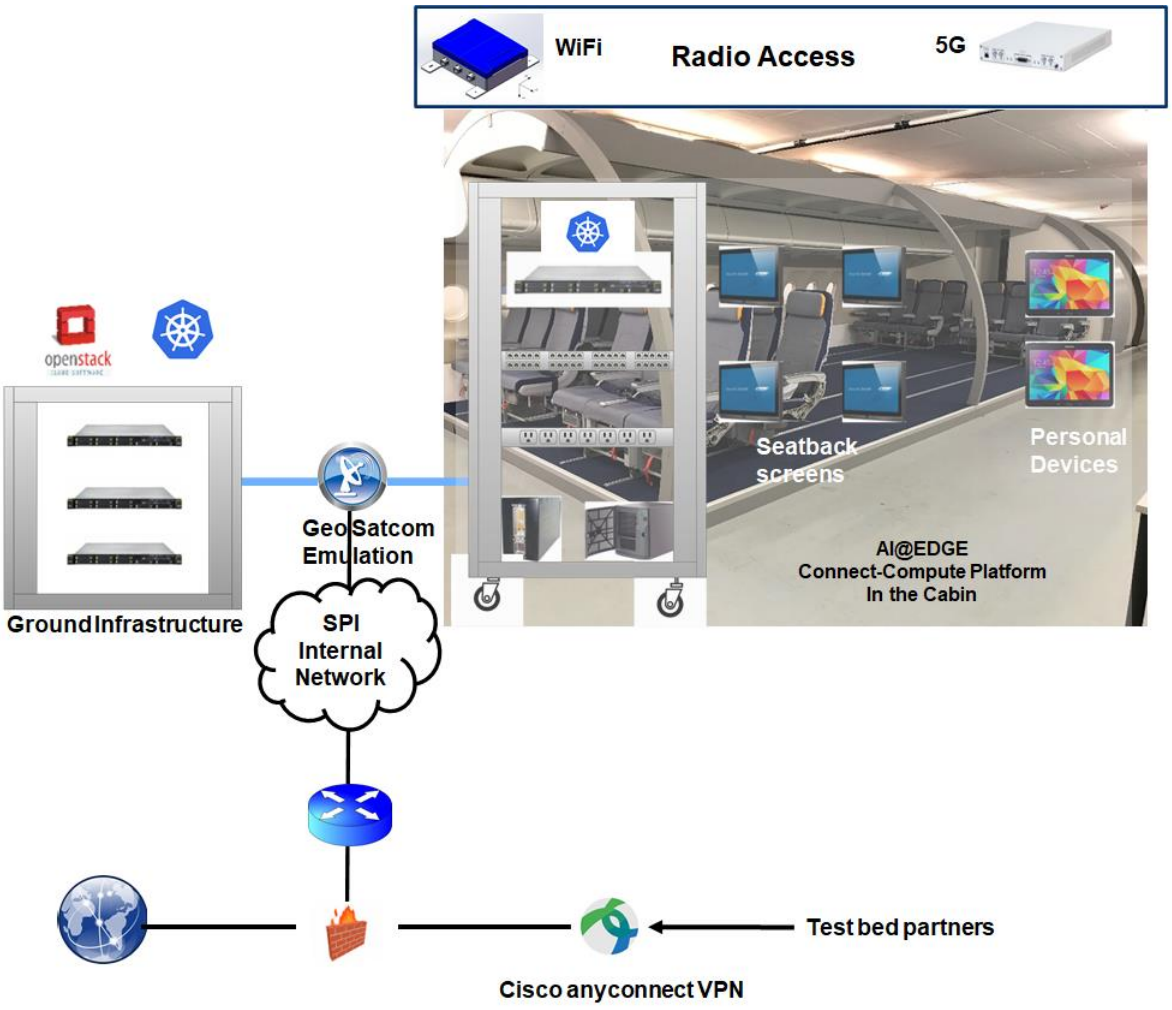


Figure 10 Use case 4 main hardware components

6.5.1 Platforms features requirements

The general AI@EDGE platform that shall be used and tailored to reach the main objectives of use case 4 (see Section 6.2) shall comply with the technical requirements mentioned below. The requirements are mapped taking into account that the existing IFEC system commercially rolled out does not carry any of the features mentioned since 5G connectivity, edge computing and on-board content curation are very novel in the context of passengers' entertainment and connectivity.

- MEC connect-compute platform and interfaces, or any part thereof.
- 5G RAN connectivity and 5G Core Network with the option of local breakout.
- Relieve utilization of satcom bandwidth pre-selecting on ground the content to bring on board and content curation.
- Adoption of a RAN intelligent controller.
- Fast deployment service, service monitoring and fast service healing time.
- AI-enabled AI@EDGE platform for self-healing and self-optimizing infrastructure with minimum human intervention of non-specialized people (e.g. cabin crew).
- Performance monitoring at platform and service levels.

6.5.2 Hardware equipment

The testbed that will be developed by UC4 shall comprise different hardware capabilities as already mentioned. The hardware components provided below shall be considered as the main ones currently identified and as such may be subject to variations and/or adaptations during the use case development.

- SPI aero certified IFE server (at least one),
- SPI aero certified seatback screens (at least eight),
- SPI aero certified Wi-Fi access point (one),
- A320 cabin mock up (one)
- Wi-Fi access points (one – two)
- Supermicro X10SDV-12C-TLN4F COTS servers (at least one),
- Intel NUC small servers with Intel i7-6770HQ processor (at least two),
- Fujitsu Primergy servers with Intel Xeon X650 processor (at least one),
- Off-the-shelf 5G client devices (two – four),
- Software-defined radio X310 (two),
- GPU NVIDIA PCIe card (at least one),
- SIM cards (several),
- SIM card programmer (one).

6.5.3 Application and software components

The Smart Content & Data Curation for In-flight Entertainment Services use case shall deploy the following software components that enable the main objectives described in Section 6.2.

- Software for AI-based content curation (i.e. recommendation system) for passengers,
- AI@EDGE edge-cloud management system based on Kubernetes,
- AI@EDGE platform automation tools based on artificial intelligence functions,
- Platform monitoring tools,

- Optional external orchestrator,
- Near RT and Non-RT RAN Intelligent Controller modules,
- 5G virtualized mobile core network with the option of local breakout,
- 5G base station software,
- Software constructs for multi-path connectivity.

6.5.4 Preliminary testbed deployment and access time-line

Figure 11 shows the use case 4 initial testbed set up that will be the extension of the testbed used in two previous H2020 projects: Sat5G [29] and 5G ESSENCE [30]. The infrastructure is split between a ground segment and an aircraft infrastructure segment. The testbed playground is provided by the cabin mock up, a realistic reproduction of a single aisle A320 airplane cabin. Both ground and aircraft infrastructures are part of an Openstack-enabled cloud and are connected together through an emulated GEO satellite connection. The ground segment is made of four rack servers of the Fujitsu Primergy series, whereas the aircraft segment includes different types of server. Namely, one system control unit that stands for an aero-certified server manufactured by SPI, two Supermicro servers X10SDV-12C-TLN4F and four Intel NUC small servers equipped with Intel i7-6770HQ processor. Apart from the SPI manufactured server, all the others use Linux OS Ubuntu distribution 20.04 server. In terms of existing access network, two Wi-Fi access points connected to 5G EmPOWER [53], one 4G CASA Systems small cell and one National Instruments USRP B210 are made available to deploy a 4G cell based on SRS software. Clients include SPI manufactured aero certified seatback screens currently enabled with both ethernet and Wi-Fi connectivity, several Raspberry Pi 3, commercial Android-based tablets and smartphones that are enabled with Wi-Fi and 4G connectivity. Access to the existing SPI testbed infrastructure can be granted to AI@EDGE partners through a Cisco anyconnect VPN connection.

It is worth emphasizing that the AI@EDGE UC4 infrastructure shall depart from the current one and progress toward the testbed shown in Figure 10, still retaining some key elements both on hardware and software sides, whereas the edge-cloud management & orchestration for the aircraft infrastructure shall be developed within AI@EDGE.

In terms of testbed access timeline, a preliminary plan includes the following incremental steps.

- Remote access to the infrastructure described above and shown in Figure 10 can be made available immediately for preliminary tests,
- The first release of the cabin infrastructure evolution shown in Figure 11 is expected in **M13**,
- Initial access to the evolved test infrastructure shown in Figure 11 is expected to be made available from **M20** onward,
- Integration with partners' services from **M22**,
- Integrated use case 4 rehearsal of demonstrations from **M30**.

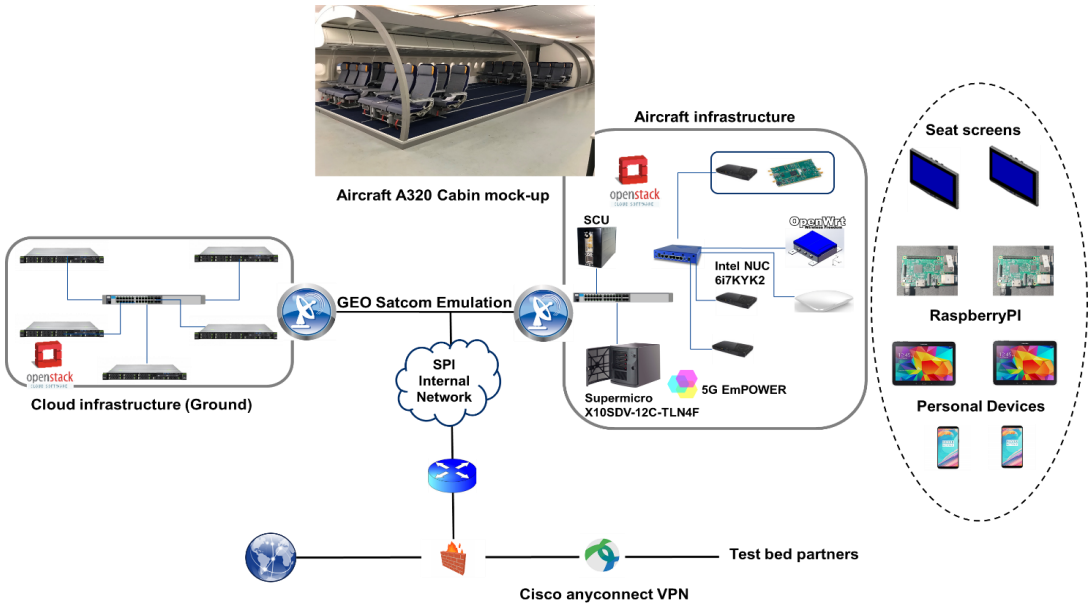


Figure 11 Preliminary testbed for use case 4 development

6.6 Evaluation criteria

In the development of use case 4, SPI has devised different types of evaluation criteria that include edge-cloud infrastructure and use case specific achievements. As such, during the development of use case 4 the following milestones are currently envisaged, which shall also be aligned with the estimated testbed access timeline provided in Section 6.5.4.

- Assembly of the evolved test bed infrastructure as shown in Figure 10,
- Installation of Docker engine and Kubernetes on COTS servers, Aero certified server and seatback screens upon necessary modifications/adaptations of the operating system,
- Adaptation of Aero specific software (e.g. screens operating system) to accommodate the AI@EDGE management & orchestration platform based on Kubernetes,
- Integration of a GPU card for hardware acceleration,
- Deployment and testing of the content curation AIF in the on board edge-cloud platform for infrastructure management,
- Deployment and test of 5G mobile core and radio access for superior data rate performance,
- Deployment of hardware/software specific components for multi-path connectivity tests,
- Making available in the testbed a pool contents on which to apply the curation system,
- Demonstration and test of a content curation system,
- Delivery of a personalised package of media contents to different seatback screens.

6.7 Risks and potential issues

SPI, as use case 4 leader and one the worldwide leading vendors of IFEC has preliminary identified the following risks provided in Table 6 on the way to develop the use case, alongside with initial mitigation effort. A continuous monitoring process and interaction with use case and project partners shall take place to ensure proper development of the technical activities, as well as to identify unexpected problems.

Table 6 Use case 4 identified d risks and envisaged mitigation actions

Use case 4 Identified Risks	Mitigation Actions
Core tools of the AI@EDGE platform turn available late in the project.	It is envisaged that project-wide decisions shall be required in the occurrence of such a situation.
AI@EDGE platform integration with aero certified hardware/software turns out to be more difficult than expected.	In this case the use case development will rely on COTS hardware, and where possible develop adaptation of the operating systems installed in the Aero certified equipment in order to deploy the AI@EDGE platform.
The edge-cloud management platform does not meet the specific use case requirements.	Focus on AI@EDGE platform features that better suit the use case and its technical requirements.
AI-enabled platform features require additional work to adapt to use case 4 hardware/software specifications.	Focus on key platform features only.
Target use case KPIs cannot be met.	Relax some KPIs whether possible without compromising use case effectiveness and demonstrability.
5G standalone mode requires extra effort.	Begin testing activities with 4G and 5G NSA.

7 From Requirements to Preliminary Specifications

We summarize in Table 7 the UC technical requirements specificities, differentiating between the Device (at/next to end-users), RAN, Core Network, Edge and Remote cloud, and the AIF domains.

Table 7 Synthetic view on AI@EDGE use case technical requirements specificities

Technical domain	Use case 1	Use case 2	Use case 3	Use case 4
Device	Driving emulator. Monitoring of psychological human signals. On-board telematic boxes.	Standard sensors. Automated guided vehicles (AGVs).	Drones with stereo and FPV cameras. Ground control station with NVidia Xavier NX or AGX Xavier GPUs.	Aero certified custom linux screens and android tablets nodes. Raspberry PI. Personal devices. Wi-Fi access points. Ethernet backhaul.
RAN	Uu and PC5 interfacing. C-V2X on the 5.9 Ghz band. Handover management needed to cope with high mobility.	AirSpan OpenRAN. N-78 Band (3.7-3.8 GHz). Handover management needed to cope with factory AGV mobility.	No band limitations. Handover management needed to cope with high mobility.	SatCom links for Ground communications. No 4G/5G band limitations. No handover management needed.
CN	Handle and forward data collected via Uu and PC5 interface. 5G virtualized core network with the option of local breakout.	Druid Core Network. Local breakout for traffic routing capabilities towards the edge cloud.	5TONIC facility.	5G virtualized mobile core network with the option of local breakout.
Edge cloud	Co-located MEC servers.	Serverless computing and orchestration. P4-NetFPGA smart-NICs	5TONIC facility. Integration of Hardware acceleration to be investigated.	Intel NUC servers. Fujitsu Primergy and supermicro servers. GPU NVIDIA cards.
Remote/far cloud	GPU equipped servers	GPU and FPGA acceleration at FL central server.	None.	Ground OpenStack cluster.
AIF	Traffic Controller AIF	Security AIF	Monitoring AIF	Content Curation AIF

An aspect that Table 7 is not detailing is which orchestration and re-orchestration actions are needed by which use-case and at which domain. Based on the current awareness of use-case technical challenges, scheduling adaptation at the RAN layer, scaling at the edge cloud and CN domains, adaptive AIF duplication at the edge cloud and remote cloud layer. The orchestration aspects, and related requirements on a per-use case and per-layer will be further investigated in WP3, WP4 and WP5.

In terms of security and privacy aspects, Table 8 summarizes specific security and privacy requirements for the four AI@EDGE use cases.

Table 8 Synthetic view on AI@EDGE use case security and privacy requirements

Security/Privacy requirements	Use case 1	Use case 2	Use case 3	Use case 4
Network and System Security	The traffic controller AIF can suffer from radio jamming.	Attacks against federated learning to be emulated to assess AIF robustness.	The drone control channel must have the highest security level possible in order to limit interferences as maximum.	Certified equipment highly secure against attacks.
Privacy	Privacy preservation needed in the exchange of messages, because sensitive and personal information may be exchanged. Stored that of the digital twin system should also be proetcted.	Analyzed data privacy preserved by the locality of that at/next collection point thanks to federated learning.	Recorded video-images subject to data protection. Drone data transmitted (video, datalink) needs to be protected.	Anonymised datasets coming from data collected on real flights in collaboration with different airlines that are currently customers of SPI

We summarize in Table 9 the use case KPIs, grouped in terms of networking, computing, AIF precision, reliability and environmental factors.

Table 9 Synthetic view on AI@EDGE use case KPIs

KPI group	Use case 1	Use case 2	Use case 3	Use case 4
Networking	Sensor-to-vehicle latency lower than 100ms.	Standard 5G targets for mMTC.	Control Signal latency lower than 50 ms. Video latency between 50 and 100 ms.	Data rate/client for streaming: > 15 Mbps (4K video), Data rate/client for content loading: ≥ 200 Mbps.
Computing	The above latency budget includes data processing latency. One configuration to support a virtual Road Side Unit on MEC side can be a 8-core ARM 64-bit CPU with GPU and 32GB RAM	State of the art virtualization servers with at least two PCI slots for the NetFPGA and with no specific minimum computing requirements.	The above latency budget for the control signal includes data processing latency. For the compute nodes: state of the art processor with a minimum capacity equivalent to a NVIDIA Quad-core ARM A57 @1.43 GHz - RAM equivalent to 2GB and 25 GB/s	Service deployment time: few minutes. Curated content delivery time: ≤ 180 s.
AIF	Vehicle trajectory features estimation with an accident avoidance rate higher than 99%.	Attack detection lag from few ms (known attacks) to few minutes (zero-day attacks).	mAP@.5 >= 0.6	Content curation system precision of recommendation >= 80%
Reliability	System reliability higher or equal to 99%.	Attack detection accuracy ≥ 97%.	Control signal packet loss lower or equal than 1%.	Conventional URLLC requirements. Content delivery service recovery time: ≤ 180 s
Environment	At least 50 vehicles per km2.	-	Reach of at least 20 Km range.	Average aggregate throughput: ≥ 20 Mbps/sqm.

The previous three tables are therefore spotting the specific requirements that the AI@EDGE fabric shall meet, going beyond baseline 5G service requirements. In the next section we detail the preliminary activity started toward the network automation and connect-compute platform design, making reference to these use case requirements to justify the technical decisions taken so far, where needed.

8 System-level and functional architecture

In this section, we provide a preliminary description of the AI@EDGE system architecture. These preliminary considerations form the main corpus for project milestone M2.1 (due at M6, June 2021). Their development, in both the algorithmic implications (learning, orchestration) and connect-compute fabric specifications, will be henceforth the object of activities in:

- Task 2.2 and later documented in D2.2 ("Preliminary assessment of system architecture, interfaces specifications, and techno-economic analysis"), for the system architecture and interface specification,
- WP3 and D3.1 ("Initial report on systems and methods for AI@EDGE platform automation"), for the federated learning and automation loop challenges, and
- WP4 and D4.1 ("Design and initial prototype of the AI@EDGE connect-compute platform") for the connect-compute fabric.

8.1 Preliminary system architecture

The AI@EDGE platform is meant to support all aspects of network and service management including the full orchestration and lifecycle management (onboarding, deployment, migration/scaling, and termination) of the AI-enabled applications over a fully distributed facility including also the ancillary tasks that needs to be performed to enable such lifecycle management, e.g., the creation and management slices.

To achieve this goal, AI@EDGE combines a mix of cutting-edge cloud computing (cloud-native, serverless computing, and hardware acceleration) and 5G concepts (disaggregated RAN and multi-connectivity) with a secure and privacy preserving AI/ML layer with the goal of providing a network and service automation platform that can be used to both manage a beyond 5G network infrastructure as well as to manage the value-added application and services running on top of it. Figure 12 depicts the reference AI@EDGE system architecture.

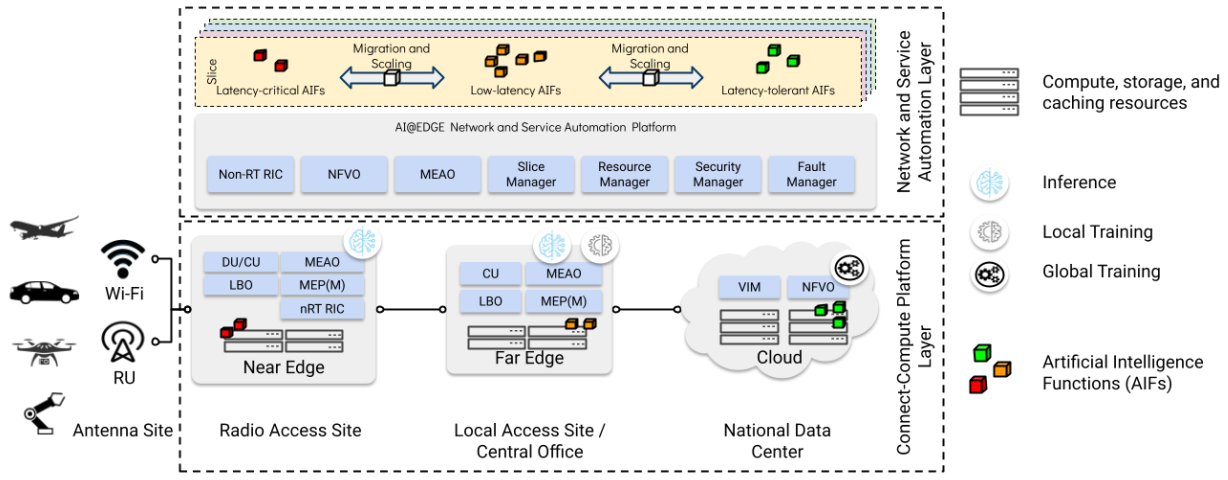


Figure 12 AI@EDGE Reference System Architecture

AI@EDGE leverages a multi-layer cloud/edge architecture spanning from end-user terminals to the centralized public/private cloud. Multiple edge computing layers are also foreseen. In particular as it can be seen in the reference system architecture, the near edge is deployed at radio access sites with the goal of providing coverage at city block level, while the far edge cloud is deployed at the local access site or at the central office site which provides coverage at the city or metropolitan level. Finally, the centralized cloud could be either a public cloud (e.g., Amazon Web Services, Microsoft Azure, etc.) or the MNO's regional/national datacenter. The reference system architecture also indicates how AIFs and AI/ML tasks such as inference, local training, and global training could be deployed across the AI@EDGE architecture. More details about these aspects will be provided in the following sections. Notice how both far and near edge clouds feature a Local Break Out (LBO) which can be implemented using an User Plane Function in case of a 5G core or a S/P-GW breakout in case of a 4G Core.

Section 8.2 will discuss the aspects of the AI@EDGE architecture related to the network and service automation platform. It will do so by first covering the challenges associated with the realization of closed control loops. Then it will introduce the AIF conceptual model leveraged by the AI@EDGE network and service automation platform and will conclude by describing the challenges associated with federated and distributed ML. Section 8.3 will discuss the aspects of the AI@EDGE architecture related to the connect-compute platform. The section will analyze the various sub-components of the platform, including the serverless connect-compute fabric, the provisioning of AI-enabled applications, the multi-connectivity aspects, and finally the hardware acceleration features.

8.2 Network and Service Automation Platform

This section will report on the technical and architectural aspects as well on the research challenges tackled by the AI@EDGE network and service automation platform. The section will begin with the introduction of the conceptual model of the AIFs, then it will describe how they fit into the AI@EDGE Closed Control Loop Intelligence, and finally it will describe the aspects of the network and service automation platform related with distributed and federated learning. The AI@EDGE network and service automation platform will be discussed in greater detail in D3.1 and D3.2.

AI@EDGE leverages the concept of reusable, secure, and trustworthy AI for network and service automation in industry relevant multi-stakeholder environments. In order to achieve this goal AI@EDGE will prototype and validate a network and service automation platform capable of supporting flexible and programmable pipelines for the creation, utilization, and adaptation of secure and privacy-aware AI/ML models.

AI@EDGE aims at enabling the full potential of mobile edge computing, which will be done by developing the mechanisms required for scalable distributed and federated learning in a 5G/6G context. Distribution of training/inference tasks, of AI-enabled applications, and of closed loop control tasks are key to realize AI/ML-driven applications at scale. To achieve this, AI@EDGE will develop a platform for closed-loop automation for deploying AI/ML compute infrastructures over the edge.

Altogether, the final goal is to enable zero-touch service management and network operations through the end-to-end creation, utilization, and adaptation of reusable AI/ML models also accounting for the resource

availability and failure resilience. The platform needs to ensure the support for the (re)scaling of distributed AI/ML algorithms and the associated control loops to ensure the required application performances under the varying and heterogeneous edge resource availability conditions.

Artificial Intelligence Functions Conceptual Model

AI@EDGE promotes the vision of a new generation of AI-enabled applications obtained through the chaining of multiple AIFs across a converged connect-compute platform. With the term AIF, we refer to the AI-enabled end-to-end applications subcomponents that can be deployed across the AI@EDGE platform.

AI@EDGE aims at provisioning AI-enabled applications over a distributed connect-compute platform where applications and services are dynamically orchestrated. In AI@EDGE we introduce a conceptual model for AIFs capable of capturing and representing in generic terms the common aspects of a component of an AI-enabled application. Figure 13 depicts such a model from the point of view of the interfaces exposed externally.

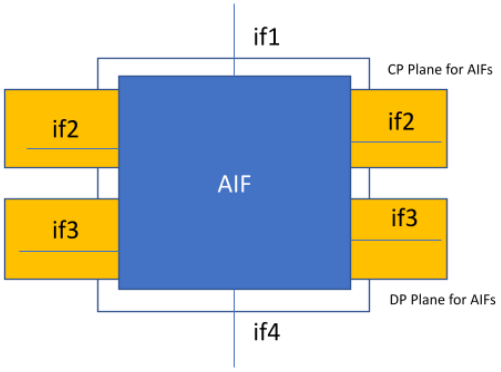


Figure 13 The reference AIF model

In particular, we identify the following interfaces:

1. **if1**, is the northbound interface used for (re)configuring the AIFs. Its semantic is defined by the specific component the AIFs is implementing. For example, in the case of a mobility management application this interface could be used to set signal strength threshold below which a mobility management event should be triggered.
2. **if2**, is the ML control plane interface used to exchange model parameters. This is meant to support distributed and/or federated learning scenarios.
3. **if3**, is the ML data plane interface used to exchange the data on which the ML model is actually applied. For example, in the case of a load balancing application this could be the stream of radio channel quality (e.g., RSRP/RSRQ) measurement originating from the RAN.
4. **if4**, is the ML southbound interface used to (re)configure another entity. This could be for example the hardware on which the ML model is running or an external SD-RAN controller. The format of the interface is the one exposed by the external entity.

In Figure 14 we report an example of an application built using the reference AIF model. The example models a load balancing application. In this application downlink traffic to a given group of UEs needs to be load-balanced between multiple radio access technologies. However, in order to do this efficiently it is important to estimate the downlink bitrate that each of the radio access technologies will be able to deliver. As one can see from the figure, the stream of RAN data coming from different O-RAN nrt-RICs [24] is ingested by the local downlink bitrate prediction AIF using the if3 interface. These AIFs operate in federated learning fashion and consolidate their model with a centralized AIF using the if2 interfaces. The global downlink bitrate prediction AIF is in charge of building the final model and of redistributing the updated model weights to the local downlink bitrate prediction AIFs. Finally, the global bitrate prediction AIF can feed the estimated expected bitrate to the load balancing AIF which can take the appropriate load-balancing decisions and then reconfigure the distribution of client terminals by interacting with the O-RAN nRT-RIC.

The reference AIF model is to be considered as preliminary. Future iterations will refine the definition of the interfaces. Special attention will be put in describing AIFs not only from the functional point of view but also considering other capabilities (e.g., computation, communication, storage, hardware acceleration) the complement those of the AIFs, as well as any other aspects related to the constraints necessary to enable their dynamic orchestration, i.e., the service requirements.

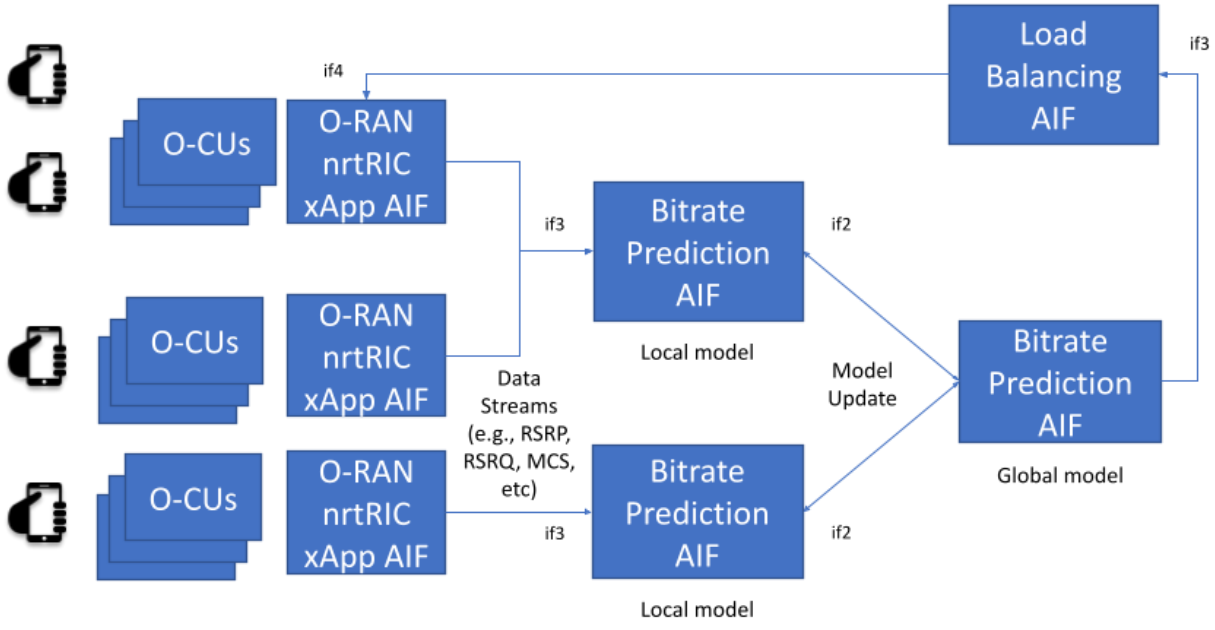


Figure 14 Sample load-balancing applications

As future work to be conducted, we foreseen the investigation of topics such as composition and configuration of AIFs and in particular how a complex AI-enabled application can be composed starting from elementary AIFs including the decomposition of large monolithic AI-enabled application leveraging for example complex neural networks into smaller and components that can allow us to achieve the project-wide goal of reusable and trustworthy AI/ML pipelines.

Furthermore, we will also focus on security and privacy aspects of the AIFs. Thus, we will look into security and privacy aspects related with federated and distributed learning techniques with specific attention to the distribution of models and the sharing of models' parameters in such a way to preserve the confidentiality of the data used during the local training phase. Adversarial ML attacks will be used to inject malicious inputs crafted to fool the AIFs. Then, to make AIFs more resilient to such attacks, vulnerabilities in the model will be identified and countermeasures will be devised (e.g. outlier detection filtering).

An important activity to be undertaken in the frame of WP3, WP4 and WP5 activities is to determine how the four identified AIFs for the four use case (Traffic Controller AIF, Security AIF, Monitoring AIF, and Content Curation AIF) would map to specific AIF architecture challenges (WP3, Task 3.1) and algorithmic challenges (WP3, Taks 3.3), computing system and networking challenges (WP4, Task 4.2) and use case testbed experimentations (WP5).

The view of AI@EDGE is that use case edge computing applications can leverage on the Connect-Compute fabric to run in-network artificial intelligence algorithms instrumental for the use case applications and to adapt the fabric to changing states and data variations. The AI logic is therefore meant to be distributed in the network to run distributed AI algorithms. Among them, federated learning is emerging as an efficient algorithmic framework for distributed learning.

Closed Loop Network Intelligence

In this section we will look into the role of closed loop control in the context of the AI@EDGE Network and Service Automation Platform. A closed-loop control for network intelligence starts from the monitor collecting information from network sensors and terminates at performing the actions by means of, e.g., deploying AIFs, SDN/NFV actuators, and etc. Once an intelligent network task is required or a network problem/anomaly is detected or predicted through analyzing the network and/or service data, an intelligence control loop can be triggered.

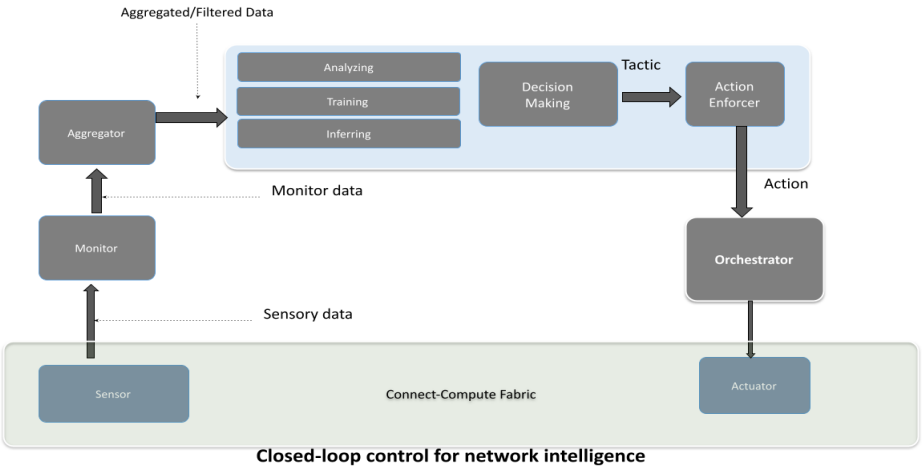


Figure 15 Network and service automation control loop

The involved functional blocks through the whole intelligence control loop as well as its information flow are illustrated briefly in Figure 15. The figure shows the draft version of the AI@EDGE closed loop for network and service automation. This loop will be further refined by WP3 during the rest of the project. Its first official version will be reported in D3.1. Each functional block receives the required information from its previous block, derives a higher level of information and passes it to the next functional block. The information processing of each functional module will be later detailed in deliverable D3.1, where we will also present the current state of the art in the topic at standardization bodies and open-source initiatives.

Federated and Distributed Learning

We elaborate in the following a reference scenario for an AI algorithm based on federated learning, evidencing its possible AIF topology characteristics as well as the possible orchestration points.

Let us first provide a quick view on what federated learning is and how it differs from general machine learning. In conventional distributed ML approaches, the data used for training must be collected from different sources/devices into a server cluster before training takes place. However, these approaches suffer from several issues:

1. Data privacy: owners of data are increasingly privacy sensitive and data privacy legislations are introduced to limit data collection, storage and sharing.
2. Unacceptable latency due to long propagation delays in standard AI/ML approaches.
3. Connection unreliability and high cost of using bandwidth in moving data.

As edge devices acquire powerful sensing and computing capabilities combined with growing privacy concerns, the concept of Federated Learning (FL) has been proposed. Federated Learning is a distributed ML paradigm where several devices are collectively participating in training global ML models locally under the orchestration of a central server. Each device's data is stored locally and not exchanged or

transferred, instead only model parameters are updated for aggregation. FL training process can be used for different ML models such as neural networks, Support Vector Machines (SVMs) and linear regression.

Generally, there are two main components in a FL system: the data owners (i.e., end devices or participants) and the model owner (i.e., the FL server), as represented in Figure 16.

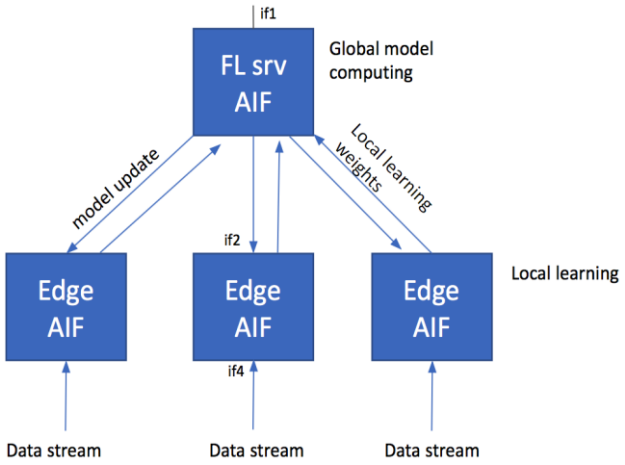


Figure 16 AIF graph representation of standard federated learning communications

Training in FL is done through the following three steps:

1. The FL server AIF selects a subset of the available edge node AIFs and specifies them the ML model (e.g., neural network weights, linear regression weights) to use. In the very beginning it is a starting ML model along with the hyper-parameters to use for building locally at the edge node the ML model (e.g., learning rate); then, it is an ML model update that is sent to the selected participants.
2. Each edge AIF participant uses its local data to update the local model parameters. The local data samples are grouped in batches of a given size B: every B samples, the local model is updated. This is repeated for a number of E of learning epochs: every E epochs, the updated local model parameters are sent to the FL server AIF.
   Some edge FL nodes may not be able at a given round to do the local learning or to send the local model to the server; these nodes are called "stragglers".
3. The FL server AIF aggregates the updated local models from the edge AIFs and then sends back to them the updated global model parameters back to the participants.

Eventually, the learning outcome allows performing classification and prediction tasks, depending on the specific ML algorithm. For instance, if the ML model is a deep neural network used for anomaly detection, the classification precision depends on the configured neural networks weights; if a linear regression is used for prediction, the prediction error can also be affected by the weights update. The principle in FL is that having a global view instead of a simple local view can grant better classification and prediction outcomes indeed. There is therefore room for customizing its FL framework according to the application requirements, in terms of the expected performance as well as the timing requirements. Depending on the frequency of data stream samples, the batch size (B) and the number of epochs (E) are to be tuned to meet such requirements. The AIF graph traffic model also changes with respect to these settings, as the links between edge node AIFs and FL server AIF have a bitrate that is inversely proportional with both B and E.

The conventional basic FL algorithm is Federated Averaging (FedAvg) that computes the new model weights by averaging edge node weights [1].
Since data on participant's devices are produced locally, this data tends to be unbalanced and highly non-identically distributed (non-IID). Imbalanced data leads to a deterioration in model accuracy. Although it is argued that FedAvg handles both IID and non-IID, it does that without convergence guarantees on non-IID data. Another sources of heterogeneity in collected data is that the storage, computational, and communication capabilities of each device in federated networks may differ due to variability in hardware (CPU, memory), network connectivity (3G, 4G, 5G, Wi-Fi); this could dramatically exacerbate challenges such as straggler mitigation and fault tolerance.
In order to mitigate the effect of system and statistical heterogeneities new FL algorithms are proposed that make modification to the FedAvg objective function. FedProx [2], tackles systems heterogeneity by incorporating partial updates that usually conform to the underlying systems constraints; it makes a small modification to the objective function by adding a tunable proximal term to it. Experiments in [2] show an overall 22% improvement in accuracy than FedAvg.
Another algorithm based on FedAvg objective function is q-FedAvg [3]. Inspired by α-fairness for fair resources allocation in wireless communication, it reweighs the objective function in FedAvg to assign higher weights in the loss function to devices with higher loss to encourage less variance in the final accuracy distribution.

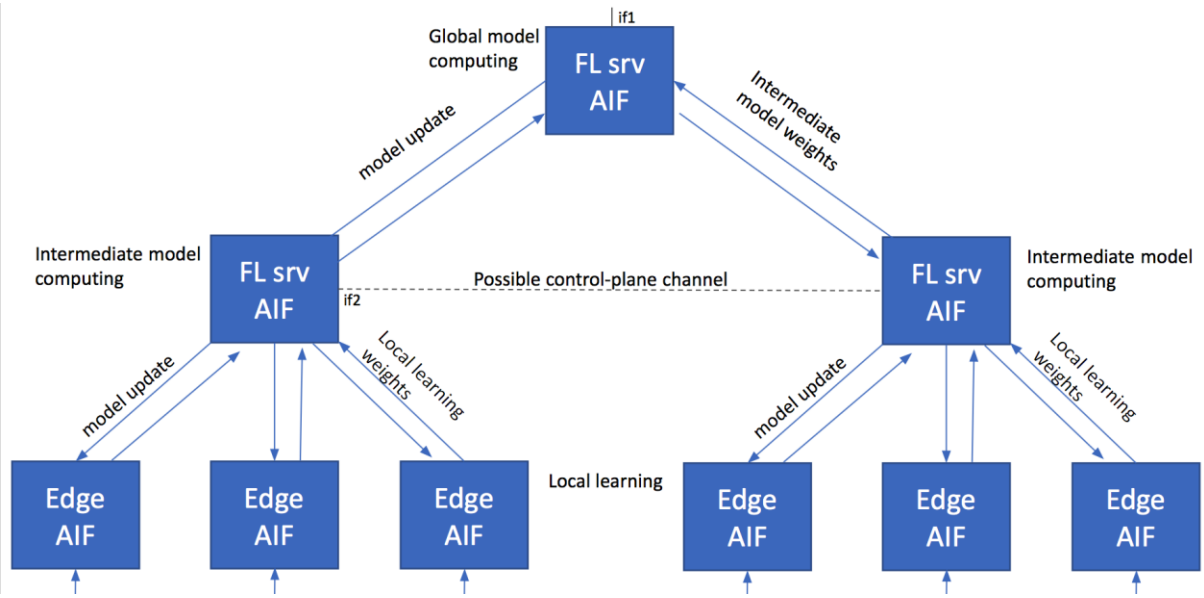


Figure 17 AIF graph representation of hierarchical federated learning

Other FL framework variations exist and will be considered in the AI@EDGE project; namely, solutions exist to deal with communication bottlenecks [4-6], and to protect against privacy violations [7-9]. In the frame of mobile access edge networks, the HierFAVG proposal is made in [10]; it allows multiple edge servers to perform partial model aggregation in the overall training process. In this approach an edge intermediate server sits between edge nodes and the central server and aggregates models collected from edge node updates. After a predefined number of edges, server aggregations the edge server then communicates with the cloud for global model aggregation. The HierFAVG setting is resumed in Figure 17 under the AIF forwarding graph perspective. HierFAVG has two benefits over FedAvg: it reduces communication costs, and relieves the burden on the remote cloud if it is distant from the edge nodes, or highly loaded. HierFAVG represents an interesting framework for the AI@EDGE project, and it may reveal useful for some use cases.

AI/ML problems

The focus of WP3 will be to investigate the classification and prediction problems arising in the AI@EDGE fabric and how AI algorithms and ML frameworks such as FL can help in meeting the stringent use case requirements. Among the currently identified AI problems we can mention:

- *traffic prediction for network QoS improvement:* in some use cases, predicting the traffic volume or the traffic pattern permits us to take infrastructure reconfiguration decisions in advance, in a proactive fashion, hence going beyond the reactive approaches. For instance, in UC4 the challenge of using multiple access radios for downlink communications with airplane multimedia content users calls for a proactive downlink traffic scheduler able to anticipate congestions in the available interfaces and radio channels.

- *infrastructure state clustering for anomaly detection*: the large heterogeneity of components in a softwarized infrastructure making use of SD-RAN, NFV, SDN and cloud-native technologies is a challenge that can hardly be solved by modeling the network with standard network planning and traffic engineering models. Softwarized network components today come with a plethora of features derived from computing resource usage (CPU, RAM, storage), network interfaces, at both physical and virtual/container level. Hence, an AI/ML approach to qualify the running network state to detect anomalies appears at the forefront of the networking research agenda. Monitoring the complete stack of a virtualized infrastructure calls, on the one hand, for a large amount of data to be collected and, on the other hand, to possibly allow for distributed anomaly detection to meet the stringent requirements in terms of infrastructure reconfiguration. Adopting a federated learning approach for softwarized infrastructure anomaly detection is a direction currently investigated in WP3. Particularly relevant for UC2, anomaly detection can then trigger reorchestration of a softwarized edge network infrastructure to go back to a nominal working state.

- *classification problems for pattern recognition and context analysis*: UC1 and UC3 AIFs are meant to solve a set of classification problems for context analysis and recognition of patterns. For instance, in UC1, each connected vehicle, both autonomously and humanly driven, generates information that is sent to the network for various reasons (e.g., maintenance, traffic management, dangerous situations, infomobility). Depending on the geographical area, there are different situations, from dense areas of vehicles with potentially dangerous situations for pedestrians and drivers to areas with few vehicles not particularly dangerous. Being able to identify and classify these vehicular traffic situations through the data generated by the vehicles becomes essential to be able to dynamically allocate and deallocate network resources capable of satisfying traffic density KPIs and the latencies necessary for high-risk services. In UC3, the identification and georeferencing of the elements of the infrastructure being monitored, done through AI classification, will enable to recognize the infrastructure incidents. In this way the monitoring task becomes more automated and visual, facilitating the work of the drone operator. Application example: the drone is flying and identifies a concentration of vehicles indicating that a traffic jam is occurring. The drone follows the traffic jam until it finds its origin (a fallen branch, an accident, etc.). This information maintains the drone operator constantly updated, helping him to make the best decisions.

A complete set of AI/ML problems addressed by the project, their formal definition and preliminary modeling and resolution approaches will be documented in D3.1. A particular focus will be given to security and privacy of the AI/ML approach and corresponding data collection and processing steps, in view of adversarial machine learning attacks making surface for both centralized schemes and FL schemes.

8.3 The connect-compute platform

In this section we give an account of the minimum set of functionalities that the connect-compute platform introduces into the overall AI@EDGE architecture.

The AI@EDGE connect-compute fabric combines Function-as-a-Service (FaaS)/serverless computing, hardware acceleration (GPU, FPGA, and CPU), and a cross layer, multi-connectivity-enabled disaggregated RAN into a single connect-compute platform allowing developers to take advantage of the new capabilities offered by 5G using well established cloud-native paradigms. The resulting system will allow workloads to be intelligently spread and scaled across the connect-compute fabric according to their requirements. The connect-compute platform leverages heterogeneous hardware acceleration solutions based on GPU and FPGA, to optimize energy consumption, performance, and security for specific AI-based workloads types.

The connect-compute platform components and interfaces will be discussed in greater detail in D4.1 and D4.2 ("Results on the validation of the AI@EDGE connect-compute platform"), while in the following sections, an overview of the main platform architecture components is given.

The connect-compute platform main elements and functionalities are described in the following sections.

The connect-compute platform basic functionalities

This paragraph covers basic functionalities of the connect-compute platform, and specifically the *Application Deployment and Migration* functionalities and *Slicing* functionalities.

Application Deployment and Migration: The Connect-Compute platform will provide the ability to deploy applications and AIFs in an evolutionary path towards cloud-native application deployment. To this end, the platform shall support current lightweight virtualization infrastructure such as containers and integrate alongside FaaS solutions to provide a homogeneous way to expose and use virtualized resources.

AI@EDGE will extend the current ETSI MEC/NFV architectures [27] to embed in the Connect-Compute platform the heterogeneity given by the different application building domains and the various edge layers. The edge platform will be distributed across various layers (for instance, a near edge located at the aggregation points and a far edge deployed at the central office) comprising different capabilities and resources. While the platform located at the nearest edge will count with more limited computational resources and minimal hardware acceleration capabilities (if any), to name a few, the platform at the far edge will comprise a greater computational capacity and more advanced hardware acceleration units (GPUs and FPGAs).

Given the heterogeneous nature described above, the current orchestration frameworks available in ETSI MEC/NFV architectures are insufficient for the AI@EDGE platform. On the one hand, currently available and widely-used Network Function Virtualization Orchestrator (NFVO) and MEC Orchestrator (MEAO) solutions cover a single type of virtualization infrastructure (in certain cases, limited support for simultaneously managing virtual machines and containers could be offered) and are often not suitable for multi-site or distributed environments.

On the other hand, in view of the diversification of resources and hardware acceleration options, the applications to be deployed in the platform could be provided in different versions, for instance in the cases where one of those applications relies on a certain hardware acceleration solution such as FPGAs, which require a dedicated implementation. As such, the complexity of the orchestration of these applications may highly increase. In addition to these functionalities, the consortium is evaluating the requirements of the orchestration components in order to consider the level of automation expected in the AI@EDGE. Such automation operations may comprise not only the deployment and scaling of applications at different sites according to the requirements and resources available, but also their migration across different sites (e.g., from a near edge to a far edge) driven by diverse events such as user mobility and resource outage. The type of the migration performed (stateless or stateful), as well as the orchestration components and interfaces involved in the process will be discussed in greater detail in D4.1 and D4.2.

Due to the above reasons, at the initial stage, the applications to be instantiated in the AI@EDGE platform will be pre-deployed at the required sites using the dedicated version needed for that site (i.e., depending on if hardware requirements apply on that site). To that end, the resources will be reserved in the virtualized environment without employing any kind of orchestration features. A deeper description on the deployment options of the applications will be given in D4.1.

Network Slicing: A slice is a logical partition of the overall network infrastructure that provides full network functionalities and spans through all network segments: radio access, core, transport, cloud, and support systems. The connect-compute platform will provide means to support end-to-end slicing, reserving resources for the new service instances that require it. Currently, providing hard slicing with total resource isolation is very costly, reaching the point of having to replicate a full platform for each new instance in order to achieve complete isolation. To avoid that and still be able to offer the required slicing levels, we have to balance the options in each stage, from the storage hosted in the core data centers or the MEC infrastructures, to the radio, maximizing the resource isolation without compromising the connect-compute platform functionalities.

The connect-compute platform aims to take advantage of the serverless paradigm at the MEC platform, but the abstraction level reached, collides head-on with the slicing concept: Serverless platforms, offer FaaS, facilitating quick response to the services and the UEs, but at the cost of losing some control over the infrastructure where they are deployed, making difficult the resource isolation and therefore, the slicing level provided. Slicing of Serverless Platforms is not commonly provided out of the box, therefore solutions that effectively combine resource effective consumption and isolation will be considered. Furthermore, solutions that rely on the duplication of Serverless Platform services may involve non-efficient resource utilization. During the project, partners will investigate the best balance between solutions, and the availability of tools providing isolation will be one of the selection criteria for the Serverless Platform to be integrated.

The core network (CN) plays an important role in the practical realization of network slicing, enabling resource partitioning, isolation, and the satisfaction of specific quality-of-service (QoS) requirements. Dedicated network segments are realized in 4G networks by differentiating Access Point Names (APNs), mapping them to particular VLANs, being the CN the main element responsible for separating and directing the traffic accordingly. This works, for example, for private network scenarios, such as a manufacturing enterprise that wishes to dedicate a segment of its network exclusively to control its robots' operations,

another for smartphones, and a third one for video surveillance. Assuming that all the involved end-devices support multiple APN configuration, such a solution allows segregation of traffic, adapted security, different assignments of QoS levels, and end-to-end prioritization.

Nonetheless, fully flexible slice deployment and dynamic resource allocation require more advanced and innovative solutions, both for private and national or large-scale networks. Looking at the evolution of slicing in 5G from the CN perspective, the 3GPP standard associates a 5G network slice with a so-called "Single – Network Slice Selection Assistance Information" (S-NSSAI) and a Packet Data Unit Session to a Data Network Name (DNN), equivalent to the APN. The capability to handle S-NSSAIs theoretically allows a 5G CN to support any number of slice instances of a given slice type. Moreover, at least three intertwined technological and architectural enablers play a key role: the virtualization of network functions (begun with 4G), their assembly into a service-based architecture (SBA), and their distributed deployment that can (and more and more will) involve edge sites. In particular, implementing Multi-access Edge Computing (MEC) will tangibly complement and bring value to the public carrier network slicing model defined in 3GPP standards.

As in the connect-compute platform that AI@EDGE is developing, a 5G system can support applications running at MEC servers by deploying a User Plane Function (UPF) directly at the edge, not necessarily collocated with the rest of the CN functions. This allows to maintain localized part of the (or the whole) data traffic and contributes to complying with specific Service-Level Agreements (SLAs), like those typically related to Ultra-Reliable and Low-Latency Communications (URLLC). Even more complex scenarios can be handled by a 5G CN that deploys intermediate UPFs (I-UPFs) into chains that are more naturally suitable for data traffic differentiation over distinct slices, adapting to a variety of use cases. Analogous multiple or redundant deployments can be conceived also for other CN functions, towards a higher communication and service resilience or to enable the coexistence of slices with potentially contradictory performance requirements. In this sense, virtualization is crucial for an easy instantiation of the CN, for the optimization of the resource exploitation (especially when they are shared), and for the scaling of each network slice. In addition, compared to older generations, 5G CNs are by design more "malleable" and suitable for supporting each slice's peculiarity. They can be tailored so that dedicated CN functions are deployed within individual slices to enable specific services in a customized fashion (like a multimedia broadcast services function in a slice modelled for AR/VR live content broadcasting), allowing a dynamic reshaping of the network according to its evolving needs and the environment changes. Finally, but not less importantly, the virtualized nature of 5G CN functions makes them particularly adapted to an automated instantiation and management, coherently with the modern frameworks for network slice management and orchestration based on or supported by AI and ML mechanisms.

For effective end-to-end slicing, appropriate support must also exist in the RAN. The slice manager and/or the orchestrator must be able to control the relevant parts of the network to correctly allocate slice resources as needed. In particular, in AI@EDGE, srsRAN, and more importantly srsENB (both provided by SRS), supports this via its highly customizable config files and scheduling options, which allows a slice manager to create on start-up a slice with the necessary resources. srsENB currently offers two schedulers: a proportional fair scheduler and a Round-Robin scheduler. The interface to the eNB is designed such that it can be easily customized to meet the requirements of the network. Similar customizations could be done to allow re-allocation of resources on the fly with a high level of control. QoS is also supported by srsRAN,

which would allow finer control over slices by assigning varying levels of priority to the traffic in each. Nonetheless, further control of network resources for slicing, beyond scheduling and the configuration files, is possible. This could be done by exposing metrics and features within the code at various points, or the relevant interfaces. As this is not currently supported out-of-the-box, modification to the code-base would be required. To do this, the desired metrics/ interfaces/ network functions would first need to be identified.

Potential issues with network slicing from a RAN perspective are found in the dynamic allocation of bandwidth. It is possible to allocate bandwidth on start-up via the config files, which will allow the RF-frontend to correctly reserve the desired amount. For 4G networks it is not possible to then reallocate bandwidth during run-time as this would require a hard-restart of the radio. This is not an issue stemming from srsRAN, but is a hardware issue that cannot be overcome. This is overcome in 5G networks by configuring a wideband carrier and then dynamically allocating bandwidth within this as needed. This then removes the need to restart the radio each time bandwidth is reallocated. This feature is not yet available within srsRAN, but is expected to be released over the course of the project.
In view of the above, the consortium is currently examining the requirements of both the radio access and the core network to support the types of slicing described before. Regardless of the two main approaches depicted in Figure 18 and Figure 19, it is critical to bear in mind that their isolation does not depend strictly on the capabilities of RAN and core, but also on the design performed at the level of the MEC infrastructure.

On the one hand, Figure 18 shows two alternatives of the network slicing design based on DNNs. The first alternative (in green, depicted on the top of the figure) presents a system where all the end-to-end slices share a unique MEC platform, each of such slices being deployed on its own DNN and containing its own UPF and MEC applications. The second alternative (in pink, on the bottom of Figure 18) is similar to the first one with the difference that in this case, each slice would also have a dedicated instance of the MEC platform. While this option may provide greater isolation across slices, it would add a greater degree of complexity on the orchestrator, which would be forced to handle a higher number of MEC platforms, increasing computational and network management complexity. Note that the approach in Figure 18 considers one SMF shared by multiple slices, which can be envisioned only if a data-path between the SMF and the UPFs is made available. Alternatively, multiple SMFs (one or many per slice) could be used, as depicted in Figure 19. This first approach would be also applicable to 4G core networks in 4G and 5G NSA scenarios, replacing the AMF by the MME, the SMF by the SGW-c plus PGW-c, and the UPF by the SGW-u plus PGW-u. In this case, since slicing is not natively supported by 4G networks, this approach would be based on traffic segregation. For simplicity, other core NFs are omitted in the figure.

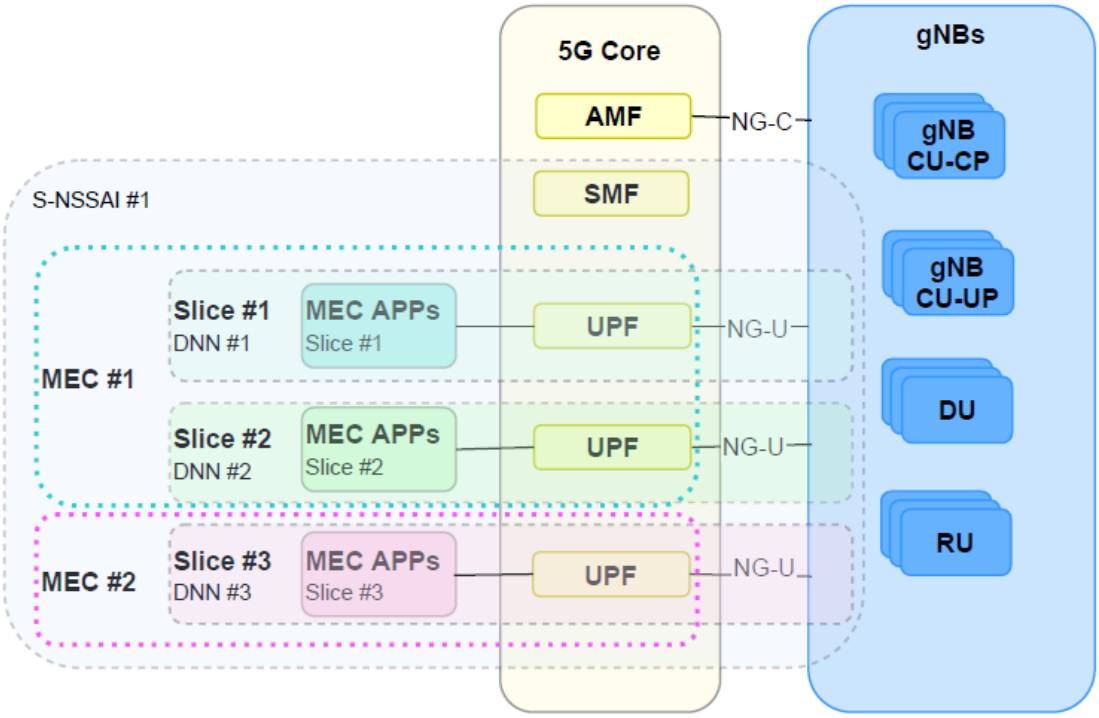


Figure 18 Alternatives of network slicing based on DNNs

On the other hand, Figure 19 provides a similar view comprising two alternatives for the network slicing design based on a combination of S-NSSAIs and DNNs. Similarly, to Figure 18, the two alternatives represent the cases where the same MEC platform is shared by several slices using different DNNs within the same S-NSSAI (i.e. shared SMF, slice-specific UPFs) (represented in green in Figure 19), and the case where the MEC platform is dedicated to a single slice instance with slice-specific SMF and UPF (represented in pink in Figure 19). At the radio side, this network slicing architecture could be complemented by the instantiation of different CUs serving different slices according to the S-NSSAI parameter.

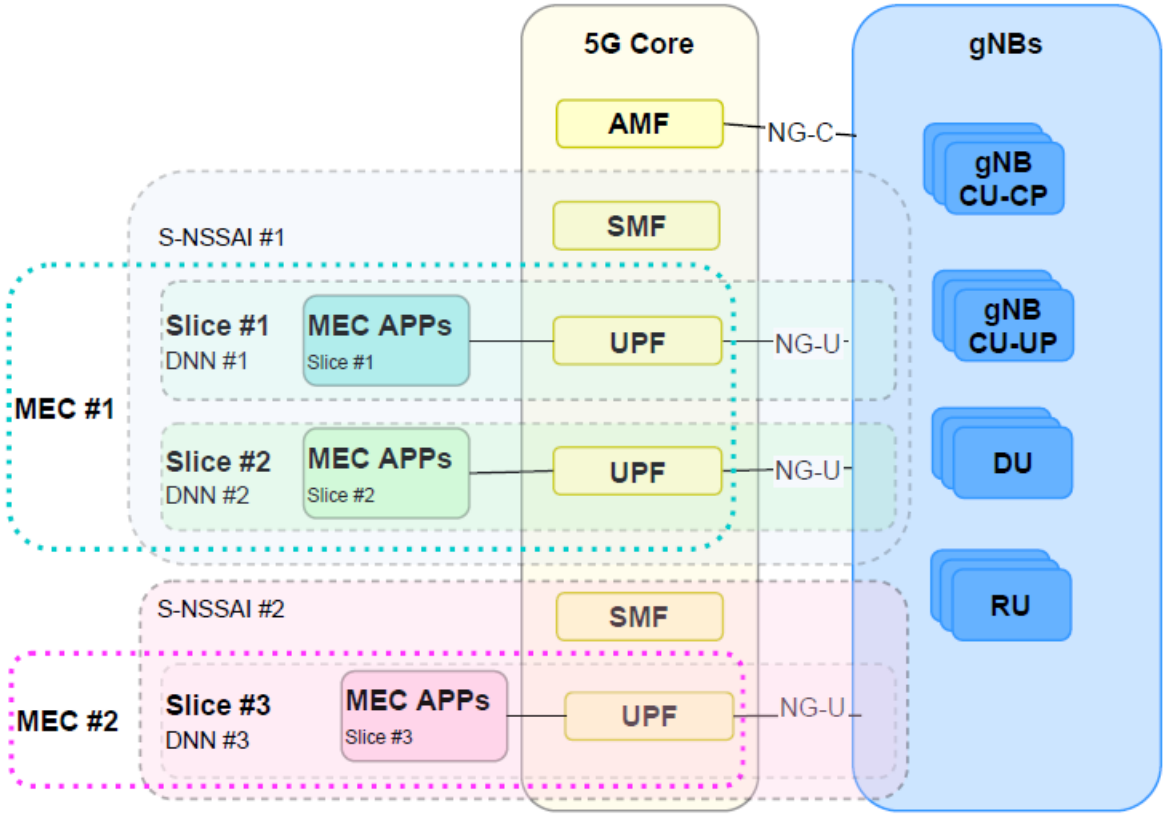


Figure 19 Alternatives for network slicing based on S-NSSAIs

The consortium is currently evaluating the aforementioned described options, the requirements imposed by the connect-compute platform and the capabilities of the RAN, 5GC and MEC components. A more detailed view of the status and a specific decision made on the network slicing design will be provided in D4.1.

Distributed and decentralized serverless connect-compute platform

The connect-compute platform provides serverless support to all use cases, and in particular to UC2, which expects to leverage the serverless capabilities of the AI@EDGE fabric to scale with changing computing demands.

Serverless platform support: The connect-compute platform will integrate with a serverless platform in order to provide FaaS functionality. The choice of the Serverless Platform to be integrated will depend on various factors, including the availability of Open Source licencing, performance and flexibility of integration in the current LightEdge framework [52]. With FaaS, developers will be able to develop functions ("Serverless Functions"), in various programming languages (depending on the chosen platform) that can be executed in response to events without dealing with the complex infrastructure typically associated with building and launching microservices applications. In this model, the infrastructure provider takes the responsibility of managing the underlying infrastructure and dynamically allocating

enough resources to auto scale the applications and services based on the demand. Serverless computing offers a variety of benefits over traditional computing, including the ability to scale to zero without charging the customers for idle time, zero server management, and autoscaling, making it an appropriate technology for a number of use cases such as stream data processing, chatbots, stateless HTTP applications, etc.

Serverless Functions: Serverless Functions are code units that have the property of being event-driven, stateless, and often short-lived. Serverless Functions can be implemented in any programming language between those supported by the Serverless Platform. The Serverless Function is triggered by an event, such as a REST API call or a MQTT message. The output format of the Serverless Function may vary, can be another event, or an entry on a database, or the response to a REST API. The serverless platform implements event listeners that catch the event and calls the Serverless Function associated with the event passing the parameters, if any. Developers have to just write the code for the function, all the burden of preparing and managing it is taken care of by the Serverless Platform, from setting up the triggers, to dundling, deployment and autoscaling.

Serverless Function deployment: Serverless Functions are deployed when an event happens. The platform commits just the required amount of resources to a particular application/task (as many instances as necessary, but only when needed), and utilizes the resources for just the time needed to execute an invoked function. When there is no demand, the platform can scale near to zero (i.e. no resources used), while it scales to as many instances (with some limits) as needed to meet the traffic demand. Depending on the function, starting from zero can require a sensible delay. This is known as the cold start delay problem. Strategies may be put in place to avoid high cold start delays, when necessary. Such strategies include caching and prediction. A common strategy is that the applications consuming the Serverless Functions preload them before using, e.g. at the application start, in order to have them already available when needed.

Migration of Serverless Function across Architecture: Serverless Functions are meant to be stateless, and short-lived. For this reason, migrating a Serverless Function while it is running may not be a good strategy. However, when migrating an application using Serverless Functions the cold start delays associated should be considered. Adopting strategies to assure a smooth migration between Edge Nodes should consider avoiding cold start delays on the target note, e.g. pre-loading necessary functions as soon as possible.

AI-enabled application provisioning and orchestration

The AI@EDGE platform will include data-driven service lifecycle management solutions for the deployment, management and monitoring of end-to-end AI-enabled applications. To this aim, an end-to-end decentralized and distributed orchestration solution of AIFs will be researched, thus supporting the development and deployment of the AI-enabled applications.

Provisioning of AI-enabled applications will start from “de facto” standards for the orchestration of cloud and edge services, such as Docker and Kubernetes and from their emergent variants (e.g., FaaS) to cover more and more extreme provisioning scenarios in terms of hardware and software resources.

We expect that some of the AI-enabled applications will require that orchestrating mechanisms consider additional factors, apart from the code and data locality, when making appropriate placement decisions at MEC servers, for example to satisfy stringent latency and data rate requirements. Since these applications

will be composed of multiple AIFs, AI@EDGE will research innovative solutions for the end-to-end orchestration able to satisfy the diverse requirements of each AIF and to ensure that the QoS requirements of the applications are satisfied while the network resources are used in the most efficient manner. To this aim, the solution will leverage the AI@EDGE network and service automation platform developed in WP3, in which monitoring solutions tackling the specific requirements of optimized hardware, edge devices, communication infrastructures, and cloud services will be designed and developed.

Current service orchestration solutions primarily focus on centralized machine learning algorithms. In AI@EDGE, the orchestration framework of the AI-enabled applications (AI functions) will consider also the orchestration of AI-enabled applications and AIFs based on distributed learning approaches, such as:

- Federated AIFs: where AI functions store locally their own data and locally run the ML algorithms. Each AI function has specific characteristics of storage, speed, accuracy of ML performance, and/or solve a specific portion of the task.
- Cooperative AIFs: where AI functions communicate and share insights as parameters (not raw data) to solve complex cooperative tasks.
- Composite AIFs: where different AIF instances perform a subset of the AIF operations as (partially) concatenated and chained AIFs.

The provisioning of AI- enabled applications is envisaged by all the UCs. At this stage, the AIFs envisioned by the UCs, and that will provisioned by the system, are resumed as follows:

- In UC1, the Traffic Controller AIF is a distributed AI Function that collects and shares information to create a digital view of the roundabout and its surrounding, and the orchestration framework should take into account the specific requirements of this function to provision the application in the most effective way.
- In UC2, the AI security function (Security AIF) adopts federated learning to share trained models while preserving the data confidentiality among different stakeholders, in multi-stakeholder IIoT environment.
- In UC3, composite and distributed monitoring Artificial Intelligence Function (Monitoring AIF) will be used to carry out infrastructure monitoring, executed by drone, in real time.
- In UC4, a composite content curation artificial intelligence function (Content Curation AIF) is used to sense and manage the infrastructure. This AIF is expected to be deployed with a (possibly distributed) set of AIF instances, possibly making use of hardware acceleration.

The AI-application provisioning framework will also consider solutions for provisioning AI enabled applications that concatenate and integrate multiple AIFs of different nature (Composite AIFs), to achieve its goals. In Composite AIFs, different AIF instances perform a subset of the AIF operations as (partially) concatenated and chained AIFs.

AIFs Reference Model: For being able to provision the AIFs across the connect-compute platform, an AIFs description must be provided. To this aim, a reference model for AIFs, based on advanced knowledge representation techniques, will be defined. The reference model will be designed to capture and represent the heterogeneity of AIFs at the different levels of the technology stack. It will describe the AIFs from a functionality point of view, similarly as it happens in catalogues. Additionally, it will also contain the information related to AIFs' capabilities and constraints (e.g., computation, communication, storage,

hardware acceleration) necessary to support their dynamic orchestration over the AI@EDGE platform. Similarly, the reference model shall also offer capabilities to describe methods and approaches to support the provisioning and orchestration of the AI-enabled applications developed in WP5.

The reference model will be developed as a network of interconnected modular ontologies, implemented in standard knowledge representation languages (e.g., OWL[2] - Web Ontology Language). Such ontology networks will combine ontologies focused on the description of datasets, software entities, algorithms, models, workflows, evaluation testbeds and frameworks, appliances and hardware. It will be developed according to well-known state-of-the-art methodologies for ontology engineering and will be properly documented and published under open licenses so that they facilitate the generation of metadata for all those artefacts in such a way that the AIF catalogue can be generated and populated. These ontologies will be based on existing ontologies and metadata profiles, as much as possible, so as to ensure compliance with as many existing catalogues as possible.

AIFs Catalogue: As part of the connect-compute platform, a catalogue of AIFs, based on the reference model for AIFs defined in T4.2, will be produced. It will incrementally be filled with the AIFs produced during the project (e.g. in WP5), and by others available sources (e.g., the AI4EU platform [31]). The description of AIFs will not only include their functional description, but also additional capabilities (e.g., computation, communication, storage, hardware acceleration) as well as those constraints necessary to support their dynamic orchestration.

End-to-End quality of AI-enabled applications: The connect-compute platform will provide tools for measuring the end-to-end quality of AI-enabled applications, whenever these challenges need to adopt a cross-cutting approach with respect to the levels in AI@EDGE technology stack. The monitoring subsystem will collect raw information to obtain the Quality of Service (QoS) indicators that provide insights regarding the correct behaviour of multiple AIFs orchestrated and linked to create complex AI-enabled applications. The quality indicators will cover traditional metrics for IT systems (e.g., performance) but also specific items to assess that AIFs operate according to their initial design. At the same time, the collected metrics and the quality indicators will ensure the capacity to oversee the operation of orchestrated AIFs and even to integrate AI-based functionalities to detect abnormal situations or to implement predictive maintenance. The performance monitoring has the purpose of supporting the quality assurance of the AI-based applications (e.g. monitoring and predictive maintenance, reliability, and security) and will be covered in synergy with WP3.

***Continuous configuration of AI-enabled applications*:** Provisioning and deploying AI-enabled applications assumes the configuration of the latter and the AIFs that compose it. Some parts of the configuration might be guided by functional needs expressed by the users. For example, in case of an application reporting some predictive metrics about the status of the system, the frequency of the updates can be configured or even set as continuous or real-time. However, other parameters are specific, in particular the hyper-parameters of AIFs. Configuring the parameters of AI algorithms is known to be highly challenging and very sensitive. Default parameters could be given with AIFs in the catalogue. However,

[2] Please see: https://www.w3.org/OWL/

customization is often required. Configuration is thus an essential step and can rely on various techniques to optimize the efficiency of the applications (brute-force search, heuristics…). Configuration is necessary at the deployment stage but might be refined depending on contextual changes (adding users, new applications deployed, changes in network traffic…). Therefore, configuration must be continuous. To enable fast reactivity, a set of default configurations rather than a unique one could be provided along with a description of the contexts they fit with. *End-to-end quality* evaluation mentioned can be used to automatically trigger a reconfiguration process (identification and setting of new parameters). These different options must be analyzed in synergy with WP4.

Cross-layer, multi-connectivity radio access

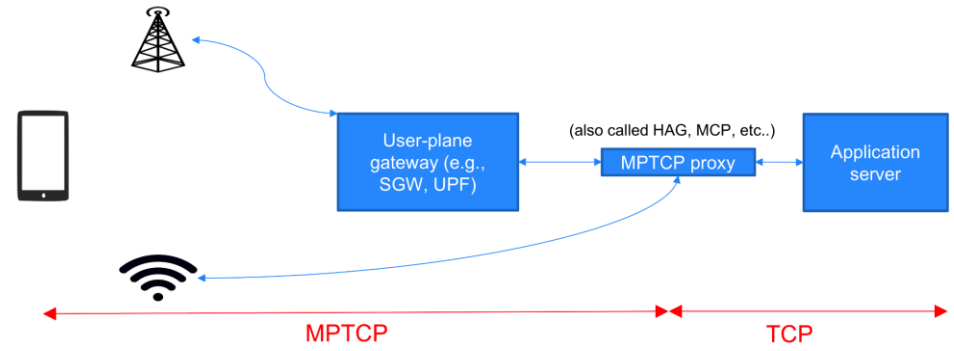
In UC4, the user-level aggregation within a single RAT (by using the Packet Data Convergence Protocol, PDCP) and across multiple RATs and wireline technologies (4G, 5G, Wi-Fi, Ethernet) is envisioned in order to meet throughput, reliability and latency requirements in high spatial density situations. Two main end-user devices are envisioned, wireless tablets or mobile devices and wired screens. Both types of devices are meant to be multihomed; for instance, tablets can dispose of both Wi-Fi and 5G interfaces. On-seat screens can dispose of ethernet and Wi-Fi interfaces. Any other multi-RAT setting can be envisioned, making use also of 4G radios in addition to Wi-Fi and 5G ones, which could be interesting for bring-your-own-device scenarios.

Within the cellular radio, at layer 2 PDCP can be used with two main purposes: carrier aggregation, to increase the overall bandwidth made available to UEs, and increasing reliability by packet duplication, considered for URLLC services. WP4, and in particular Task 4.3, will investigate how tacking PDCP decisions can be a joint or an independent strategy from aggregating different RATs at layer 4.

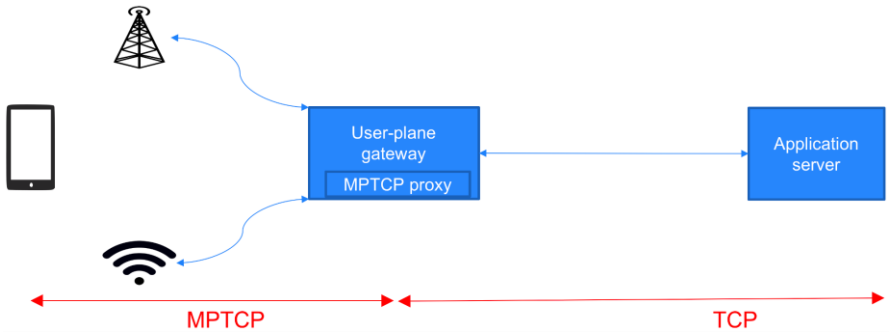
Aggregating multiple RATs can be done at the IP level aggregating the RAT-specific IP paths in the edge network. The idea of using multipath bonding in IP-based access networks is not new and dates back to 2013 [11], later discussed at the IETF [12,13]. Works in this area started at the EIT ICT-Labs, in use cases with Orange and TIM to aggregate wired links (IP, DSL lines) for remote service access [14]. The common denominator of these works is the usage of the Multipath Transmission Control Protocol (MPTCP) extension of TCP [15], and in particular MPTCP proxies in the data-plane paths.

Commercial products currently exist based on MPTCP proxies, commercialized by Tessares, Korea Telecom, OVH, for instance. The first two companies namely propose solutions for mobile cellular access, coupling 4G and 5G radios with Wi-Fi. At the end-point side, MPTCP has been supported by MACOSX, iOS and Linux operating systems for a few years, and it is expected to be soon ported to mainstream Android systems. Note that, as of today, in MACOSX and IOS systems MPTCP is used only for some selected (e.g. Apple) services.

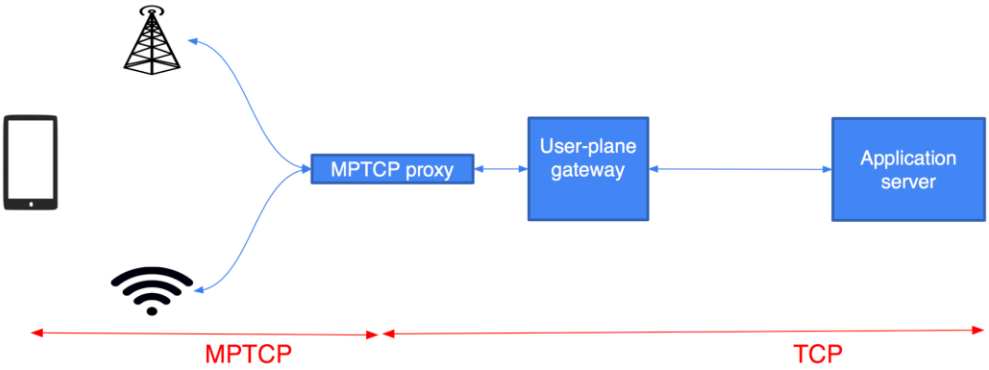
For cellular access scenarios, the UE is meant to be MPTCP capable, the end-point server does not need to be MPTCP capable. The MPTCP proxy therefore receives MPTCP subflows from the multiple RAT-specific IP paths and aggregates them to a single TCP connection on the way to the server. In [12], two network models are described for the integration of MPTCP proxy to aggregate cellular radio with an additional non-3GPPP RAT (typically Wi-Fi). They are depicted in Figure 20.



(a) off-path model



(b) on-path model



(c) on-path model variant

Figure 20 On-path and off-path models for MPTCP proxy usage in multi-connectivity scenarios

- *Off-path model (Figure 20 a):* the MPTCP proxy sits after the user-plane cellular core gateway, i.e., the subflows join after the gateway. The position of the proxy can be anywhere along the IP path from the user-plane gateway and the server, and therefore the multiple RATs can be under the control of different access operators. The address of the proxy may need to be configured in the UE in case of additional paths not crossing the proxy in the uplink direction.

- *On-path model (Figure 20 b):* the MPTCP proxy sits within the user-plane gateway and the subflows join at the gateway. The multiple RATs are therefore meant to be under the control of the same operator. A variant in Figure 20 c consists in having the MPTCP proxy before the user-plane gateway as a network function independent from the UPF (but it would be unfeasible in standard 4G/5G settings due to GTP tunneling, which may however not be a strong requirement in UC4 scenarios as handover may be absent or limited to few known cells).

The integration of MPTCP proxies in 5G systems has been envisioned since Release 16 [16] following the on-path model. It is referred to as ATSSS (Access Traffic Steering, Switching and Splitting) and the MPTCP proxy is integrated within the UPF. A complete 5GC ATSSS system description is expected for Release 17. The current ATSSS specification is depicted in Figure 21; it encompasses the integration of an MPTCP proxy for multi-RAT bonding at the UPF level, with the exploitation of the PCF (Policy Control Function) to regulate the scheduling over the RATs, and the PMF (Performance Measurement Function) to gather real-time packet-level measurements to allow dynamic MPTCP scheduling update.

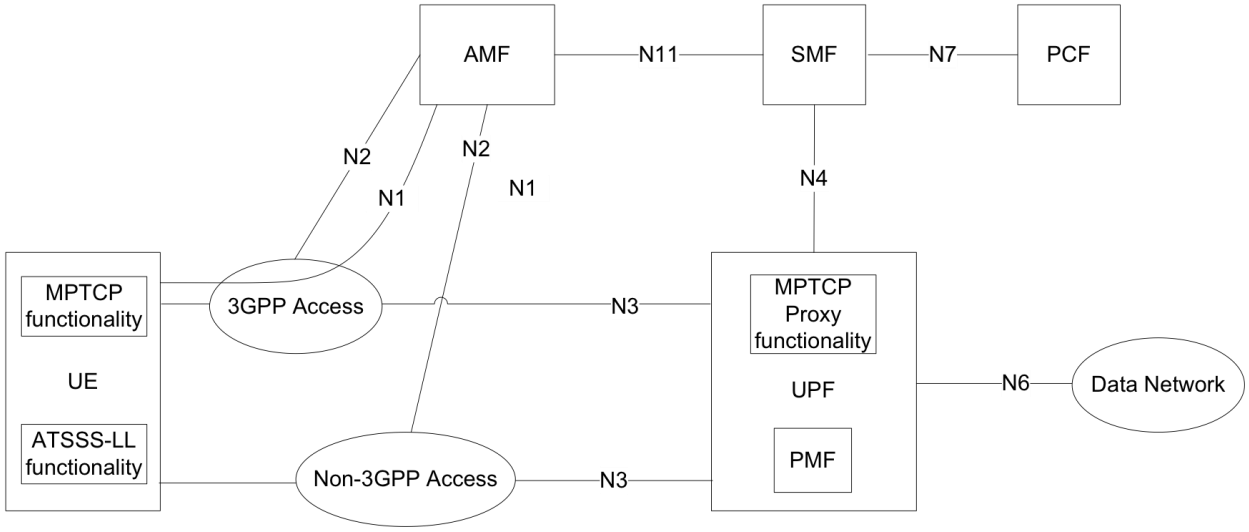


Figure 21 Access Traffic Steering, Switching and Splitting (ATSSS)-capable 5GC system. Source: [16]

In the framework of AI@EDGE, and in particular Task 4.3 and UC4 activities, we plan to design a gradual integration of ATSSS in the AI@EDGE platform and to design predictive scheduling algorithms for downlink communications at the UPF MPTCP-proxy level. Indeed, current MPTCP schedulers are reactive schedulers, changing the decision on which packet to send over which subflow upon sub-flow state changes. Available open-source implementations use packet-level latency and buffer occupancy measures collected in real-time at the socket level, as per the MPTCP standard [15].

The current plan of T4.3 is to contribute to the integration of this multi-connectivity innovation in the AI@EDGE platform, and to experiment novel predictive schedulers for the MPTCP proxy functionalities. We plan first to demonstrate its usage following the off-path model, with a programmable reactive scheduler exposing through a dedicated API its configuration. Then, we plan to move to the on-path model,

using the project near-RT RIC to integrate (part of) the PMF functionalities and the MPTCP proxy as a function sitting before the UPF along the user-plane path. Possible evolutions then include the integration of the PMF and MPTCP proxy functionalities at the UPF, depending on its availability.

Hardware accelerated platform for AI/ML

The connect-compute platform will be integrated with HW accelerators for AIF computing, i.e., FPGA and GPU, in order to provide high-performance AI/ML capabilities to the system. In terms of HW equipment at the "far edge" and/or cloud sites, the HW accelerators will come in the form of ordinary PCIe boards for 1U servers, e.g., Xilinx ALVEO U50 and NVIDIA V100. At the “near edge” sites, embedded devices can be examined (trading lower power consumption for lower acceleration). In terms of SW tools, the AI accelerated functions will be developed with mainstream frameworks and vendor tools from Xilinx [47] and NVIDIA [48]. The programming/development will be realized offline and the resulting executable files will be loaded to our custom library/repository, which will then be deployed in the connect-compute platform (and will be updated continuously).

FPGA acceleration: The FPGA acceleration in the server will involve a high-performance high-density chip communicating through PCIe with the native processor of the server. Such a board is specifically designed for server acceleration with a low profile form factor of half height, half length, 300-400g, single-slot. Different FPGA devices and server CPUs can be combined allowing future adaptation of the connect-compute platform to various deployment requirements. In the context of AI@EDGE, owing to the current availability of tools and frameworks, we will utilize FPGA devices from Xilinx. More specifically, we will utilize an Ultrascale+ [49] device hosted on an ALVEO [50] board offering huge amount of programmable resources, e.g., 1M LUTs (Look-up Tables), 10K DSPs (Digital Signal Processing), 35MB on-chip RAM, 8GB HBM memory, all passive cooling at 75W power consumption.

When used for AI/ML, such an FPGA board can provide acceleration in the area of 10x for common AI cases when compared to the native CPU of the server, e.g., Intel®Xeon®Gold 6138 (figure based on in-house testing). The specifics of each network and dataset affect this speedup considerably, and thus, further study is needed to provide guidelines on when/how to apply FPGA acceleration in the connect-compute platform (per case/application). The FPGA accelerated function will be provided in a bitstream form, i.e., as an executable, which will be loaded to the FPGA device automatically upon request. In combination with a custom SW executable and standard PCIe drivers on the CPU side, the data will be forwarded to the FPGA and the results will be returned to the CPU for further post-processing. This HW/SW co-processing will be wrapped by a SW function/API to replace the original SW-only function and provide the aforementioned acceleration. The development of the bitstream will rely on a series of steps and Xilinx EDA (Electronic Design Automation) tools. More specifically, starting from standard AI models (e.g., TensorFlow) and/or custom C/C++ code, we will use High-Level-Synthesis (HLS) and/or automated Xilinx tools. We will employ the VitisAI [51] tool for standard supported models and Vitis for our more customized HLS. During our design steps, additional care will be given to parallelization and optimization of the digital circuits. The resulting netlists are converted to bitstreams by the Xilinx Vivado[3] tool and, occasionally, careful tuning of tool parameters is needed (e.g., utilization balancing, placement directives).

[3] Details at: https://www.xilinx.com/products/design-tools/vivado.html

In terms of development time, the circuit design can require days and the compilation/synthesis can last hours; therefore, the FPGA development will be performed offline.

UC2 will consider the use of GPU and standard FPGA systems to offload some of the learning computational effort in the remote or far cloud. In UC2, one or more Edge servers and some network nodes will be equipped with P4-NetFPGA smart-NICs that will be used for metrics computation and collection.

GPU acceleration: Similarly to the aforementioned FPGA accelerator, the GPU card in the connect-compute platform will utilize a state-of-the-art high-performance device from the most prominent vendor, namely NVIDIA. More specifically, we will target a HW/SW subsystem customized for server acceleration, e.g., the V100 GPU [39] with CUDA [40] SW and drivers. The device is based on NVIDIA Volta architecture with 5120 cores (640 tensor cores) and 32GB HBM2 memory, while it connects to the server CPU via PCIe Gen3. Its form factor is PCIe Full Height/Length, single-slot, and the power consumption is in the area of 250W. The acceleration for AI/ML workloads is expected to be in the area of 20x compared to the server CPU. The SW development will be performed offline starting from common frameworks, e.g., TensorFlow, and carefully applying optimization techniques and C/C++ coding with CUDA. Fully automated solutions for code generation also exist and will be examined in the context of AI@EDGE.

The V100 device can be used, due to its power consumption requirements (250W), in the "far edge" and "cloud" environment. In the case of a "near edge" environment, the NVIDIA T4 [41] device can be considered as a possible candidate. Indeed, T4 is based on NVIDIA Tensor architecture, providing 2560 CUDA cores (320 tensor cores) but having only 70W power consumption as a requirement. Considering the "near edge" environment with higher power consumption constraints or, in addition, high end user equipment (i.e., drones), it can be considered to move from an architecture that hosts the GPU functionality on the PCIe bus, to an architecture that hosts the GPU function leveraging an SoC (System on Chip) solution. In this case, the SoC integrates the CPU function (typically ARM based) with the GPU function being provided as a HW module to be integrated in the end user equipment. The NVIDIA Jetson family of modules [42] includes devices within a power consumption range between 5W and 30W.

The important aspect to be highlighted is about the SW compatibility that can be obtained when migrating applications among different devices. Indeed, the CUDA Toolkit provides a common development environment for creating GPU-accelerated applications using an abstraction layer. When moving to AI based applications NVIDIA makes available in addition the CUDA-X toolkit [43], built on top of CUDA. CUDA-X embeds a complete deep learning software stack, AI libraries and a Deep Learning framework, including TensorFlow, Pytorch and MXNet.

GPU hardware acceleration is envisioned to be used by UC1, and UC2 in the remote or far cloud, and in UC4 at the edge. Possible usage in UC3 will also be considered.

HW acceleration support to Serverless Functions: AI@EDGE will support heterogeneous acceleration capabilities for Serverless Functions, especially the AIFs, through the utilization of the GPU and FPGA accelerators of the connect-compute platform. This will be achieved in a cooperative manner between containerization and resource orchestration. Given that serverless functions, at the very end, are running inside containers, the AI@EDGE platform will provide a repository with several "flavors" of container compositions according to the target node, i.e. containers will be extended with specific runtimes and the

corresponding drivers from NVIDIA and Xilinx, thus enabling the direct and transparent access to the specific HW acceleration resources, GPUs or FPGAs, from containerized functions mapped onto nodes with acceleration capabilities. However, careful containerization solves only half of the problem, i.e. how to access the GPU or the FPGA when the mapping (the assignment of the container to the node) has been performed. To fully support heterogeneous HW acceleration for Serverless Functions, the AI@EDGE resource orchestrator will be extended to with proper device plugins [17][18], enabling the discovery of the specific accelerators across the nodes forming the orchestrated resources, their representation as schedulable resources, their monitoring as well as their secure sharing among containers. During a FaaS deployment, each function requiring access to a specific HW acceleration resource is expected to be invoked in a proper container populated with the HW specific runtimes, while the orchestrator will control the rest of the resource allocation and container placement according to the current system state and the requested resource affinity. Several details concerning more advanced features like function co-location on HW accelerators, function multi-versioning across differing accelerators etc., will be evaluated in conjunction with the AI@EDGE use cases and be discussed in WP4 deliverables.

8.4 Relationship with 5G PPP activities

The AI@EDGE consortium counts multiple partners with strong expertise in 5G-PPP projects and a long track record of successful collaborations. Several AI@EDGE partners have been, or are actively involved in 5G-PPP Phase 1, Phase 2, and Phase 3 projects and act as liaisons between AI@EDGE and these projects. This ensures a smooth flow of information among the projects and the successful integration of relevant outputs from other projects into the AI@EDGE architecture. This knowledge transfer from ongoing or recently finished 5G-PPP projects is key to ensuring that the AI@EDGE platform takes into account the most recent 5G technological developments and a safeguard of the smooth transition of communications to the beyond 5G era.

Figure 22 The projects that are a part of Phase-1 5G-PPP, as listed in the 5G-PPP website

Figure 23 The projects that are a part of Phase-2 5G-PPP, as listed in the 5G-PPP website

Figure 24 The projects that are a part of Phase-3 5G-PPP, by category

In order to better understand the relationship between Phase 1-3 projects with the AI@EDGE, the scope of projects whose objectives are most closely related to those of AI@EDGE, are presented in this section:

5G-CLARITY (*Beyond 5G multi-tenant private networks integrating Cellular, Wi-Fi, and LiFi, Powered by ARtificial Intelligence and Intent Based PolicY*) aims at converging multi-technology access networks and addressing the challenges in spectrum flexibility, delivery of critical services, and autonomic network management. 5G-CLARITY develops and demonstrates a beyond 5G system for private networks integrating 5G, Wi-Fi, and LiFi technologies, and managed through AI based autonomic networking. 5G-CLARITY aims to be instrumental in order to secure the leadership of Europe in the growing markets of private 5G networks, and 5G for factory automation.

5G-CLARITY brings forward the design of a system for beyond 5G private networks that addresses the challenges in spectrum flexibility, delivery of critical services, and autonomic network management. The project is based on two different technological pillars: the first pillar is a heterogeneous wireless access network that integrates three technologies: 5G beyond R16, Wi-Fi, and LiFi. The second pillar is a novel management plane based on the principles of Software Defined Networking (SDN) and Network Function Virtualization (NFV), and powered by Artificial Intelligence (AI) algorithms, in order to enable network slicing for neutral hosts, and autonomic network management.

Much like AI@EDGE, 5G-CLARITY is AI-powered for a significant amount of its technological advances. In particular, 5G-CLARITY supports AI-driven management. 5G-CLARITY enables effective provision of slices, managing and optimizing their performance. By supporting faster fulfilment of user and business intents while enabling optimal resource sharing during the entire lifetime of slices, AI-driven management drives network automation by greatly reducing the need for human intervention. Relevant outputs from this project like multi-connectivity and AI for network resource management will be considered while designing the AI@EDGE platform to enable a resilient zero-touch service management system.

5GZORRO (*Zero-touch security and trust for ubiquitous computing and connectivity in 5G networks*) aims at combining zero-touch automation solutions and distributed ledger technologies to enable a secure, flexible, and multi-stakeholder combination and composition of resources and services in 5G networks. Furthermore, 5GZORRO introduces a security and trust framework that is integrated with 5G service management platforms, to demonstrate Zero Trust principles in distributed multi-stakeholder environments, as well as automated security management in order to ensure trusted and secure execution of offloaded workloads across domains in 5G networks. In addition, 5GZORRO defines a Smart Contract ecosystem anchored on a native distributed ledger to allow commercial and technical data provided by 3rd-party users to be standardised and mapped into Smart Contracts, which can be initiated "at will" between multiple untrusted parties.

AI-based techniques for zerotouch network orchestration are valuable inputs to be considered in AI@EDGE which aims at building a resilient, secure, and elastic management of end-to-end slices.

SESAME (*Small cEllS coordinAtion for Multi-tenancy and Edge services*) is a Phase 1 project whose field of study focuses upon the concept of CESC (Cloud-Enabled Small Cell); a Small Cell (SC) with shared

virtualized resources between several operators. Computing capacity in Light Data Centres (DCs) is one of the main pillars of SESAME. Those Light DCs allow the deployment of Virtualized Network Functions (VNFs) and support "self-x" operations and management to execute applications in a flexible clustered edge inside the network, by using low-power processors and hardware accelerators for low-latency purposes.
SESAME Light DC features low-power processors and hardware accelerators for time critical operations and builds a high manageable clustered edge computing infrastructure. This approach allows new stakeholders to dynamically enter the value chain by acting as 'host-neutral' providers in high traffic areas where densification of multiple networks is not practical.

The optimisation of the CESC management is the "key" innovation and challenge in SESAME, where orchestration, NFV management and management of resource and radio access help in the development of a neutral network that offers access to diverse providers, especially for those cases where networks cannot be easily integrated in dense areas.

The AI@EDGE near-RT RIC development leverages the 5G-EmPOWER platform, which has been developed within SESAME and other H2020 projects. The platform will be further evolved in AI@EDGE, by aligning with ORAN specification effectively making it the first open source near-RT RAN Intelligent Controller. Furthermore, the SESAME Light DC server will be integrated with AI@EDGE GPU HW acceleration capability to support AI capability.

5G-CARMEN (*5G for Connected and Automated Road Mobility in the European Union*) leverages the most recent 5G advances to provide a multi-tenant platform that can support the automotive sector delivering safer, greener, and more intelligent transportation with the ultimate goal of enabling self-driving cars. The key innovations in 5G-CARMEN are centred around developing an autonomously managed hybrid network, combining direct short range V2V (vehicle to vehicle) and V2I (vehicle to infrastructure) communications with long-range V2N (vehicle to network) communications.

To realize its goals, 5G-CARMEN employs different enabling technologies such as 5G New Radio, C-V2X (Cellular vehicle to everything), and secure, multi-domain, and cross-border service orchestration system to provide end-to-end 5G enabled CARMEN services. These technologies are integrated in the LightEdge platform, along with cross-border interworking. In AI@EDGE, the connect-compute platform builds upon the foundations of the LightEdge platform in order to implement support for serverless computing and the Function-as-a-Service paradigm.

The COHERENT (*Coordinated control and spectrum management for 5G heterogeneous radio access networks*) project has been focused upon building advanced network abstractions concepts in order to enable an efficient and scalable solution for network-wide coordination in Heterogeneous Mobile Networks (HMNs). Effective and proactive resource management of heterogeneous RANs is a crucial part of implementing reliable 5G services with guaranteed performance.

In particular, COHERENT's scope has been to design, develop and showcase a novel control framework for 5G heterogeneous radio networks, which leverages the proper abstraction of physical and MAC layers

in the network and a novel programmable control framework, to offer operators a powerful means to dynamically and efficiently control wireless network resources; this could significantly improve capacity, spectrum reuse efficiency, energy efficiency and user experience in the operators' increasingly complex HMNs. The abstraction of network states and functions provides a base for the development of COHERENT Software Development Kit (SDK), which enables programmable control and coordination in heterogeneous radio access networks.

The output of the project has strong relevance for enabling AI/ML-driven automation and coordination over integrated edge infrastructures. AI@EDGE leverages the expertise developed within COHERENT. Furthermore, the aforementioned 5G-EmPOWER platform is also a common link between AI@EDGE and COHERENT, accompanied by AI@EDGE improving on the platform by working within the ORAN specifications.

5G-Xhaul (*5G-XHaul: Dynamically Reconfigurable Optical-Wireless Backhaul/Fronthaul with Cognitive Control Plane for Small Cells and Cloud-RANs*) project main objective is to build a converged optical and wireless software-defined (SD) network solution that is able to transport both backhaul and fronthaul traffic over the same infrastructure, coping with the requirements of 5G RANs. The 5G-XHaul considers a C-RAN model, where Remote Unit (RUs) are connected to Central Units (CU) through high bandwidth transport links, supporting the application of different functional splits options (from fully-distributed to fully-centralised) to perform the Base Band functionality.

In the data plane, the 5G-XHaul considers that in some scenarios the Small Cells will make use of a wireless transport segment, based on mmWave and sub-6GHz technologies, to carry the traffic until the fiber attachment points. In particular, due to its novelty, special effort has been applied to the mmWave solution, including the evaluation of capabilities to support access, fronthaul and backhaul networks and also innovations regarding functional split implementations and mesh networking. In addition, significant evaluations and innovations have been performed in the optical domain, which is based on WDM-PON and TSON technologies. In the control plane, main topics are softwarisation, slicing and multi-tenancy. Applying a hierarchical and scalable architecture, the different heterogeneous transport networks are managed by their specific SDN controllers, while these controllers are coordinated by a higher-level controller which manages the connectivity between the different networks. Finally, a Top Controller supervises the whole architecture, managing the end-to-end paths according to the tenants and the slices. Data transport through the network is achieved by encapsulating the frames at the edge of the different networks into specific tunnels, which permits connectivity between VNFs from common slices but allocated in different networks.

Although AI@EDGE is oriented towards RAN, "outputs" of the 5G-XHaul can provide building blocks for RAN splitting and RAN coordination from the cSD-RAN Controller. In particular, the objective of network slicing is an ongoing problem studied by both 5G-XHaul and AI@EDGE.

VINEYARD (*Versatile Integrated Accelerator-based Heterogeneous Data Centres*) aimed to develop an integrated platform for energy-efficient data centres based on new servers with novel, coarse-grained and fine-grained, programmable hardware accelerators and build a high-level programming framework for

allowing end users to seamlessly utilize these accelerators in heterogeneous computing systems by using typical data-centre programming frameworks.

The deployment of energy-efficient hardware accelerators was used to significantly improve the performance of Cloud-computing applications and reduce the energy consumption in data centers. In the years that passed since project start, the VINEYARD vision has seen its realization as the deployment of hardware accelerators in the Cloud: in 2017, hyperscalers like Amazon, Huawei, Alibaba and Baidu offered FPGA resources to their Cloud users.

VINEYARD developed novel energy-efficient platforms by integrating two types of hardware accelerator: A new-generation dataflow-based accelerator and a novel architecture for FPGA-based (control-flow) accelerators. The dataflow engines (DFEs) are suitable for high-performance computing (HPC) applications that can be effectively represented with dataflow graphs while the latter are used for accelerating applications that need tight communication between the processor and the hardware accelerator(s).

AI@EDGE advances VINEYARD results with the scalable implementation of AI/ML algorithms. Following VINEYARD's paradigm, the project explores FPGA options for accelerating algorithms and pays attention to the energy-efficiency aspect of accelerators. Furthermore, AI@EDGE explores the functionality of processors in the Function-as-a-Service programming paradigm, which has been widely adopted by developers and featured in research and industrial environments.

SELFNET (*A Framework for Self-Organised Network Management in Virtualised and Software Defined Networks*) project has been working in developing a smart autonomic network management framework for NFV/SDN environments that incorporates Self-Organising Network (SON) capabilities. For that purpose, the project has defined a multi-layered architecture that encompasses the following layers: (i) Infrastructure layer, which encompasses both the virtualized and physical network functions managed by the SELFNET framework; (ii) Data network layer, which is in charge of forwarding the data and represents an architectural evolution towards SDN; (iii) SON control layer, which includes the SON sensors, capable of collecting data from the network, and the SON actuators, which enforce actions into the network; (iv) NFV orchestration and management layer, which corresponds to the ETSI NFV MANO and orchestrates the virtual functions embedded in the SON control and data network layers; (v) SON autonomic layer, which provides the mechanisms to enable network intelligence by collecting information about the network behaviour, using that information to "diagnose" the network condition, and deciding what must be done to accomplish the system goals; and the (vi) SON access layer, which encompasses the interface functions that are exposed by the SELFNET framework to external systems, such as Business Support Systems (BSS) and Operational Support Systems (OSS).

The main element to incorporate network intelligence in the management processes is the autonomic manager at the SON autonomic layer. It includes a diagnosis tool to identify the root cause of network problems, a decision-maker to choose the corrective and preventive tactics to deal with the detected and emerging network problems, and an action enforcer that provides a consistent and coherent scheduled set of actions to be enforced in the network infrastructure. The autonomic manager incorporates different types of artificial intelligence (AI), data mining and stochastic algorithms for diagnosing and decision-making.

Specific types of algorithms considered in the project include knowledge inference, prediction and pattern recognition.

To assess the proposed framework, the SELFNET project considered three individual use cases. The first one deals with self-healing capabilities against network failures by detecting and repairing the fault condition of a malfunctioning VNF. The second one deals with providing self-protection capabilities against distributed cyber-attacks by detecting devices that can be hijacked by botnets. Finally, the third individual use case deals with self-optimisation to improve network performance and users' QoE dynamically for 5G users with high-quality video streaming applications. The three use cases have been also assessed in a more complex use case, when the three individual use cases run simultaneously.

SELFNET Framework for self-organized network management, particularly its features for anomaly detection, can serve as a foundation for AI@EDGE. The focus of SON functions in SELFNET lies in the management of NFV, which is an issue under investigation by AI@EDGE.

Summarizing, AI@EDGE builds on the foundations of past and ongoing 5G-PPP-related projects in order to address a significant amount of challenges. AI@EDGE draws inspiration from other projects in particular to tackle issues regarding AI automation and coordination in cloud/edge environments, NFV management and network orchestration. Furthermore, the AI@EDGE connect-compute platform will leverage on a number of existing software components for the different technical domains, as it will be detailed in D4.1. The AI@EDGE consortium has been members of other 5G-PPP projects and act as strong liaisons to other research efforts.

9 Conclusion and next steps

This deliverable represents the first technical deliverable of the AI@EDGE project. Its edition allowed partners to brainstorm on common project-wide objectives and use case specific challenges and objectives. It is the first step of an ambitious scientific and technical program that will be continued within the project as follows.

Within WP2, Task 2.2 will start from the preliminary system and interface specifications to be documented in deliverable D2.2, to come up with a complete specification allowing WP3 and WP4 to take over on the corresponding tasks. Task 2.3 will further specify KPI requirements and will further elaborate on the use-cases socio-economic impact.

Within WP3, Task 3.1 will take over the definition of the architecture of the AI@EDGE network automation platform including the closed-loop network automation aspects. Task 3.2 will focus on the secure and scalable data pipelining while the methods and algorithms for automation in edge and cloud systems will be handled by Task 3.3. Finally, Task 3.4 will take care of implementing the software prototypes and of delivering them for integration in WP4. D3.1 will report on the outcomes of these efforts.

Within WP4, the connect-compute platform will be further detailed and documented in D4.1, including a precise description of the technological enablers, with a particular focus on serverless computing and fabric orchestration, multi-connectivity and hardware acceleration challenges. In particular D4.1 will report the outputs of Task 4.2, on the provisioning of AI-enabled applications, the AIFs reference model, and solutions for the end-to-end decentralized and distributed orchestration of AIFs; it will also report the outputs of Task 4.3, on cross-layer, multi-connectivity, and disaggregated radio access, and of Task 4.4. on GPU and FPGA hardware acceleration. Task4.1, will design, develop, and integrate the AI@EDGE connect-compute fabric.

In WP5, Task 5.1 will start from the D2.1 elements for defining the use cases development and testing process as well as the associated planning. Furthermore, the evaluation procedures, the tests schedule, and the related milestones will be defined. Finally, the task will manage the adaptation and set-up of the use cases subsystems, preparing the test sessions taking as input the AI@EDGE platform developed by WP4 delivering a platform to be integrated on a per-use case basis in T5.2, T5.3, T5.4, and T5.5.

References

[1] H. B. McMahan et al. Communication-efficient learning of deep networks from decentralized data. in Proc. 20th Int. Conf. Artificial Intelligence and Statistics, 2017.

[2] T. Li et al. Federated optimization in heterogeneous networks. In Proc. of Conference on Machine Learning and Systems, 2020.

[3] T. Li et al. Fair resource allocation in federated learning. In Proc. Int. Conf. Learning Representations, 2020.

[4] L. Liu et al. Edge-assisted hierarchical federated learning with non-iid data. arXiv preprint arXiv:1905.06641, 2019.

[5] J. Konecby et al. Federated Learning: Strategies for Improving Communication Efficiency. arXiv preprint arXiv:1610.05492, 2016.

[6] Z. Tao, Q. Li. ESGD: Communication efficient distributed deep learning on the edge. In Proc. Of USENIX Workshop on Hot Topics in Edge Computing {HotEdge 18}, 2018.

[7] N. Carlini et al. The secret sharer: Evaluating and testing unintended memorization in neural networks. In Proc. USENIX Security Symposium, 2019.

[8] R. Bost et al. Machine learning classification over encrypted data. In Proc. Network and Distributed System Security Symposium, 2015.

[9] C. Fung et al. Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866, 2018.

[10] L. Liu et al., Client-Edge-Cloud Hierarchical Federated Learning. arXiv preprint arXiv:1905.006641, 2019.

[11] G. Detal, C. Paasch, O. Bonaventure. Multipath in the middle (box). In Proc. of the 2013 workshop on Hot topics in middleboxes and network function virtualization. 2013.

[12] X. Wei et al. MPTCP proxy mechanisms. draft-wei-mptcp-proxy-mechanism-00, 2014.

[13] M. Boucadair et al. Extensions for Network-Assisted MPTCP Deployment Models. draft-boucadair-mptcp-plain-mode-10, 2015.

[14] Y. Benchaïb, S. Secci, CD. Phung. Transparent cloud access performance augmentation via an MPTCP-LISP connection proxy. In Proc of 2015 ACM/IEEE ANCS.

[15] A. Ford, et al. TCP extensions for multipath operation with multiple addresses. RFC 6824. Internet Engineering Task Force (2013).

[16] 3GPP TS 23.501. System architecture for the 5G system.

[17] Xilinx FPGA Kubernetes plugin: https://github.com/Xilinx/FPGA_as_a_Service/tree/master/k8s-fpga-device-plugin

[18] NVIDIA GPU device Kubernetes plugin: https://ngc.nvidia.com/catalog/containers/nvidia:k8s-device-plugin

[19] C. Fung, C. J. Yoon, and I. Beschastnikh, “Mitigating sybils in federated learning poisoning,” arXiv preprint arXiv:1808.04866, 2018.

[20] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, “Analyzing federated learning through an adversarial lens,” arXiv preprint arXiv:1811.12470, 2018.

[21] H. Kim, J. Park, M. Bennis, and S.-L. Kim, “On-device federated learning via blockchain and its latency analysis,” arXiv preprint arXiv:1808.03949, 2018.

[22] A. Blaise, M. Bouet, V. Conan, S. Secci, "Detection of zero-day attacks: an unsupervised port-based approach", Computer Networks, Vol. 180, Oct. 2020.

[23] S. Ibanez et al. “The P4->NetFPGA Workflow for Line-Rate Packet Processing” In Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '19).

[24] O-RAN Alliance (websit): https://www.o-ran.org/

[25] M. Liyanage, J. Salo, A. Braeken, T. Kumar, S. Seneviratne and M. Ylianttila, "5G Privacy: Scenarios and Solutions," *2018 IEEE 5G World Forum (5GWF)*, 2018, pp. 197-203, doi: 10.1109/5GWF.2018.8516981

[26] E. Tufan, C. Tezcan and C. Acartürk, "Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network," in IEEE Access, vol. 9, pp. 50078-50092, 2021, doi: 10.1109/ACCESS.2021.3068961 [Titel anhand dieser DOI in Citavi-Projekt übernehmen]

[27] ETSI GR MEC 017 V1.1.1 (2018-02) Mobile Edge Computing (MEC); Deployment of Mobile Edge Computing in an NFV environment

[28] 5TONIC laboratory (website): https://www.5tonic.org

[29] H2020 Sat5G project (website): https://www.sat5g-project.eu

[30] H2020 5G-ESSENCE (website): https://www.5g-essence-h2020.eu

[31] AI4EU platform (website): https://www.ai4eu.eu

[32] P4 language framework (website): https://opennetworking.org/p4

[33] TCP Replay software (website): https://tcpreplay.appneta.com

[34] C-V2X Use Cases Volume II: Examples and Service Level Requirements. 5GAA White Paper. 20 oct. 2020. https://5gaa.org/wp-content/uploads/2020/10/5GAA_White-Paper_C-V2X-Use-Cases-Volume-II.pdf

[35] Airspan software (website): https://www.airspan.com

[36] Druid software (website): https://www.druidsoftware.com

[37] Jetson Xavier NX Devkit (website): https://developer.nvidia.com/embedded/jetson-xavier-nx-devkit

[38] Jetson AGX Xavier Developer Kit (website): https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit

[39] NVIDIA V100 (webpage): https://www.nvidia.com/en-us/data-center/v100

[40] NVIDIA CUDATollkit (webpage): https://developer.nvidia.com/cuda-toolkit :

[41] NVIDIA Tesla T4 (webpage):]https://www.nvidia.com/en-us/data-center/tesla-t4

[42] NVIDIA Jetson Modules (webpage): https://developer.nvidia.com/embedded/jetson-modules

[43] NVIDIA Cuda X (webpage): https://www.nvidia.com/en-us/technologies/cuda-x

[44] SciKit Learn software (website): https://scikit-learn.org/stable

[45] Tensorflow software (website): https://www.tensorflow.org

[46] Pytorch software (website): https://pytorch.org

[47] Xilinx (website): https://www.xilinx.com/

[48] Nvidia (website): https://www.nvidia.com/en-us/

[49]: Xilinx Virtex Ultrascale plus (website): https://www.xilinx.com/products/silicon-devices/fpga/virtex-ultrascale-plus.html

[50] Xilinx Alveo (website): https://www.xilinx.com/products/boards-and-kits/alveo.html

[51] Xilinx Vitis platform (webpage): https://www.xilinx.com/products/design-tools/vitis/vitis-platform.html

[52] LightEdge platform (website): https://lightedge.io/

[53] 5G-Empower platform (website): https://5g-empower.io/