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

D2.2 Preliminary assessment of system architecture, interfaces specifications, and techno-economic analysis

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

D2.2 Preliminary assessment of system architecture, interfaces specifications, and techno-economic analysis	
WP	WP2 – Use cases, requirements analysis, and system design
Responsible partner	i2CAT Foundation (i2CAT)
Version	1
Editor	Miguel Catalan-Cid (i2CAT)
Authors	Miguel Catalan-Cid (i2CAT), Estefania Coronado Calero (i2CAT), Nicola di Pietro (ATH), Arif Ishaq (ATH), Marco Centenaro (ATH), Omar Anser (INRIA), Jonathan Proietto (INRIA), Stefano Secci (CNAM), Flávio Brito (EAB), Neiva Linder (EAB), Per Ödling (ULUND), Cristina Costa (FBK), Arfan Wahla (FBK), Luigi Girletti (ATOS), George Avdikos (8Bells), Stelios Koumoutzelis (8Bells), Antonino Albanese (ITL), Bredan McAuliffe (SRS). Babak Mafakheri (SPI), Marco Marchetti (CRF), Miguel Rosa (AERO), Nimish Sorathiya (DFKI), George Lentaris (ICCS), Bengt Ahlgren (RISE), Giampiero Mastinu (POLIMI), Nicola Pio Magnani (TIM)
Reviewers	Jovanka Adzic (TIM), Neiva Linder (EAB), Roberto Riggio (UNIVPM), Cristina Costa (FBK)
Deliverable Type	R
Dissemination Level	PU
Due date of delivery	31/12/2021
Submission date	31/01/2022

Version History				
Version	**Date**	**Authors**	**Partners**	**Description**
0.1	04/10/2021	Miguel Catalan-Cid, Estefania Coronado Calero , ALL	I2CAT, ALL	ToC, First Draft and initial contributions from all the involved partners
0.2	10/12/2021	Miguel Catalan-Cid, Flàvio Brito, ALL	i2CAT, EAB, ALL	Consolidated Sections (1,2,3,4,5,6) for first review
0.3	15/12/2021	Miguel Catalan-Cid, Cristina Costa, Jovanka Adzic, Per Ödling, Neiva Linder, Flávio Brito	I2CAT, FBK, TIM, EAB	Content revised after first review, acronyms revised, references added
0.4	17/12/2021	Miguel Catalan-Cid, Estefenia Coronado-Calero, Neiva Linder, Flávio Brito, Nicola di Pietro, Cristina Costa, Arfan Wahla, Antonino Albansese, Bredan McAuliffe, George Avdikos, Omar Anser, Jonathan Proietto	I2CAT, EAB, ATH, FBK, ITL, SRS, 8BELLS, INRIA	Additional contributions, preparation for second review
0.5	20/12/2021	Jovanka Adzic,	TIM	Second review
0.6	22/12/2021	Miguel Catalán-Cid, Arfan Wahla, George Avdikos, Omar Anser	I2CAT, FBK, 8BELLS, INRIA	Modifications according to review inputs
0.7	23/12/2021	Jovanka Adzic	TIM	Final review
0.8	25/01/2022	Roberto Riggio, Cristina Costa, Miguel Catalán Cid, Jonathan Proietto, Neiva Linder, Flavio Bristo	UNIVPM, FBK, i2CAT, INRIA, EAB	Quality Control Review and modifications according to the inputs
1	30/01/2022	Irene Facchin	FBK	Final review

Disclaimer

Table of Contents

List of Tables

List of Figures

Glossary	
3GPP	3rd Generation Partnership Project
4G, 5G, 6G	Fourth, Fifth, Sixth Generation of cellular networks
5G	5th Generation of mobile communication networks
5GAA	5G Automotive Association
5GC	5G Core Network
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AIF	Artificial Intelligence Function
AMF	Access and Mobility Management Function
APN	Access Point Name
AR	Augmented Reality
ATSSS	Access Traffic Steering, Switching and Splitting
AUSF	Authentication Server Function
BVLOS	Beyond Visual Line of Sight
C-V2X	Cellular Vehicular communication
CapEx	Capital Expenditures
CCP	Connect-Compute Platform
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPU	Central Processing Unit
CRUD	Create, Read, Update, and Delete
CU	Centralized Unit
DE	Deliverable Editor
DL	Downlink
DNN	Data Network Name
DNS	Domain Name System

DSP	Digital Signal Processing
DU	Distributed Unit
DP	Dynamic Payback
EDA	Electronic Design Automation
ETSI	European Telecommunications Standards Institute
FaaS	Function as a Service
FL	Federated Learning
FPGA	Field-Programmable Gate Array
FPV	First Person View
gNB	gNodeB – 5G radio base station
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
IRR	Internal Rate of Return
ISG	Industry Specification Group
I-UPF	Intermediate User Plane Function
KPI	Key Performance Indicator
LUT	Look Up Table
LCM	Life Cycle Management
mAP	Mean Average Precision

MDAS	Management Data Analytics Service
MEC	Multi-access Edge Computing
MEAO	Multi-access Edge Computing Application Orchestrator
MECP	Multi-access Edge Computing Platform
MECPM	Multi-access Edge Computing Platform Element
MEO	Multi-access Edge Computing Orchestrator
MEP(m)	Multi-access Edge Computing Platform management
MEPM-V	MEC Platform Manager - Network Function Virtualization
ML	Machine Learning
mMTC	massive Machine-Type Communication
MNO	Mobile Network Operator
MQTT	Message Queuing Telemetry Transport
MTC	Machine Type Communications
MTO	Multi-Tier Orchestrator
MPTCP	Multipath Transmission Control Protocol
NAS	Non-Access Stratum
NBi	NorthBound Interface
NF	Network Function
NFO	Network Function Orchestrator
NFV	Network Function Virtualization
NFVI	Network Function Virtualization Infrastructure
NFVO	Network Function Virtualization Orchestrator
NG	Next Generation
NPV	Net Present Value
NR	New Radio
NRF	Network Repository Function
Near-RT RIC	Near-Real-Time RAN Intelligent Controller

Non-RT RIC	Non-Real-Time RAN Intelligent Controller
NSA	Non-Standalone (5G)
NSAP	Network and Service Automation Platform
NSSAI	Network Slice Selection Assistance Information
NSSF	Network Slice Selection Function
NWDAF	5G NetWork Data Analytics Function
OAM	Operation And Maintenance
OpEx	Operational Expenditures
OSS	Operations Support System
OWL	Web Ontology Language
PCF	Policy Control Function
PCIe	Peripheral Component Interconnect express
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PFCP	Packet Forwarding Control Protocol
QoS	Quality of Service
RAN	Radio Access Network
RAM	Random Access Memory
RAT	Radio Access Technology
REST	Representational State Transfer
RIC	RAN Intelligent Controller
RNIS	Radio Network Information Service
ROI	Return On Investment
ROS	Robot Operating System
RRU	Remote Radio Unit
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
SA	Standalone (5G)
SBA	Service-Based Architecture
SCTP	Stream Control Transmission Protocol
SDG	Sustainable Development Goals
SDN	Software-Defined Networking
SDR	Software-Defined Radio
SMF	Session Management Function
SMO	Service Management and Orchestration
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
TRL	Technology Readiness Level
UC	Use Case
UDM	Unified Data Management
UDR	Unified Data Repository
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-reliable Low Latency Communications
VIM	Virtual Infrastructure Manager
VNF	Virtual Network Function
NVFM	Virtual Network Function Manager

VR	Virtual Reality
V2I	Vehicle to Infrastructure
V2N	Vehicle to Network
V2V	Vehicle to Vehicle

Executive Summary

This deliverable details the intermediate AI@EDGE system architecture, describing its main components, interfaces and workflows (M8). Together with D3.1 and D4.1, it provides a complete view of the key technical challenges and contributions of AI@EDGE project and defines the scope of the software prototypes to be used in the trials. This deliverable also reports on the preliminary techno-economic analysis and impact assessment (MS9) and extends the use case related KPIs, introduced in D2.1, to present a consolidated draft of AI@EDGE KPIs (M7).

1 Introduction

This deliverable provides the intermediate AI@EDGE system architecture, including the specification of the main interfaces and protocols available, and the preliminary techno-economic analysis, which includes a first definition of the addressed KPIs.

The first part of the deliverable details the initial contributions of Task 2.2, which aims to design and specify the functional components of the end-to-end system architecture and the communication interfaces between them. In particular, the deliverable reports the progress of the first stage of the task, which has been focused on:

- Studying existing technological platforms, state-of-the-art protocols and frameworks for network automation and edge computing, giving preference to open-source solutions and outputs of other 5G-PPP projects.
- Analyzing the contributions of AI@EDGE project partners being defined within the scope of WP3 and WP4, the expected interactions among them, and their relationship with the use case requirements defined in D2.1.
- Deciding technical choices, fundamental components and trade-offs regarding the system architecture, interfaces and workflows.
- Designing the intermediate AI@EDGE system architecture by refining the baseline architecture presented in D2.1 according to the inputs from the aforementioned analyses and decisions.

In this sense, the achievements reported in this part of the deliverable report the progress towards fulfilling the project's overall Objective 1 *(To specify the functionalities of integrated solutions based on identified technical and business requirements towards a network automation platform converging 5G, cloud-native, and secure AI/ML for the support of highly elastic real-world use cases*) and Milestone MS8 (*Intermediate System Architecture and Interfaces available*).

The second part of the deliverable covers the initial activities of Task 2.3, which is focused on the definition of the project KPIs, the socio-economic impact assessment and the techno-economic analysis. Regarding the KPIs, the deliverable details the draft KPIs (Milestone MS7), extending the use-case specific KPIs presented in D2.1 and linking them to the different domains considered within the project: technical, societal, economic and environmental. In addition, the deliverable introduces the preliminary techno-economic analysis of the project (Milestone MS2.4), which is presented in the form of dialogues or discussions on practical applications of edge computing with the objective to enrich the discussion about technical specifications in connection with the AI@EDGE architecture. This preliminary analysis considers five categories (E-health, Transportation, European food supply, Europe's electric power system and European indigenous minorities), giving a wider perspective beyond the scope of the use cases of the project and creating the basis for fulfilling the project's overall Objective 2 *(To assess the impact of AI@EDGE from the societal standpoint and to integrate the lessons learned into the final solution. A detailed techno-economic analysis focused on OTTs and telecom operators will also be used for the definition of the AI@EDGE platform requirements.*). Future revisions of the socio-economic impact assessment and the techno-economic analysis to be provisioned D2.3 and D2.4 will be more focused on the specifics of each of the use cases.

The document is organized as follows. Section 2 introduces the AI@EDGE system architecture vision, which describes how the designed architecture addresses key research challenges and brings innovations to the beyond 5G ecosystem. Section 3 details the AI@EDGE system architecture, describing its main

components. Section 4 specifies the main interfaces and workflows of the presented architecture, focusing on network and service automation functionalities. Section 5 provides the preliminary techno-economic analysis of the project. Section 6 details AI@EDGE draft KPIs. Finally, Section 7 concludes the deliverable and draws the next steps.

2 AI@EDGE System Architecture Vision

The objective of this chapter is to revisit the AI@EDGE initial vision focusing primarily on a system architecture context. Section 2.1 starts discussing an overall AI@EDGE system wanted position. In sequence, Section 2.2 provides an overview of AI@EDGE System Architecture current baseline, equally offering the reader an architecture concept update since the preliminary AI@EDGE system architecture views described in the D2.1 deliverable.

2.1 AI@EDGE System Architecture Position

As 5G, IoT and Edge solutions gain traction in telecom networks and keep transforming business and industries, advances in AI/ML stimulate even further new customer service applications while being also a powerful tool to address raising network operational complexity. In this way, telecom operators can play a central role in providing telecom networks as the platform for whole new AI-enabled application ecosystems, where a fully AI-native network edge is a fundamental piece of the puzzle. Enabling such an AI-native network edge will require a dynamic edge orchestration platform and a new approach on how we design our telecom data driven architecture. AI@EDGE project is taking a closer look at these technology trends and aims to leverage the concept of reusable, secure, and trustworthy artificial intelligence for network and service automation. Therefore, an AI@EDGE network and service automation platform will be developed and validated, focusing on the following technology breakthroughs:

- AI/ML for closed loop automation.
- Privacy preserving, machine learning for multi-stakeholder environments.
- Distributed and decentralized connect-compute platform.
- Provisioning and Orchestration of AI-enabled applications
- Hardware-accelerated serverless platform for AI/ML
- Cross-layer, multi-connectivity, and disaggregated radio access.

We foresee AI/ML based closed-loop automation solutions will play an important role in enabling the full potential of Multi-access Edge Computing, particularly combined with AI/ML compute deployment enhanced by specialized hardware targeting AI applications to bring better performance and computing power. AI/ML computation here addresses distributed model training and inference through the AI@EDGE network automation platform combined with the capabilities of a distributed and decentralized connect-compute platform.

The combination of these technology enablers will lay the foundation toward providing a fully autonomous zero-touch network and service management platform.

An overview on technical topics:

AI@EDGE intends to bring innovation in the mobile networks by investigating technical topics that are challenging for the industry and academy. Such topics include, e.g., the usage of federated learning for network automation and supporting federated and distributed machine learning-based applications with the network and orchestration infrastructure. Distributed computing as fog computing available in end-user and vertical devices is proposed in this project that it is beyond 5G.

Regarding specialized hardware targeting artificial intelligence applications, AI@EDGE work plan includes performance and computing power evaluation. Virtualization and specialized hardware usage is expected in such a way that the end user will not perceive if a service is run on say a CPU or an FPGA: these will just appear as accelerated functions. Bringing AI capability to networked applications is a key ambition of the AI@EDGE project.

When it comes to service management and orchestration and anomaly detection, we work with two categories: attacks and infrastructure impairments. The former is addressed with AI/ML-based algorithms to detect and mitigate attacks based on active sensing. The latter needs an AI/ML solution to recognize the type and location of impairments and perform a root cause analysis.

Additional research questions are expected to be identified while advancing beyond 5G and towards 6G architecture solutions, e.g., with technical push coming from new AI/ML types of learning (e.g., federated learning), privacy and security of data, orchestration, resource layer optimization, latency critical and service reliability, etc.

As AI@EDGE will be the host of several AI applications of different fields (e.g., energy-saving, automotive, content curation, etc.), it is important to manage the dependencies between these AI functions to optimize their placement in the network. Therefore, AI@EDGE intends to support AI natively by introducing the Artificial Intelligence Function (AIF) concept, understanding relation with hardware acceleration, centralization and distribution of data vs. intelligence, etc. We remind the reader that we use the term AIF to refer to the sub-components of AI-enabled applications and services deployed over the AI@EDGE platform. The AIF conceptual model is a basic element of the project. AIF conceptual model includes both the data and control interfaces that allow them to be monitored and orchestrated.

2.2 AI@EDGE Baseline System Architecture

The AI@EDGE baseline system architecture is represented in Figure 1. The architecture is composed of two main layers: the Network and Service Automation Platform (NSAP) layer and the Connect-Compute Platform (CCP) layer. The AI@EDGE Network and Service Automation Platform contains the Multi-Tier Orchestrator (MTO), the Intelligent Orchestration Component, the Non-Real-Time RAN Intelligent Controller (Non-RT RIC) and the Slice Manager. The Connect-Compute Platform aims to bring distributed computation over the cloud, far edge and near edge. The architecture will be aligned with ETSI/MEC reference architecture [1] and will implement some components of ETSI/MEC (e.g., MEC Application Orchestrator (MEAO), MEC Platform Management (MEP(m)), Virtual Infrastructure Manager (VIM), Network Function Virtualization Orchestrator (NFVO)) and O-RAN components such as Near-Real-Time RIC (Near-RT RIC).

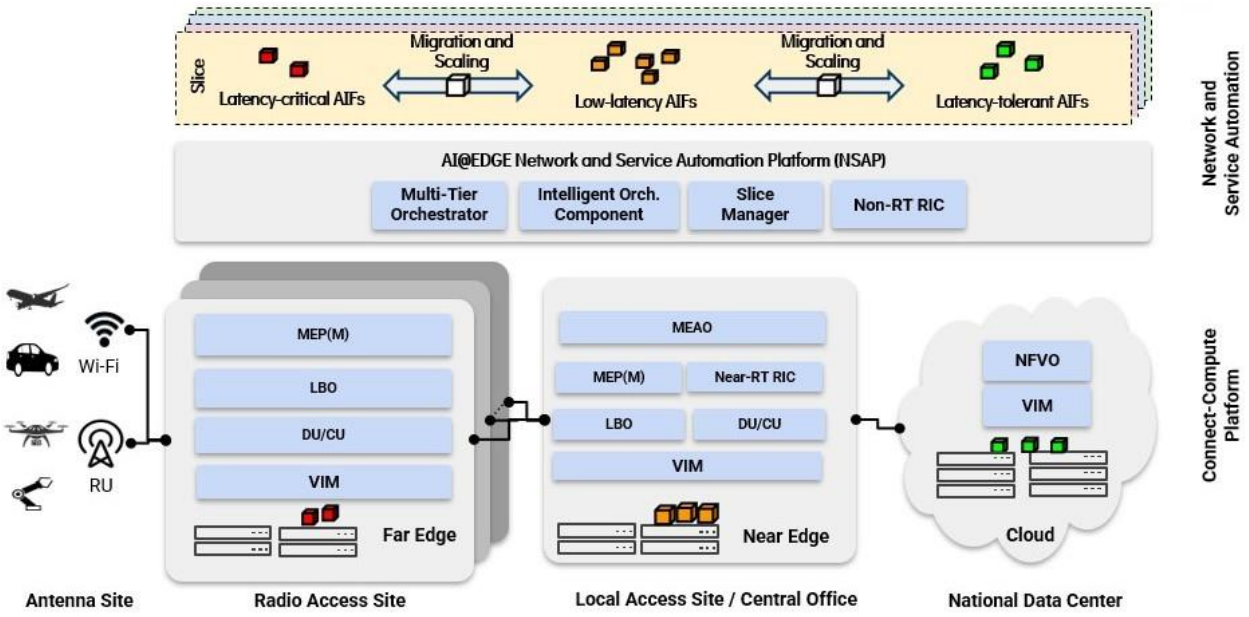

Figure 1 AI@EDGE system architecture.

Closed loops are an important enabler for automation in the AI@EDGE system architecture. Figure 2 represents how closed loops are designed as part of the system. There are three types of closed loops proposed in the system architecture:

- Resource Closed loops that are associated with domain applications and that are specific to each site. They can be deployed in the Cloud, Near Edge, and Far Edge. There is no direct relationship between each other (represented by the red closed loops in Figure 2).
- NSAP Closed loop (represented by the blue closed loop in Figure 2). It will be deployed in the NSAP domain and can interact and receive input from the Multi-Tier Orchestrator, the Intelligent Orchestration Component, the Non-Real-Time RIC and the Slice Manager.
- Cross-Domain Closed loop (represented by the orange closed loop in Figure 2). This closed loop can automate the system architecture by taking inputs from the two domains: NSAP and Connect-Compute Platform. This closed loop can interact in a master/slave scenario with the other closed loops by sending information or commands to modify the slave closed loops.

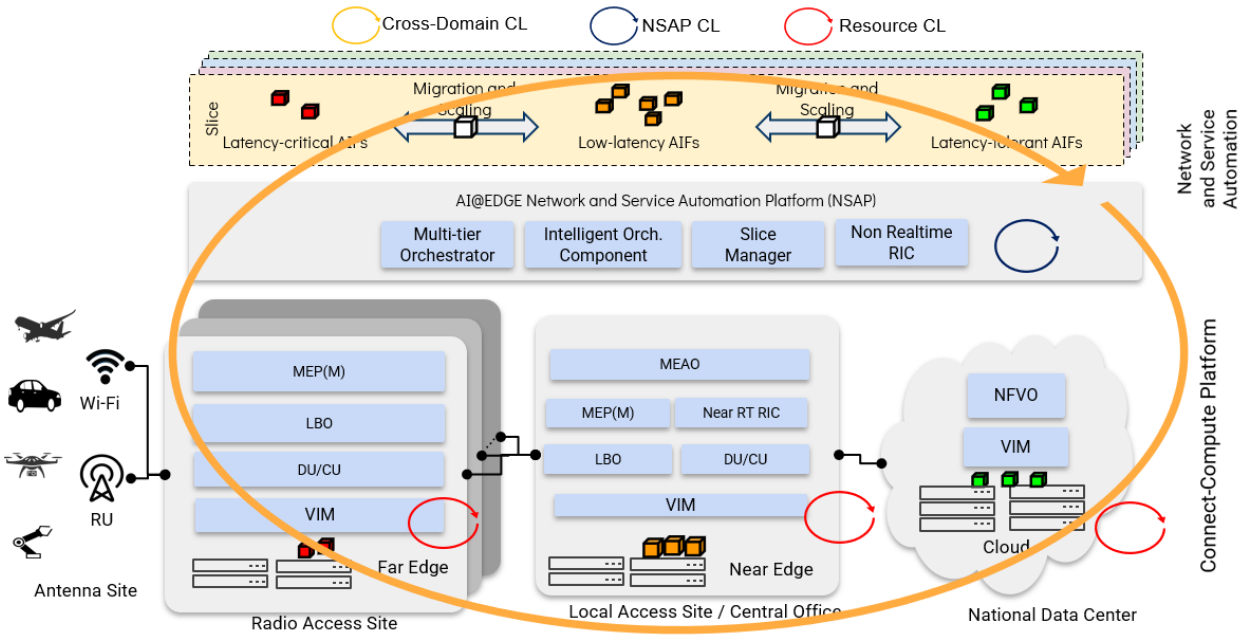

Figure 2 AI@EDGE system architecture including Closed Loops.

In the following, the AI@EDGE system architecture design components will be briefly described. Starting with the NSAP, the Multi-Tier Orchestrator, together with the support of the Intelligent Orchestration Component, will automate the management of the different Orchestrators present in the multi-tier MEC layer such as the MEAO. The second component in NSAP is the Non-RT RIC, which provides non-real-time intelligence in a RAN domain. Finally, the Slice Manager will manage MEC and 5G resources and group them in common multi-tier slices.

AI@EDGE will be based on the concept of AIFs that will be deployed through Cloud, Near Edge and Far Edge. These deployed AIFs can run different kinds of applications. Depending on the site where AIFs are deployed, different latency levels are needed since these applications need to access the desired data without suffering much delay into it. Therefore, a data pipeline system is needed in the AI@EDGE system architecture to provide data to the applications with the latency and granularity they need. More details about NSAP components can be found in Section 3.1.

In the Connect-Compute Platform, the first component listed here is the VIM that will provide the necessary virtual infrastructure to run MEC applications. The MEC applications will be managed by the MEC Platform Management. All these elements are common for the Near Edge and the Far Edge and will be described further in Section 3.2.1. However, in the Near Edge, there is the addition of the Near-RT RIC component and the cloud will contain the NFVO. AI@EDGE will also integrate some 5G components such as Multipath Transmission Control Protocol (MPTCP) Proxy and Centralized Unit (CU) / Distributed Unit (DU) split. More details about 5G system components can be found in Section 3.2.2.

AI@EDGE aims to bring innovation in components such as the MEAO, the Non-RT RIC, the Near-RT RIC and the MEC platform. These innovations will be focused on the implementation of such components in AI@EDGE platform as well as adding extensions to these elements and defining the interaction among them. In the next chapters, a discussion of the AI@EDGE Intermediate System Architecture will be

presented, including the description of these main components, the involved interfaces and the envisioned workflows.

3 AI@EDGE Intermediate System Architecture

This chapter will provide a description of the AI@EDGE system architecture with a focus on its components. It will start with the description of each component of the AI@EDGE Functional Reference Architecture diagram represented in Figure 3. This figure also details the main interfaces among each component of the AI@EDGE architecture which will be described in Chapter 4. Besides the grouping based on NSAP and CCP blocks, Figure 3 also differentiates between MEC System and 5G System levels at the CCP, and considers the utilization of Cloud, Near and Far Edge resources to place the different components, giving a view of the cross-platform and cross-system interactions needed to fulfill AI@EDGE objectives.

This chapter is organized as follows: in Section 3.1, the NSAP blocks are presented: the Multi-Tier Orchestrator, the Intelligent Orchestrator Component, the Non-RT RIC and the Slice Manager. Concluding this section, the Data Pipeline is presented. Section 3.2 focuses on the Connect-Compute Platform components such as the MEC system components: MEC Apps/AIFs orchestrator, Intelligent Orchestrator Component, MEC platform Management and the MEC host. The last major component shown in this chapter is the 5G system components such as the Near-RT RIC, the 5G RAN and Core, and the MPTCP proxy.

Since this deliverable has as objective to present a complete view of AI@EDGE System Architecture, we advise the reader to seek further technical details about the Network and Service Automation system and methods in the Deliverable D3.1, and similarly, read further about the Connect Compute Platform in the deliverable D4.1.

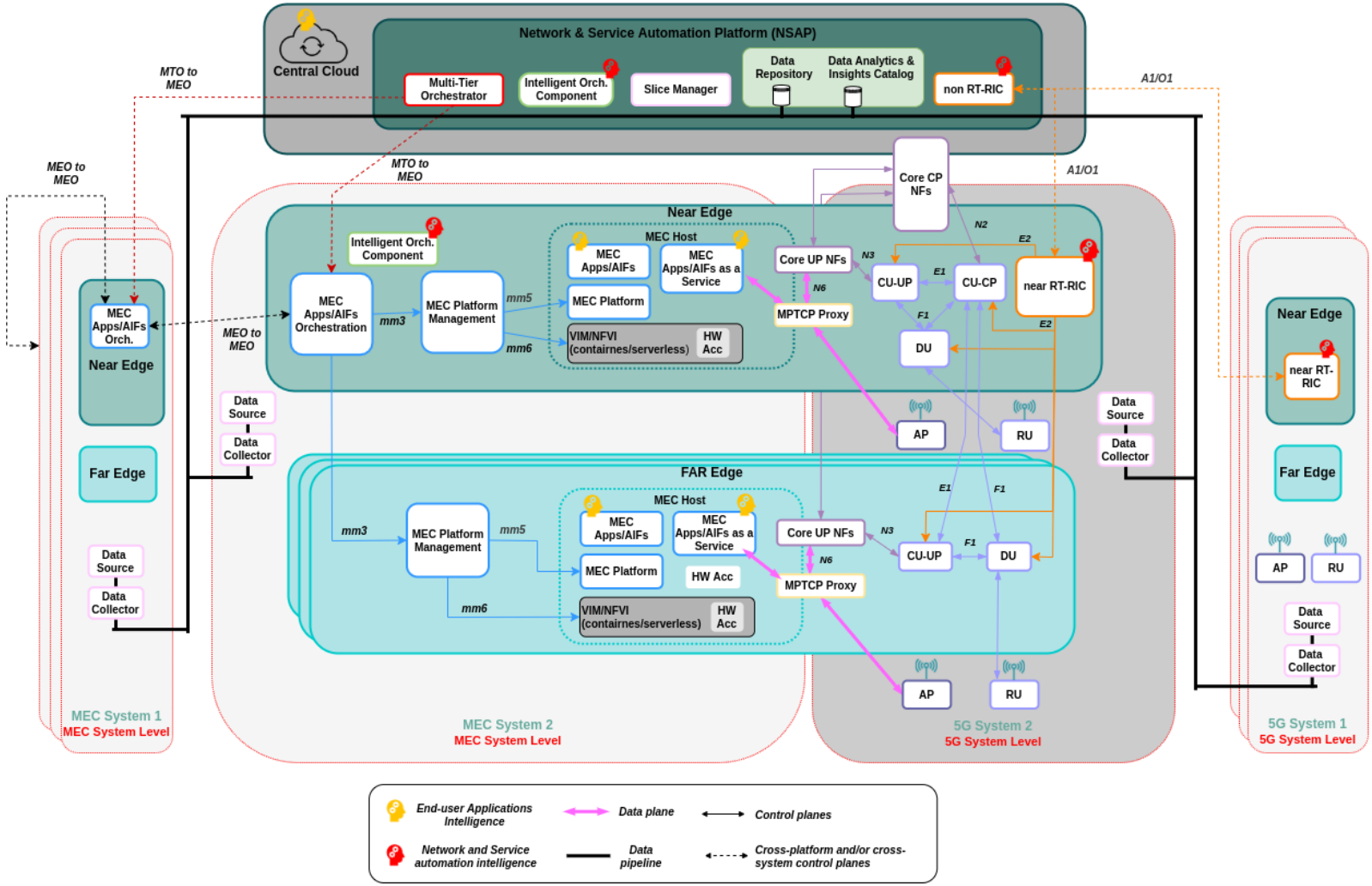

Figure 3 AI@EDGE Functional Reference Architecture.

3.1 Network and Service Automation Platform

We group under the NSAP the components located at the Cloud that provide the means to properly control and optimize the performance of the MEC and 5G Systems deployed at the Near and Far Edges. On the one hand, the MTO, together with the Intelligent Orchestration Component, automates the coordination of the different Orchestrators present in the multi-tier MEC System. On the other hand, the Non-RT RIC is the key element of the O-RAN's Service Management and Orchestration (SMO) component to enable control-loop automations at the 5G RAN. Finally, the Slice Manager intelligently manages MEC and 5G resources to create multi-tier slices. In addition to these components, a Data Pipeline system is required to enable scalable and trustworthy information exchange across computing overlays.

3.1.1 *Multi-Tier Orchestrator*

The Multi-Tier Orchestrator (MTO) represents the entry point of the Network and Service Automation Platform for the operations related to the instantiation of MEC applications. The main objective of this element is to enable the interaction with various types of orchestrators through a series of southbound clients, such as the MEC Orchestrators located at the Near Edges and the NFV Orchestrators located at the cloud. However, it does not perform any sort of onboarding, placement or migration operations, and offload such tasks to the underlying orchestrators. In this sense, when it receives an instantiation request, it is able to trigger the required API invocation chains attending to the specific orchestrator needed. The requirements of the application to be deployed are specified on a descriptor, which includes the service descriptor (including, among others, application id, application name, application provider, application descriptor, etc.) and a list of additional (and optional) requirements needed for the application to be deployed, such as MEC services used and/or necessary computing resources (RAM, CPU and disk). At the moment it is being studied the possibility of collecting, at the MTO, telemetry data coming from the various orchestrators. Given that this data would provide a view of the status of the various resources available as per the orchestrators' knowledge, it would be necessary to request to the Intelligent Orchestration Component the provision of placement decisions for the incoming requests.

3.1.2 *Intelligent Orchestration Component*

The Intelligent Orchestration Component will interact with the rest of the NSAP components to enhance their functionalities, focusing on the application of AI/ML methods. This component might leverage functionalities of fault, security and resource management, and enable the interaction between MEC and 5G Systems to provide intelligence to MTO and non-RT RIC procedures. Due to its dependence on WP3 and WP4 contributions, the concrete definition of its functionalities and interactions is still under discussion, and will be extended in future revisions of the architecture.

3.1.3 *Slice Manager*

The Slice Manager provides control over the lifecycle of network slices in the AI@EDGE platform. Starting from a slice template, it is able to create slice instances and trigger their deployment over multiple MEC Systems and 5G Systems. In particular, it enables the Create, Read, Update, and Delete (CRUD) operations over slice instances and it is in charge of creating a mapping between SLA requirements and a logical sliced network. To do so, the Slice Manager interacts with the different NFVOs and VIMs to deploy the VNFs needed for the slice, and with the Core and RAN controllers to enable the allocation of the needed network

resources to a particular slice. Further details over the set of functionalities and the workflows of the Slice Manager will be further studied in D2.3, while the implementation details of it will be defined in D4.2.

3.1.4 *Non-Real-Time RAN Intelligent Controller*

The NSAP in AI@EDGE implements a subset of functionalities of O-RAN's SMO layer [1]. In particular, as shown in Figure 3, the Non-Real-Time RIC is the key element to implement intelligent closed-loop automations related to the 5G System at the NSAP level, and managing/optimizing the 5G System automations present at the Connect-Compute Platform level. In this sense, the main functionality of the Non-Real-Time RIC is to implement the A1 interface termination, which provides the necessary methods to manage the closed-loop automations performed at the available Near-RT RICs and xAPPs. These methods or functions comprehend the management of policies (A1-P interface), the exposure of enrichment information available at the SMO level (A1-EI interface) and the management of machine learning workflows (A1-ML interface). Although this latter interface is still being specified by O-RAN [2] and we don't envision implementing it within the scope of this project, we will monitor its specification status due to its relevance to AI@EDGE topics.

The Non-RT RIC also exposes different SMO functionalities to the rAPPS that might be used to implement closed-loop automations at the NSAP level. For instance, via the O2 interface, rAPPs can access SMO's NFO functionalities for xAPPS onboarding or NFs deployment at the O-Cloud (equivalent to the 5G System in Figure 3), while the O1 interface exposes to them Operation and Maintenance (OAM) and slicing management features. Also, aggregated data from the 5G Systems managed by the SMO, or even from external applications, can be made accessible to rAPPs via the Non-RT RIC. In any case, although some of these exposed functionalities will be considered in the different workflows defined in Chapter 5, the focus of the project will be the incorporation and demonstration of the A1 interface.

Additional information on the Non-RT RIC architecture and the developed functionalities can be found in deliverables D3.1 and D4.1.

3.1.5 *Data Pipeline*

It is not desirable to use a dedicated data pipeline system for each application, given the amount of data that is increasing nowadays. An alternative approach is to deploy a shared data pipeline that can "democratize" the same data to different applications using different preprocessing tools to fit each application's requirements.

AI@EDGE aims to design and validate a network capable of deploying AI/ML solutions running in the Cloud, Far Edge and Near Edge. The AI/ML solutions that will be designed need a granularity of data that a data pipeline system shall capable of delivering. They also need to preprocess a high amount of data to meet the necessary KPIs of the applications, such as those used in the project's Use Cases (UC).

In UC1 (*virtual validation of vehicle cooperative perception*), the vehicle shall send data regarding its location to detect the surrounding traffic scenario in real-time. If the data is not delivered with the desired granularity and minimal latency, the vehicles will not be able to take the best decision in the roundabout scenario (e.g., take a different direction or reduce the speed). In UC2 (*secure and resilience orchestration of large (I)IoT networks*), to be able to respond and orchestrate the devices in an (I)IoT environment, the data access shall be facilitated. In UC3 (*Edge AI assisted monitoring of linear infrastructure using drones in BVLOS scenario*), the drones will monitor large scenarios such as farms. They will be connected to the central office through a 5G network. There, a decision-making system based on the data received from the drone will reply to the drone operator with the suggested decision to be taken. Therefore, an optimal way

to collect and preprocess the data is necessary. Concluding, for UC4 (*Smart content and data curation for in-flight entertainment services*), a data-driven approach is needed to collect and preprocess the data to make an optimal decision about which content is better to deliver to the clients.

Therefore, as the UCs need an optimal tool to collect and preprocess the data, a data pipeline system designed to deliver a high amount of data with the desired granularity is a key-enabler for AI@EDGE deployment and computation of data in a Cloud, Far Edge and Near Edge domains. More details about the data pipeline system are described in D3.1.

3.2 Connect-Compute Platform

The AI@EDGE Connect-Compute Platform (CCP) 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 platform allowing developers to take advantage of the new capabilities offered by 5G using well established cloud-native paradigms.

As was shown in Figure 3, we can group the functions provided by the CCP in MEC System Components and 5G System Components. Figure 3 also highlights the MEC System, which is a collection of MEC hosts and the MEC management necessary to run MEC applications. As per [3], it contains:

- MEC Apps/AIFs Orchestration functions (including the Intelligent Orchestration Component), corresponding to the MEO or MEAO from the ETSI MEC architecture;
- MEC Platform Management functions, including MEC Apps Life Cycle Management (LCM);
- MEC hosts.

The 5G System in AI@EDGE also comprehends the NFs that form the virtualized 5G RAN and Core. As denoted by the inclusion of the Near-RT RIC, the architecture of the 5G System is based on the O-RAN specification, which enables the application of network automation intelligence. In addition, the 5G System includes the MPTCP proxy element to provide multi-path aggregation at the transport layer in multi-connectivity multi-RAT scenarios.

3.2.1 *MEC System Components*

3.2.1.1. MEC Apps/AIFs Orchestration

The MEC Apps/AIFs Orchestration element maintains an overview of the complete MEC system. It is responsible for the placement and migrations of MEC Apps or AIFs when they are requested. The placement decision can be based on simple rules or classic optimization methods or rely on the Intelligent Orchestration Component that uses AI/ML methods. The Orchestrator can be placed at the Near Edge, and it can also be placed at another site as far as it has a direct connection with the Edge Sites (both Far Edge and Near Edge). It corresponds to the MEO in the ETSI MEC architecture, and to MEAO plus NFVO in the MEC NFV synergized architecture.

3.2.1.2. Intelligent Orchestration Component

Intelligent Orchestration Component is a plug in of the MEC Apps/AIFs orchestration module. It relies on the use of AI/ML methods for providing intelligent orchestration of the MEC Apps and AIFs inside the MEC system. As in the case of the Intelligent Orchestration Component located in the NSAP, and according

to WP3 and WP4 outputs, future revisions of the architecture will detail the expected functionalities and interactions of this element.

3.2.1.3. MEC Platform Management

This component includes the MEC specific functionalities: MEC Platform Element Management and MEC Apps rules and requests management. It is responsible for sending the MECP the configurations needed to manage the MEC Apps, such as traffic rules, DNS configurations, and services requested or provided. These functionalities are included in the MEC Platform Manager (MECPM) and correspond to the MEPM-V in the MEC NFV architecture. The MEPM-V has no direct access to the VIM. This component is also responsible for the LCM of the MEC application instances. These functionalities are included in the MECPM and correspond to the VNFMs in the MEC NFV architecture.

3.2.1.4. MEC HOST

MEC host is an entity that contains a MEC platform and a virtualization infrastructure which provides compute, storage and network resources to MEC applications [3]. The MEC host is strategically placed by the Edges of the network to provide computation and storage capabilities near the Access Network and provide, between other advantages, lower latency. To this aim, the 5G traffic is steered towards the MEC host where it can be processed (more details on the integration of the MEC host with the 5G infrastructure is given in the following paragraph). The MEC host can be therefore considered as an Edge cloud able to host MEC applications and User Application AIFs. An AIF can run on the MEC Host stand-alone or as a MEC Service, providing services to other MEC Apps on the same or other MEC Hosts. The MEC Platform provides the functionalities required to run MEC applications and AIFs, enabling them to provide and consume MEC services. The MEC Platform itself can provide several MEC services, such as the Radio Network Information Service (RNIS). The MEC host can also provide hardware acceleration services.

3.2.2 *5G System Components*

3.2.2.1. 5G RAN

3GPP Release 15 first introduced 5G New Radio (NR) technology with multiple specification drops between 2017 and 2019. This next evolution of mobile wireless brings higher performance targets for throughput, latency and scale as well as greater waveform flexibility.

The 5G NR base station is the gNodeB (gNB) with new interfaces to the core network (NG) and other gNBs (Xn). The gNB itself has a flexible architecture supporting functional splits into the Centralized Unit (CU), the Distributed Unit (DU) and the Radio Unit (RU). A variety of different functional splits for the gNB are defined, supporting different use cases and performance requirements.

Building upon the 3GPP specifications, the O-RAN alliance has introduced an architecture for the 5G RAN, including RAN controllers and a Service Management and Orchestration framework. Their Near-RT RIC interacts with the gNB over the E2 interface to enable more efficient and cost-effective radio resource management.

Within the AI@EDGE project, SRS will provide a 5G RAN via the open-source srsRAN project. The srsRAN software suite currently includes 5G Non-Standalone (NSA) eNodeB and UE applications, with the initial 5G Standalone (SA) gNodeB coming in Q2 2022 and SA UE expected at a later date.

srsRAN applications are implemented in efficient and portable C/C++ supporting a wide range of baseband hardware platforms, including x86, ARM and PowerPC. The srsRAN gNodeB will support functional split

interfaces, including splits 6, 7 and 8. Split 7 supports network deployments using commercial off-the-shelf O-RAN Remote Radio Unit (RRU) devices. Split 8 supports popular Software-Defined Radio (SDR) front-end devices such as the NI USRP family with a generic baseband I/Q interface. The gNodeB will further support the E2 interface for interaction with third-party controllers.

Further modifications in line with the project's needs may be implemented in an ad hoc fashion when needed.

3.2.2.2. 5G Core

In a 5G system, the 5G Core Network (5GC) covers two main separate roles, one related to the so-called user plane and one to the control plane: the former consists of connecting the RAN (and therefore the UEs) with external data networks (the Internet, a cloud, an application server, a LAN, etc.) via the User Plane Function (UPF); the latter, instead, consists of overseeing all the network functionalities and processes that are not related to the radio access, like user authentication, subscriber data management, policy control, connectivity and mobility management, or exposure of services towards application functions. An end-to-end 5G system is such only if its core network fully complies with the 3GPP standards [4]. The reference architecture of AI@EDGE's connect-compute platform is endowed with a fully virtualized 5GC, which guarantees increased flexibility and adaptability. The main network functions (NFs) that compose a 5GC are the following: the Access and Mobility Management Function (AMF), the Session Management Function (SMF), the UPF, the Unified Data Management (UDM), the Authentication Server Function (AUSF), the Unified Data Repository (UDR), the Policy Control Function (PCF), the Network Repository Function (NRF), and the Network Slice Selection Function (NSSF). The control-plane NFs interact via standardized service-based interfaces and are logically separate from the user plane functionalities (performed by the UPF).

Among the numerous features that characterize the 5GC, it is worth mentioning the possibility of distributing core NFs (particularly the UPF) at the near or far edge of the network, bringing them as close as possible to the RAN equipment and the UEs. This is a key enabler for edge computing since selected traffic coming from the data plane can be maintained local (with respect to the UE) and routed directly towards the servers where edge applications run, thus implementing the system architectures standardized by the ETSI MEC ISG [5]. Furthermore, the 5GC supports network slicing and its resources can be associated with different slices (in coordination and cooperation with the RAN) to isolate independent services with different performance guarantees.

3.2.2.3. Near-Real-Time RAN Intelligent Controller

The Near-RT RIC is a logical function pioneered by O-RAN Alliance to enable RAN programmability and service optimization. With an open architecture, Near-RT RIC allows on-boarding of RAN control applications for near-real-time fine-grain performance optimization and policy tuning. ML-based algorithms are implemented as external applications, called xApps. These are deployed on the Near-RT RIC to deliver specific services such as inference, classification, and prediction pipelines to optimize the per-user quality of experience, controlling load balancing and handover processes, or the scheduling and beamforming design.

The Near-RT RIC implements the logic to control and enable optimization of the RAN functions in O-CU and O-DU in near-real-time intervals through the E2 interface. The Near-RT RIC logic is implemented in the form of xApps, which are independent of the Near-RT RIC and may be provided by any third party. The E2 interface enables direct association of xApp and the RAN functionality for collecting information from the RAN. The Near-RT RIC can reconfigure the O-CU and O-DU functions dynamically on the basis

of the policies configured by the Non-RT RIC through the A1 interface, still through the E2 interface. More information regarding the implementation of near-Real-Time RIC in the scope of this project can be found in Deliverable 4.1.

3.2.2.4. MPTCP Proxy

In the presence of multiple Radio Access Technologies (RAT), such as Wi-Fi, 4G and 5G, the multi-connectivity environment can be exploited by means of multi-path aggregation using transport-layer technologies. In particular, the MPTCP extension of TCP can be used for this purpose. As envisioned in the 5G specifications under the ATSSS variant of the 5G core cluster, an MPTCP proxy can be used in the core network to aggregate multiple RAT. Going beyond the integration inside the 5GC UPF, in AI@EDGE we investigate different deployment modes of the MPTCP proxy to aggregate multiple RATs, and possible wireline technologies (namely, Ethernet), with as reference Use Case 4. D4.1 describes in detail the different variants that include the placement of the MPTCP proxy after the 5GC toward the application server, within the UPF, and possibly also before the 5GC. Its actual deployment could be within the MEC host, with an adjusted virtual link routing to evaluate its positioning at different levels with respect to the 5GC.

4 AI@EDGE main interfaces and workflows

This section introduces the main interfaces and workflows of the AI@EDGE architecture that have been identified at this stage of the project. In this sense, note that the presented interfaces and workflows only comprehend cross-platform interactions between NSAP and CCP components, and internal interactions between CCP components. Future deliverable D2.3 will discuss internal NSAP interactions, which have been left out of the scope of this deliverable due to their lack of standardization and the need for additional research and design efforts that were not possible at this stage of the project.

4.1 Main interfaces

This Section describes the main interfaces of the AI@EDGE system architecture.

4.1.1 *MEC System to MTO*

This interface can be considered a proprietary implementation of the *Mm1* interface of the ETSI MEC architecture that represents the reference point between the Multi-Tier Orchestrator and the MEC Orchestrator. This interface is extended from the initial design proposed as an output of the 5Gcity project [6], and that was used to mediate between the MTO and cloud/edge orchestration domains to trigger high-level actions for lifecycle management such as instantiation of network services. The interface is presented as an API able to trigger the actions on the corresponding orchestrators. In AI@EDGE this interface is extended to also account for the deployment of applications at the edge through a MEC Orchestrator, using an extended descriptor that includes special requirements described in the project, such as the support for hardware acceleration, etc. In addition to the instantiation of applications at different orchestrators, at the moment the consortium is also analyzing the use of this interface to concentrate on the MTO telemetry data coming from the underlying orchestrators to be used for smarter deployment decisions.

4.1.2 *MEC System to MEC System*

This interface is to be designed and implemented within the AI@EDGE project to communicate two MEC systems in a distributed way. In particular, this interface aims to interconnect the MEC Orchestrators handling the management of two MEC systems. The main functionality of this interface is related to the migration of applications across MEC systems, including migration requests and resource availability checks to be performed before initiating a migration process. The aim of this interface is to allow the local functioning of each MEC system as well as the required applications migration in a distributed manner, even if the Multi-Tier Orchestrator suffers some kind of failure. At the moment, the interface design is being undertaken taking as a reference indication provided by ETSI on inter-MEC systems coordination [7] and the current requirements of the project, and more details will be provided in Deliverable D4.2.

4.1.3 *MEC System Interfaces*

The main reference points to be considered within the MEC system are as follows:

- ***Mm3*:** This reference point relates the MEO with the several MEC Platform Managers under its control in the same MEC system. In order words, it allows explicitly keeping track of the available MEC platforms and services. In the proposed architecture, the operations envisioned to be

supported by such points include the application instantiation requests, application lifecycle management, and traffic rule management.

- ***Mm5:*** The *Mm5* reference point is placed between the MEC platform manager and the MEC platform. This interface is used to perform the configuration of the platform and of the application's rules and requirements, to support the application lifecycle procedures and the management of application relocation, etc. This reference point is not specified by the ETSI standard, since it depends on the implementation of the MECPM and MECP.
- ***Mm6:*** The *Mm6* reference point between the MEC platform manager and the virtualization infrastructure manager is used to manage virtualized resources, e.g., to realize the application lifecycle management. This reference point is not specified by the ETSI standard, since it depends on the specific implementation of the MECPM or VNF-M (in the MEC-NFVI scenario) and the NBi of the VIM. In the MEC in NFVI scenario, this interface could correspond to the Vi-Vnfm ETSI NFI reference point.

4.1.4 *Non-RT RIC to Near-RT RIC*

The *A1* interface is defined by O-RAN in [9] and [10] as the interface connecting non-RT and Near-RT RICs. As introduced in previous Sections and shown in Figure 4, this interface provides three main services or functionalities: Policy Management, Enrichment Information and ML Model Management. While the interactions related to *A1-ML* services are still under definition, *A1-P* and *AI-EI* operations and types are specified in [10] and [11], respectively.

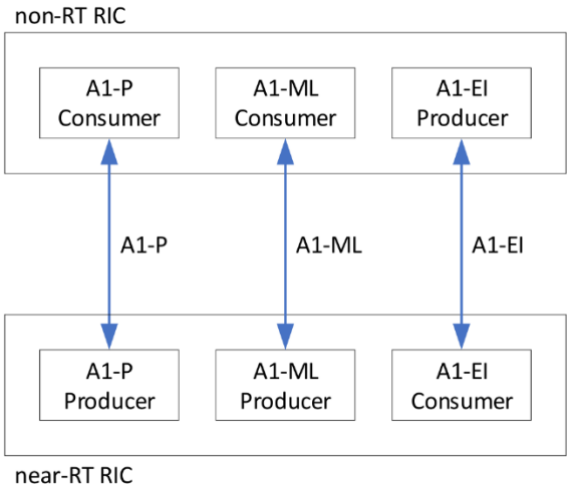

Figure 4 Service framework for the A1 Services [10].

Figure 5 shows the *A1-P* operations as specified in [10]. As described in D4.1, the main focus in AI@EDGE is on the PUT operations needed to create and update policies related to xAPPS in the Near-RT RIC. The concrete policies to be demonstrated in AI@EDGE will be detailed in the next deliverables (i.e., D2.3 and/or D4.2) and will be based on the definition of scopes and statements, as specified in [11].

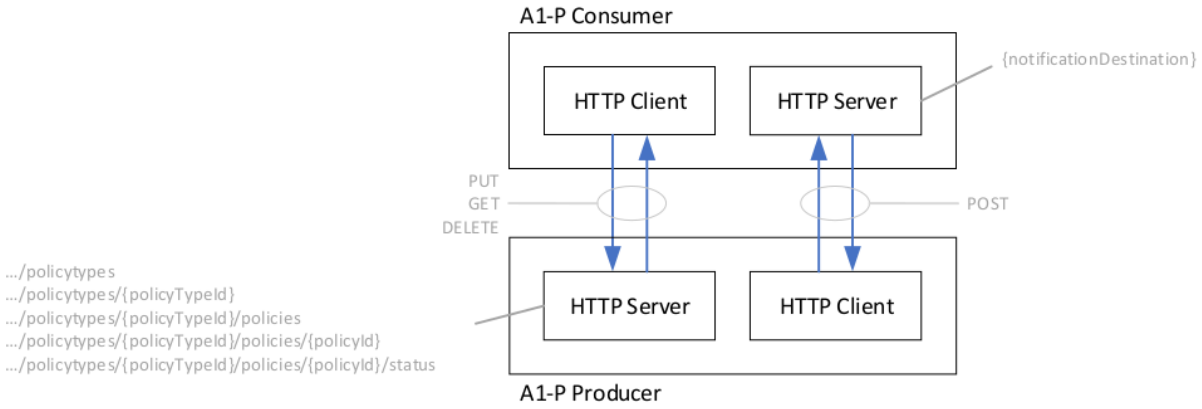

Figure 5 A1-P operations and HTTP roles [10].

On the other hand, Figure 6 shows the endpoints related to AI-EI functionalities. As introduced in D4.1, the main procedures considered in AI@EDGE will be the creation of EI jobs and the delivery of EI job results. Again, the concrete definition of EI types, EI jobs and their results will be detailed in future AI@EDGE deliverables.

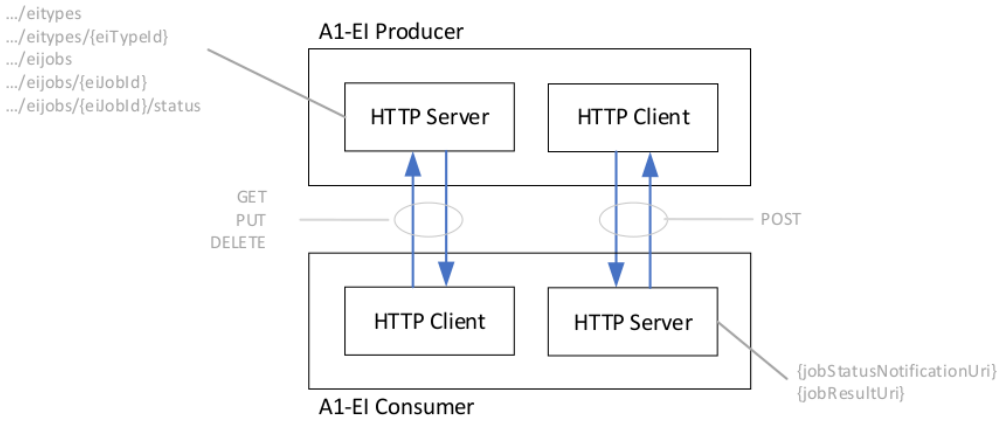

Figure 6 A1-EI operations and HTTP roles [10]

4.1.5 *Near-RT RIC to 5G RAN*

The ORAN WG3 in [12] defines the E2 as the Interface connecting the Near-RT RIC and one or more E2 nodes (O-CU-CPs, one or more O-CU-UPs, and one or more O-DUs). As specified in [12], the E2 Node consists of: (i) E2 Agent used to terminate the E2 interface and to forward/receive E2 messages. (ii) One or more RAN functions that the Near-RT RIC controls, i.e., supporting Near-RT RIC Services. (iii) Other RAN functions that do not support Near-RT RIC Services.

With respect to the E2 interface, the Near-RT RIC consists of: (i) Database holding data from xApp applications and E2 Node and providing data to xApp applications and, (ii) E2 Termination function and, (iii) One or more xApp applications. Figure 7 illustrates these interfaces and components.

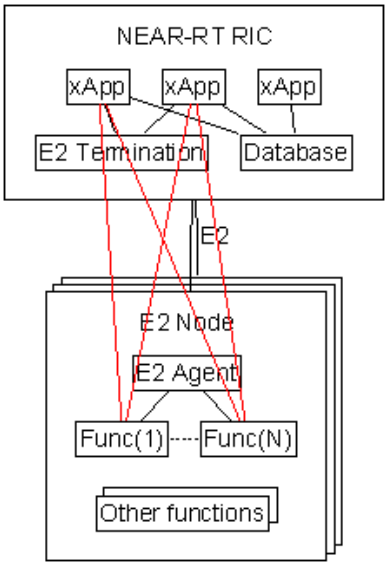

Figure 7 Relationship between Near-RT RIC and E2 Node [12] .

The E2 functions are grouped into two categories [12]:

- Near-RT RIC services: Near-RT RIC uses the following services provided by the E2 nodes: REPORT, INSERT, CONTROL and POLICY.
- Near-RT RIC support functions: Interface Management (E2 Setup, E2 Reset, E2 Node Configuration Update, Reporting of General Error Situations) and Near-RT RIC Service Update (i.e., an E2 Node initiated procedure to inform Near-RT RIC of changes to list of supported Near-RT RIC services and mapping of services to functions).

4.1.6 *5G RAN*

In the 5G RAN, a gNB may consist of a Central Unit (CU) and one or more Distributed Units (DUs). A CU and DU are connected via the F1 interface. The central unit (CU) can be further split into the control plane (CP) and user plane (UP), with the F1 interface also being split. The CU-CP is connected to the DU via the F1-C interface while the CU-UP is connected to the DU via the F1-U interface. The CU-CP and CU-UP are then connected by the E1 interface. This is illustrated in Figure 8:

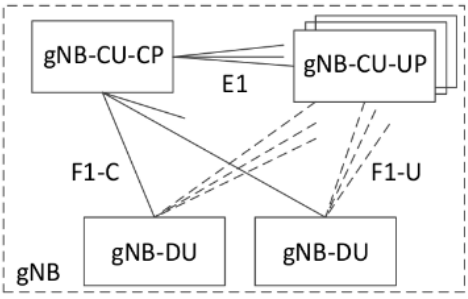

Figure 8 Disaggregated 5G RAN – Components and interfaces

The following rules must be followed for the composition of the gNB:

- There should be only one CU-CP (it is possible to have more than one for redundancy, but only one can be active at the same time).
- There can be one or more CU-UPs.
- There can be one or more DUs.
- One DU can be connected only to a single CU-CP through the F1-C interface.
- One CU-UP can be connected only to a single CU-CP through the E1 interface.
- A single DU can be connected to multiple CU-UPs under the control of the same CU-CP through the F1-U interface.
- A single CU-UP can be connected to multiple DUs under the control of the same CU-CP through the F1-U interface.

4.1.7 *5G Core Network*

Figure 9 depicts the 5GC as described in Section 3.2.2.2, the other elements of a 5G system the 5GC is connected with, and the reference points between them. According to the 3GPP standard [5], such reference points are:

- **N1:** reference point between the UE and the AMF. It carries Non-Access Stratum (NAS) messages between the UE and the AMF, transparently through the gNB. In particular, it is exploited to send to the AMF UE information concerning mobility, connection, and sessions.
- **N2:** reference point between the RAN and the AMF. It is used by the AMF mostly to control and configure the gNBs. The N2 reference point carries signalling messages exchanged via the NG application protocol over Stream Control Transmission Protocol (SCTP). Such messages support operations like PDU Session resource management, UE context transfer, configuration updates, and mobility procedures.
- **N3:** reference point between the RAN and the UPF. This is the user-plane interface between the gNB and the 5GC, used to carry the user plane PDUs towards the UPF.
- **N4:** reference point between the SMF and the UPF. It carries Packet Forwarding Control Protocol (PFCP) messages over User Datagram Protocol (UDP), used to interconnect the UP and the CP. Through the N4 interface, the SMF controls the packet processing and forwarding in the UPF.
- **N6:** reference point between the UPF and a data network. This interface provides IP connectivity from the UPF to an external Data Network. It connects a 5G network with “the rest of the world”, allowing the UE to reach the Internet, another private or public network, or a public or private cloud. Making the MPTCP Proxy reachable via the UPF through the N6 interface is one of the deployment options that AI@EDGE is considering (cf. Section 3.2.2.4).

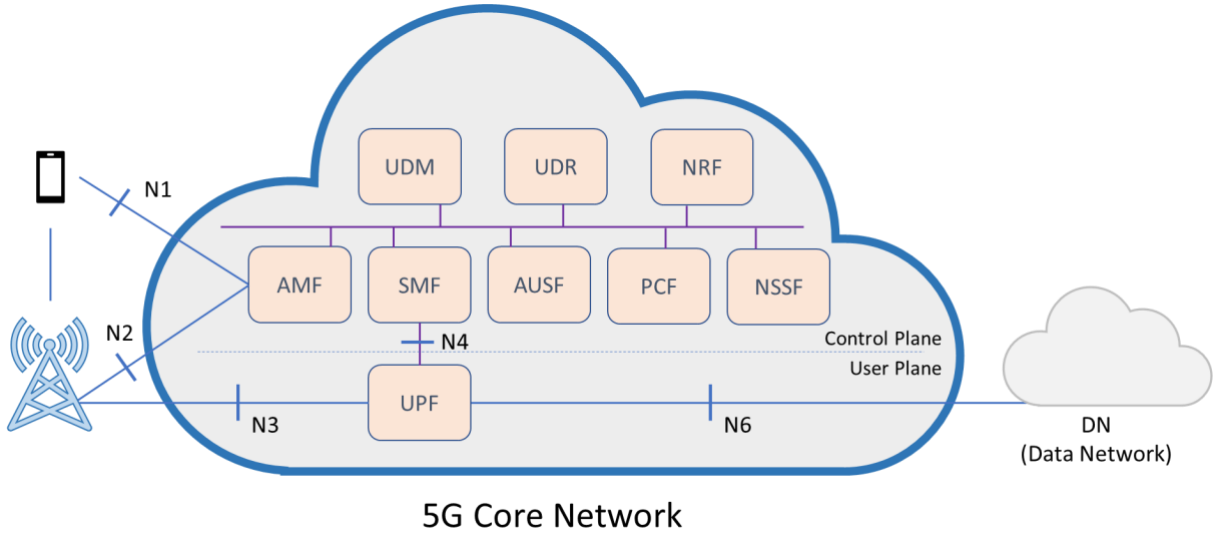

Figure 9 The reference points between the 5G Core Network and the other elements of a 5G system.

4.2 Main Workflows

This Section details the main workflows defined at the moment of writing this document. Redefinitions and new workflows may be reported in future deliverables like D2.3.

4.2.1 AIFs Orchestration

This section gives a preliminary description of the end-to-end AIFs orchestration process. At this time, there are three main phases. The first is the AIF graph definition and validation phase. Next, the willing to deploy a chosen AIF graph from those already defined and validate and do the AIFs parameters auto-configuration. And at the end, the generation of the definition/deployment files and the whole deployment part from the Multi-Tier Orchestrator (MTO).

During the validation stage, at a high level it is a question of validating both the semantics of the AIF graph, built using the user's inputs and the associated data for the construction of the AIFs.

In the case of ML functions, each AIF has to be configured with some parameters or hyper-parameters in order to provide the best possible results.

Once the AIF parameters are configured and saved, it is possible to create all the files needed for the deployment of the complete application in a run-time environment through the MTO orchestrator.

4.2.1.1. Phase 1: AIF graph and data validation

In order to deploy an AI-based application across a converged connect-compute platform, a user has first to provide a description of its application through an AIF graph. A simple application can be built using the reference AIF model, while a more complex one can be obtained through the chaining of multiple AIFs models. Therefore, an application can be represented in the form of a graph called AIF graph. Such a graph can be composed of one or more nodes linked through their interfaces to model pairwise relations between AIFs.

Since an AIF graph must be structured and semantically correct, executing a specific validation of that graph is necessary. In particular, it validates the essential points needed for the proper functioning of the AIFs to be deployed on the Connect-Compute Platform. It also checks the semantic of the AIF graph (e.g., ensuring that the AIF's interfaces are used correctly or verifying that the used AIFs in the graph are declared in the AIF catalogue).

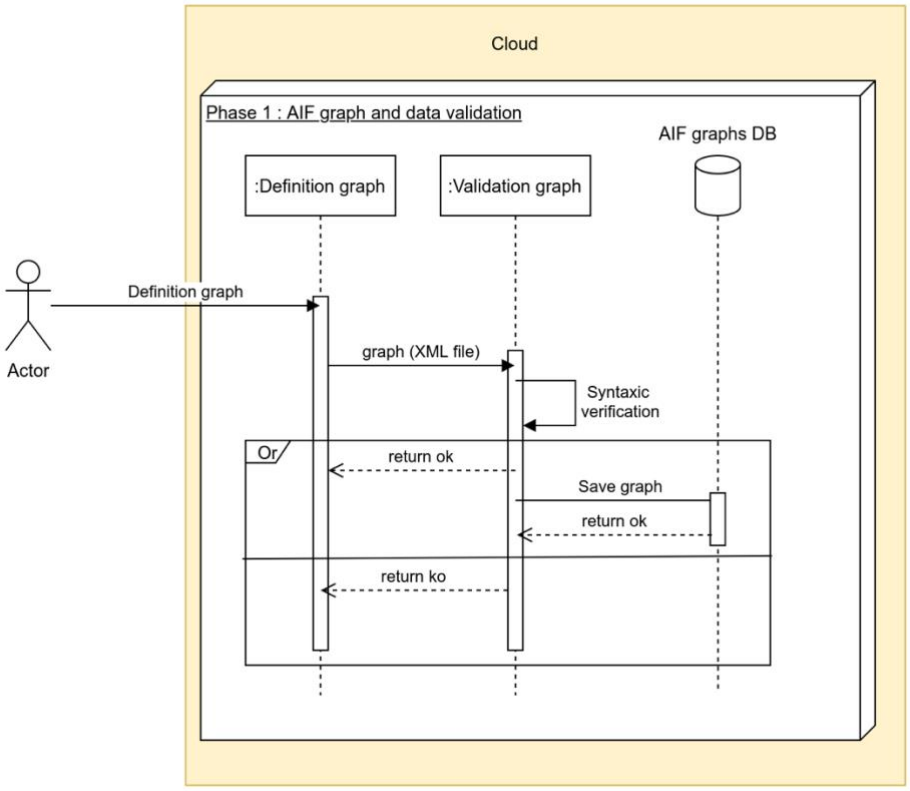

Figure 10 AIF graph and data validation workflow

4.2.1.2. Phase 2: AIFs configuration

Once an AIF graph is defined, validated, and stored in the AIF Catalog, a user can proceed with the configuration. Each AIF can have an "auto-config" property: if its value is "True", the AIF will need to be auto-configured. The data needed to do such a configuration can be provided in the following modes: by the "Data Analytics and Insights Catalogue" database filled by the Data Pipeline, by a public dataset or by a manual dataset uploaded by the user. The AIFs configuration workflow is shown in Figure 11.

In addition to parameters learned during the training process, an AIF can also have several hyper-parameters. The configuration stage aims at choosing the optimal set of these hyper-parameters. In the field of machine learning, this stage is called hyper-parameter selection [14]. Most hyper-parameter optimization techniques involve multiple training cycles of the AIF. Techniques like Hyperband and Bayesian optimization minimize the number of training cycles needed. This type of optimization needs to take into account specific requirements imposed by the environment where the AIF will run.

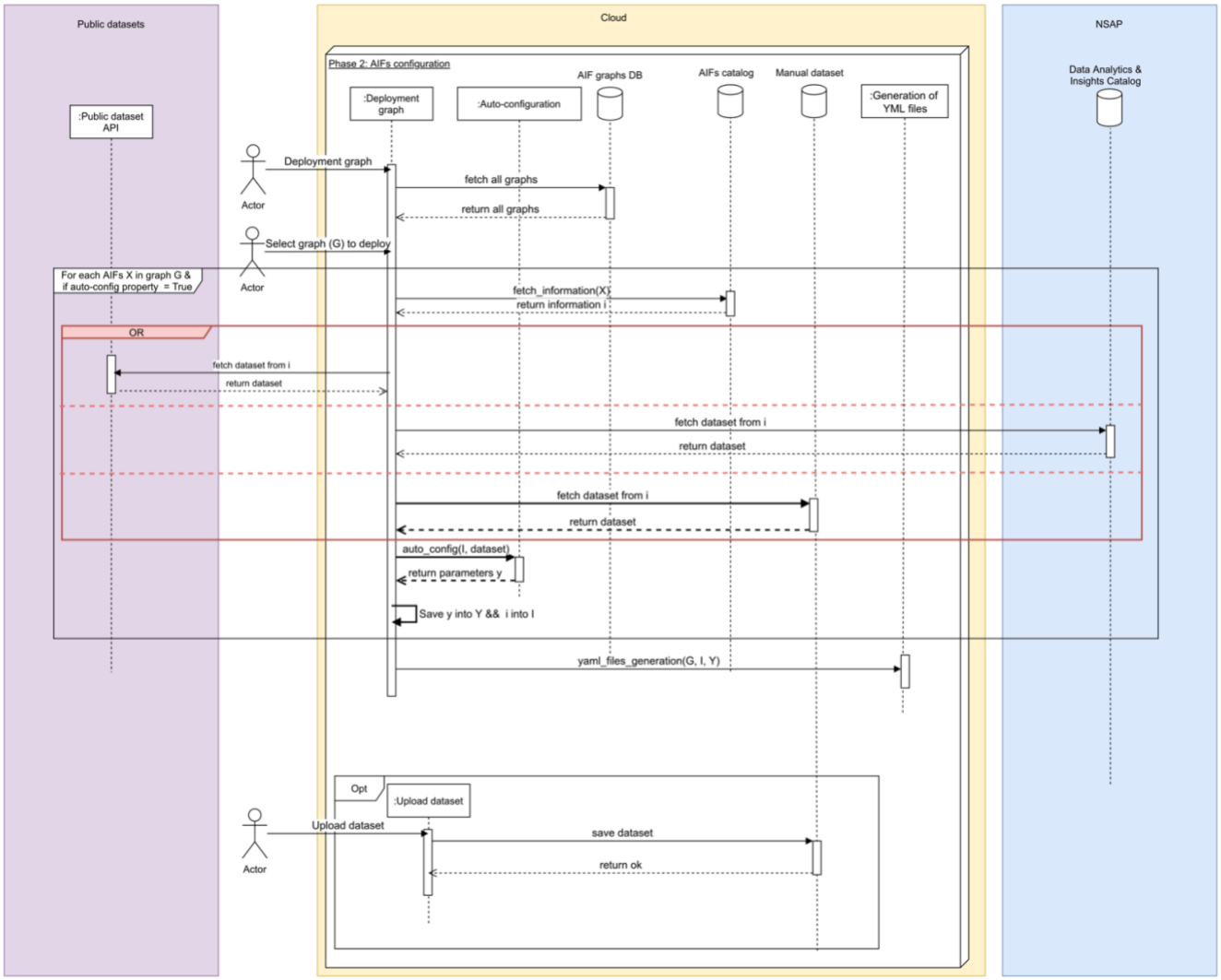

Figure 11 AIFs configuration workflow

4.2.1.3. Phase 3: Generation of descriptor files

All the files describing the application will be based on YAML, a data serialization language often used to write configuration files of applications. The application here has multiple AIFs, and each AIF will have one or several files. The first step is to fetch all the information we need (e.g., image, hardware resources need, etc.) from an AIF catalogue. Then, those fetched values are overridden with the new values from the AIF graph to have a list of final values which are used to create the descriptor files.

To deploy an application, the MTO needs some information: where to put the AIF and how many resources it needs in order to start and run properly. Thanks to the description files, the Multi-Tier Orchestrator (MTO) will know exactly all the resources and constraints the application needs (CPU/RAM resources, hardware needs (e.g., GPU), latency, ...) and will be able to deploy an application/AIF on the right place.

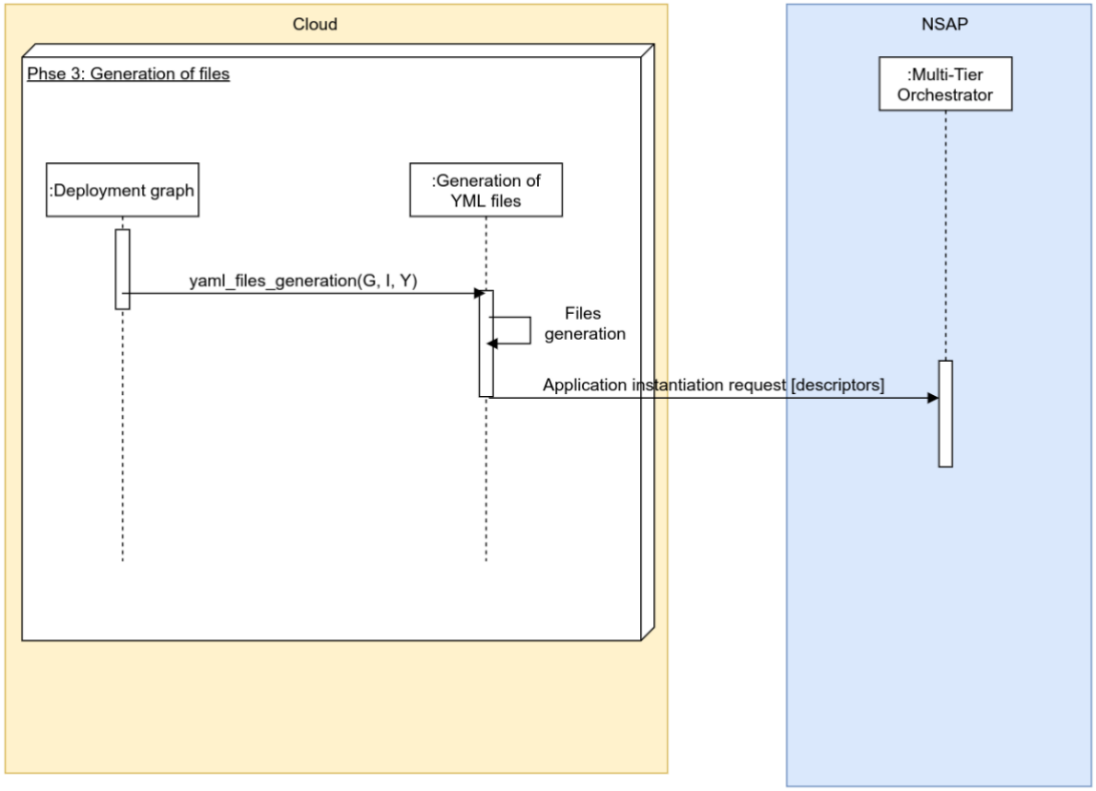

Figure 12 Generation of descriptor files

4.2.2 *MEC workflows*

4.2.2.1. Application Instantiation

The process envisioned within the AI@EDGE architecture for application instantiation is detailed in Figure 13. In particular, three main groups of modules are depicted, namely cloud, MEC system 1 and MEC system 2. The cloud site comprises the components of the NSAP as well as the orchestration capabilities in this layer (comprising an NFV Orchestrator and VIM), which are considered suitable for those applications that are less restricted in terms of latency and requiring higher computational capabilities. By contrast, the other two blocks represent the operations of two MEC systems with identical responsibilities. For the sake of clarity, only the instantiation process at one of such MEC systems is shown. The specific components of the MEC systems are the ones described in Section 3.2.1. This workflow considers that the PDU session of the UE requesting the application instantiation has been previously established, and is out of the scope of this process.

The instantiation process is triggered from the Operations Support System (OSS) when receiving a request to instantiate an application. As a result, the request reaches the entry point of the NSAP, whose role is played by the Multi-Tier Orchestrator (MTO). This module then provides the requirements specified by the new application to the intelligent orchestration module, together with the status information of each underlying system (e.g., available resources and services), in order to get an intelligent decision for the

application placement. This decision will at first instance purely indicate if the application must be forwarded to a specific MEC system or to the cloud through the MEC-System-to-MTO interface.

In the first case, the MEC orchestrator of the specific MEC system selected, receives the new request, and compose the specific descriptor as understood by the MEC platform managers. As it is the case with the main intelligent orchestration module, a local instance is present at each MEC system, with the aim of providing placement decisions on the specific edge tiers within the MEC system (i.e., Near Edge, or specific Far Edge). Before the application is instantiated, the application package must be onboarded to the system. This process is triggered as a result of the application instantiation request from the OSS and is enabled by the MEO at the MEC system or, by contrast, the NFVO at the cloud. When the MEO gets the edge placement decision response, it forwards the request to the target MEC Platform Manager to initiate the process. Besides the application requirements, the descriptor received by the MEC Platform Manager contains the traffic rules to be set on the target MEC platform. After this, the application is ready to be deployed, and to this end, the MEC Platform Manager interacts with the VIM to proceed with the deployment.

In the second case, the process follows a similar workflow, being the NFVO at the cloud the entity receiving the request from the MTO. Consequently, the NFVO will proceed with the application package onboarding and perform the resource allocation in the VIM. Once this is done, the application is ready to be instantiated by the NFVO.

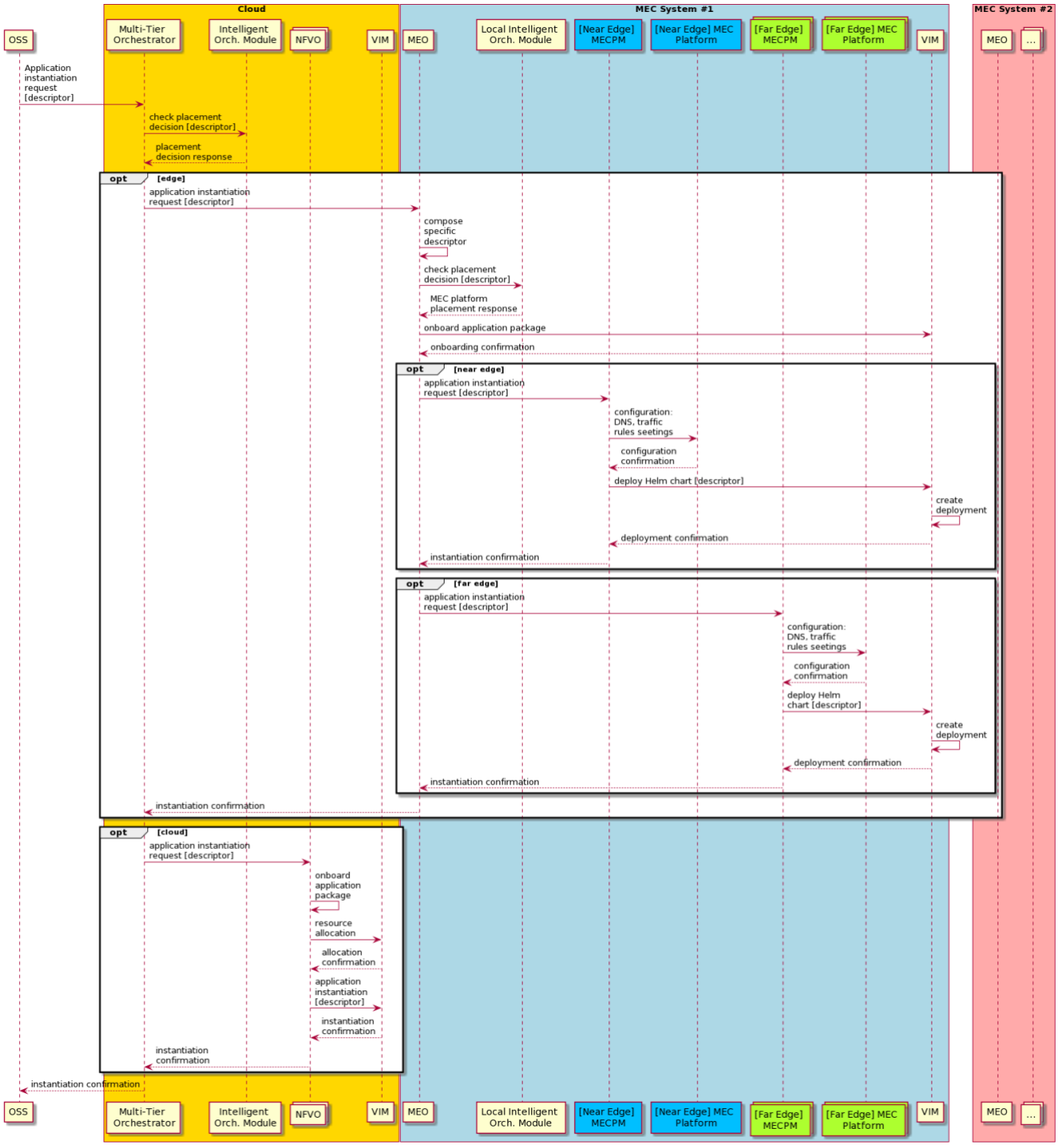

Figure 13 Workflow showing the application instantiation process.

4.2.2.2. Application migration

This section details the workflows considered for application migration processes. More in particular, two main cases are distinguished: (i) situations in which the application can be migrated to another MEC host within the same MEC system; and (ii) scenarios where no suitable candidate is found within the MEC system, and therefore the application is migrated to a MEC host belonging to another MEC system.

Figure 14 depicts the workflow of the first process when the application is migrated within the MEC system, where the MEC host located at both near and far edges share the same UPF. Therefore, in this scenario, a UPF reselection is not involved. By contrast, this workflow is triggered when the resources of the MEC host where the application is currently deployed are not sufficient to satisfy the requirements of the running applications (but they are available within the MEC system) or when the UE trajectory causes excessive delay (but still remains within the same tracking area). In the figure showcasing this scenario, it is assumed that initially the application is deployed at the near edge MEC platform. In this case, the MEO requests to the intelligent module the availability of a MEC platform that satisfies the current needs. Upon a positive response, the MEC orchestrator will start with the specific MEC platform (in this example, at one of the far edges) the same procedure followed for application deployment in the previous section. Consequently, the application will be onboarded in the VIM, and the procedure for instantiation will be carried out, including the DNS and traffic rules configuration to ensure that the UE traffic is properly routed. After receiving the confirmation from the target MEC Platform Manager, the MEO can request the original MEC Platform Manager to terminate the application, remove the traffic rules and free the allocated resources at the VIM.

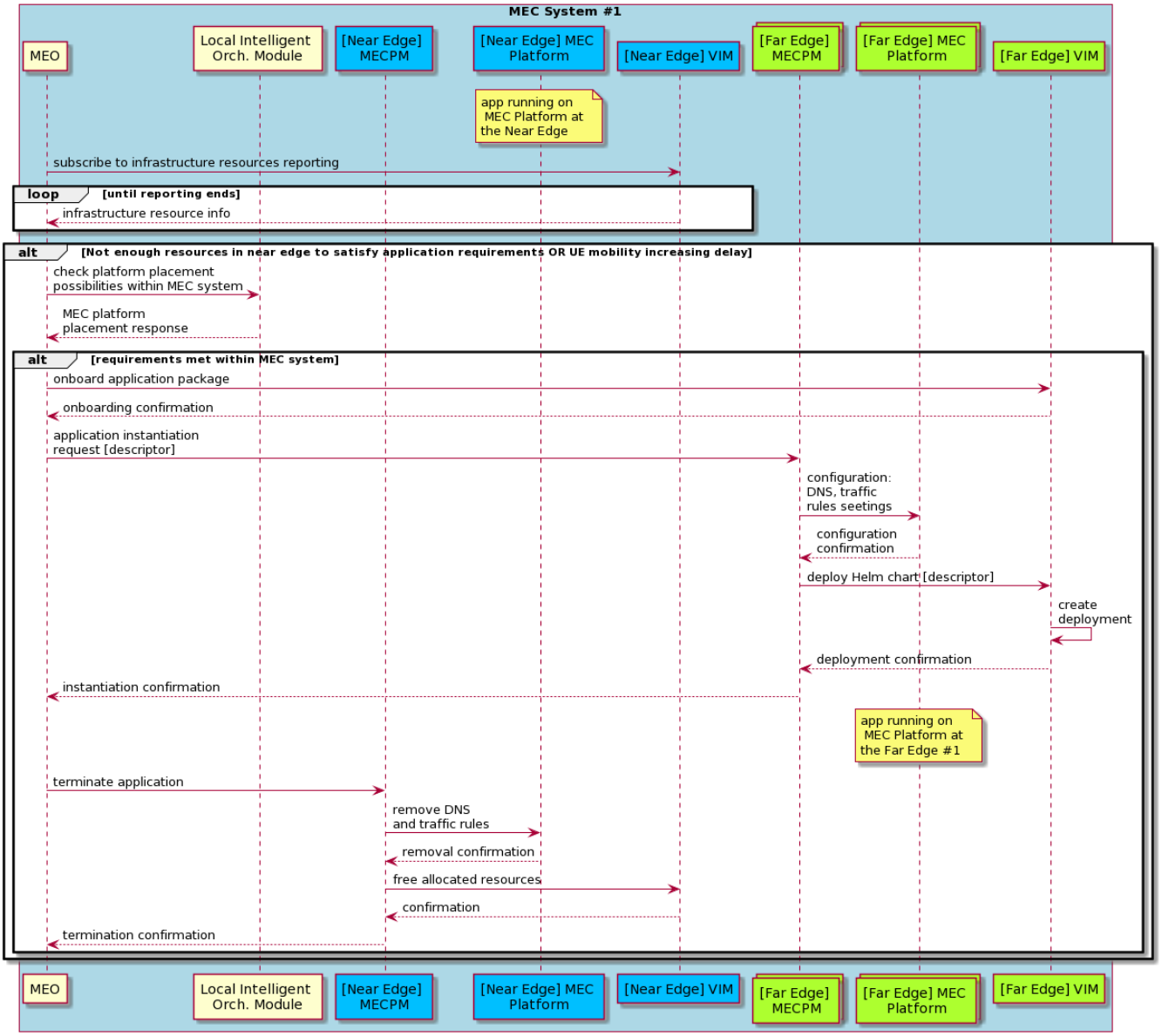

Figure 14 Workflow showing the application migration process within the same MEC system.

Figure 15 showcases the second migration case envisioned, in which the MEC system where the application is currently deployed is not able to satisfy its requirements, or the UE is moving out of the area covered by the current MEC system. In the specific example depicted in the figure, the application is deployed at the near edge MEC platform of MEC system #1 and is migrated to the near edge MEC platform of MEC system #2 when the mobility events indicate the movement out of the area of the UE. To do so, the MEO at the MEC system #1 must survey first the MEO in MEC system #2 to make sure that the migration can be performed and that enough resources are available through the MEC-System-to-MEC-System interface. Upon receiving this request, the target MEO will check the placement decision with its local intelligent orchestration module and reply accordingly to the source MEO. In case of a positive response, the target MEO will start the onboarding using the provided descriptor and will take request the instantiation and

configuration operations locally on the target MEC platform, together with the establishment of the new traffic rules. After receiving the confirmation, the source MEO will proceed with the termination process and will free the reserved resources at the VIM. Notice that the movement of the UE may also involve the reselection of the UPF by the SMF, and the corresponding adjustments of the traffic and DNS rules on the target host.

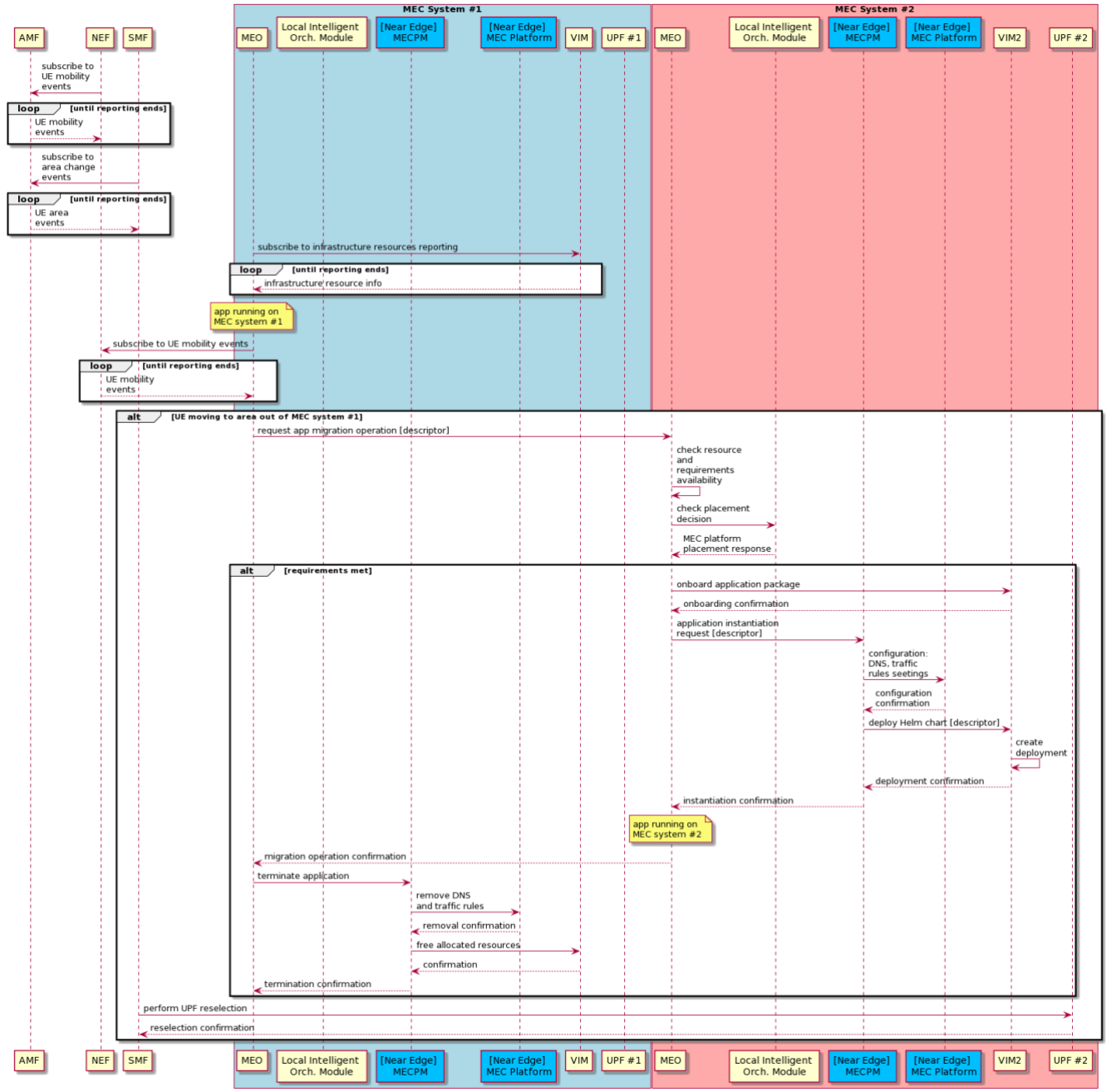

Figure 15 Workflow showing the application migration process across MEC systems.

4.2.3 *RIC-related workflows*

In this subsection we will describe two envisioned workflows involving the A1 interface between non- and Near-RT RICs. The presented workflows have been created according to AI@EDGE architecture and scope, but are compliant with O-RAN Uses Cases and requirements specified in [13]. In the following figures, the interactions marked in red denote that we will make use of O-RAN interfaces, endpoints and definitions (e.g., A1 interface for RIC communication, but also E2 and O1 for other procedures involving the RAN nodes and the SMO). On the other hand, interactions marked in blue denote procedures that are part of the workflows and might be demonstrated within AI@EDGE, but whose concrete definition is out of the scope of this deliverable (e.g., adaptations or simplifications of O-RAN methods, or procedures still not specified by O-RAN). Nevertheless, future deliverables D2.3 and D4.2 might include further details on them.

Note that in the workflows we define the rAPPs as the entities that interact with the non-RT RIC and the SMO to trigger the creation, update and deletion of policies, by means of the functionalities exposed by O-RAN's R1 interface. However, since rAPPs can be seen as part of the non-RT RIC, we could also consider the non-RT RIC as the entry point of the NSAP for these workflows, which could be triggered by an external request.

In a similar way, the xAPPs are the entities which interact with E2 nodes via the near-RT RIC according to their built-in functionalities and the performance data obtained from the RAN. They are linked to one or several policy types, enabling the creation of policy instances via the non-RT RIC to manage and optimize their operations. xAPPs are onboarded on demand by the SMO (e.g., via the O2 interface); however, other interactions could be possible (e.g., external onboarding of the xAPP via proprietary near-RT RIC APIs).

4.2.3.1. Policy Creation

Under Policy Management, we group the following procedures, which are illustrated in Figure 16:

- **Policy type management:** Before creating the policy, the rAPPs need to get the policy types available at the Near-RT RIC (A1-P). According to the existent policy types, the rAPP will create a new type (interface to be defined) or get the schema of an existent type (A1-P) to support the creation of the new policy. The schema of the policy types will follow O-RAN's polity type object specification [11].
- **Policy creation:** Once the policy type is available, the rAPP will trigger the creation of a new policy of this type through the Create Policy method of the A1-P interface. The Near-RT RIC will link this new policy with an existent xAPP, starting a control-loop operation that will involve applying the policy to the RAN resources located at Near and Far Edges according to the policy statements, and monitoring the impact of this policy on the performance according to the policy scope. These procedures will be performed via the E2 interface.

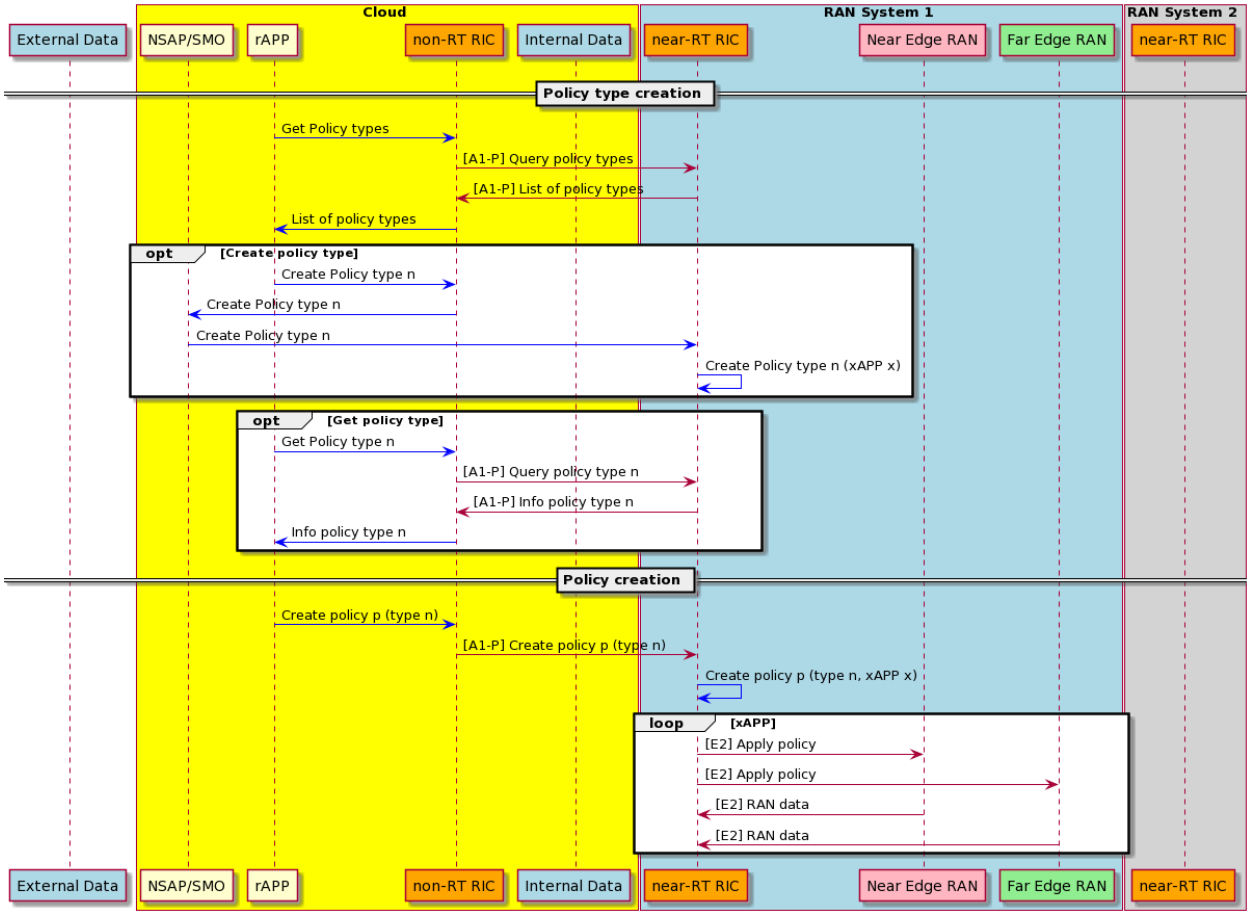

Figure 16 RIC-related workflows: Policy management.

4.2.3.2. Enrichment Information and Policy Update

Figure 17 shows the workflow envisioned for providing Enrichment Information and updating the active policies, which entails the following procedures:

- **EI type and jobs management:** The Near-RT RIC will query the Non-RT RIC for discovering the available Enrichment Information being exposed by the Non-RT RIC (A1-EI). EI Type Objects will define the type of data and how to request its delivery. According to this, the Near-RT RIC will request the creation of an EI job in the Non-RT RIC to deliver this EI (A1-EI). Since the delivery of EI will be based on the subscribe-notify paradigm [10], the creation operation will be basically a subscription to a specific EI type.

- **EI data obtention:** EI can be formed by internal data (e.g., aggregated data from E2 nodes, from available Near-RT RICs or from other components managed by the SMO), but also by external data (e.g., data from the 5G Core or from application servers). This data will be made available to the Non-RT RIC to enable its delivery to the Near-RT RIC via the EI jobs.

- **EI job delivery and application:** According to the aforementioned subscribe-notify operation, the Non-RT RIC will use push-based delivery to send to the Near-RT RIC the results of the EI job (A1-EI). Results can be delivered in a single push or repeated with regular intervals or irregularly based

on events. According to the delivered EI and the enforced policy, the Near-RT RIC might decide to modify RAN resources by means of its xAPPs.

- **Update policy:** rAPPs could also retrieve EI from the Non-RT RIC in order to take decisions on the deployed policies. This will involve the utilization of the Update Policy method (A1-P) and the appliance of this updated policy by the Near-RT RIC, leading to a control-loop automation at the Non-RT RIC level.
- **Policy deletion:** Alternatively, the rAPP can decide that the policy is no longer needed or valid, and delete it via the A1-P interface.

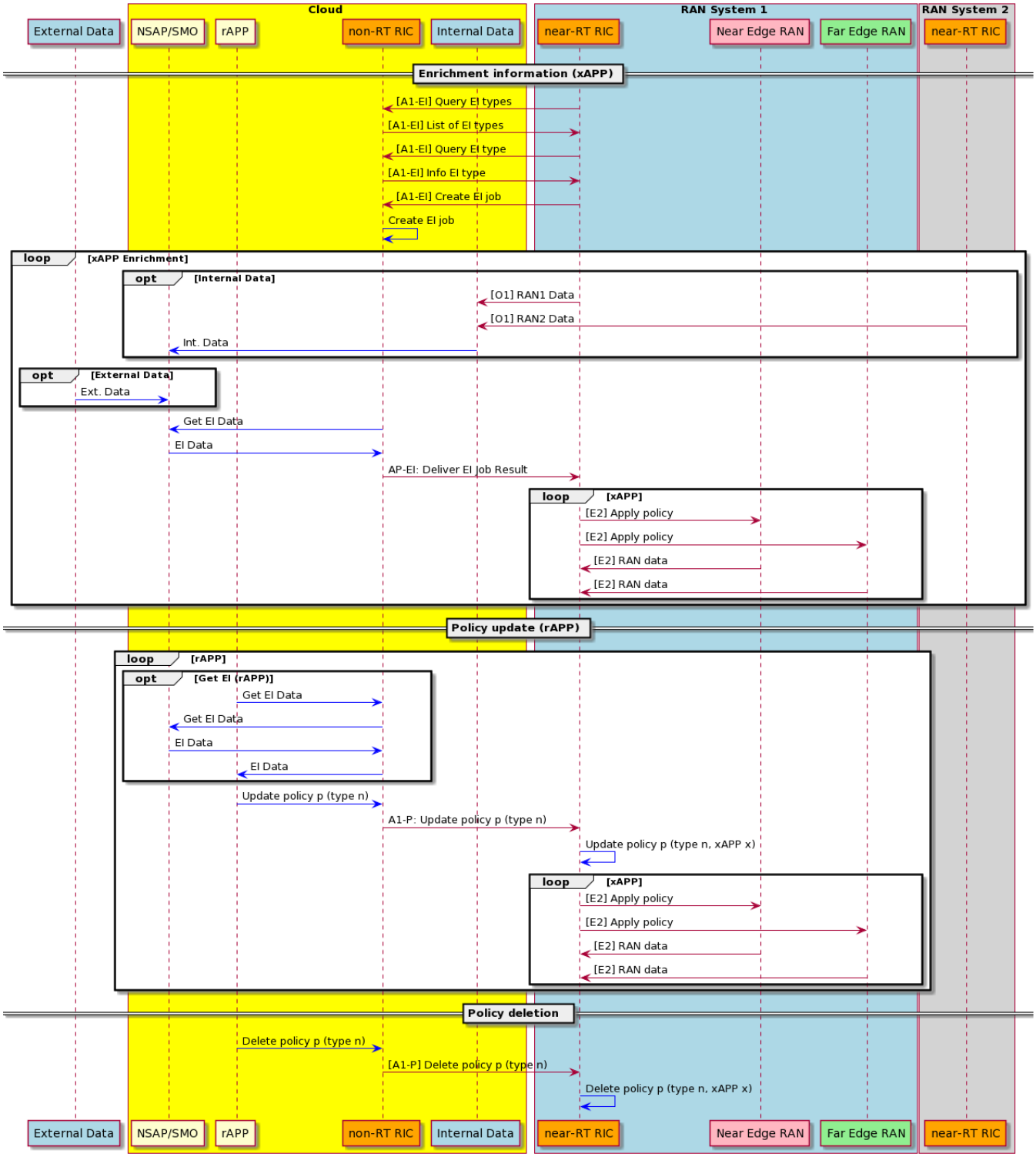

Figure 17 RIC-related workflows: EI and policy update.

4.2.4 *RAN Controller flow*

The communication between the RAN Controller and the e/gNode is carried out, in the initial phase (i.e. for the 4G/5G NSA deployments), over the 5G-OpenEmPOWER1 communication protocol, which is the native protocol used between the 5G-EmPOWER Operating System and the 5G-EmPOWER agent running at eNB that allows the 5G-EmPower controller to manage and monitor RAN elements remotely. OpenEmpower shares with O-RAN E2 the analogous design principles and functionalities, even if it is a custom protocol. Figure 18 shows the 5G-EmPOWER workflows envisioned to support RAN telemetry exposure for specific needs, e.g., monitoring, anomalies detection and predictions. As described in Deliverable 4.1 in detail, an eNB integrated with the 5G-EmPOWER agent gains the ability to interact with 5G-EmPOWER Controller (Operating System). At the moment, 5G-EmPower supports boths OpenWRT based Wi-Fi APs and srsLTE eNBs. In the scope of AI@EDGE, the RAN telemetry exposure will be used to support multi-connectivity solutions outlined in deliverable D4.1. OpenEmpower communication protocol is built around the following three major types of events:

- **Single events**: These are single standalone events, which include an initial handshake between 5G-EmPOWER Operating System with 5Gempower Agent, as well as enabling certain RAN capabilities requests. The 5G-EmPOWER Operating system decides when to initiate/schedule the next event.
- **Scheduled events**: the 5G-EmPOWER Operating System requests the eNB and the UE measurements reports. Once gathered all the requested information, it enforces the RAN slicing policy according to the capability of eNB. Once initiated, the 5G-EmPOWER agent sends these reports periodically to the 5G-EmPOWER Operating system.
- **Triggered Events**: These events are initiated by the 5gEmPOWER agent when a certain condition is detected. 5G-EmPOWER agent supports trigger messages for the UE activation and deactivation. Once a UE joins the network, the 5G-EmPOWER agent sends a trigger message to the 5G-EmPOWER Operating system, which requests for the UE report and, upon the report reception, the UE is added to the list of active UEs.

[1] https://5g-empower.io/

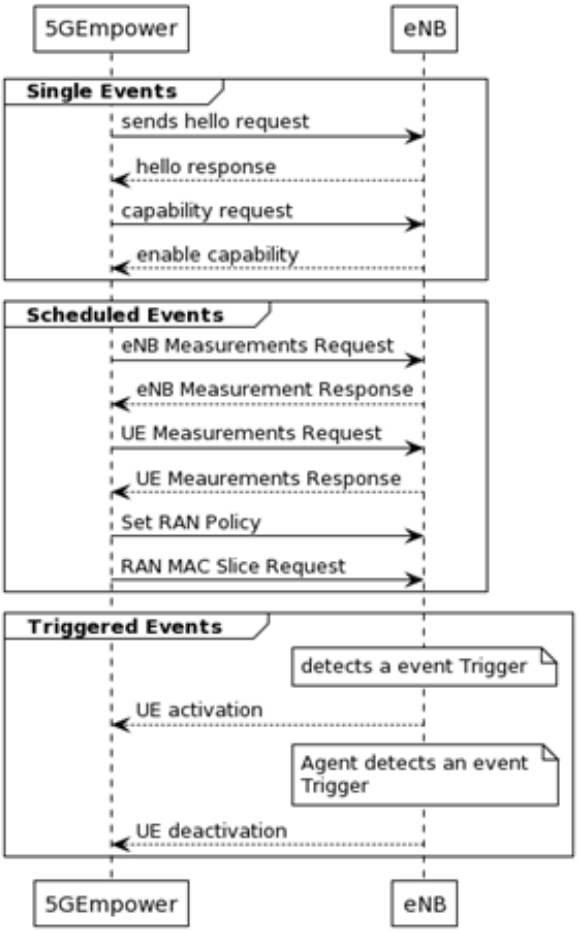

Figure 18 5G-EmPOWER Workflows.

5 Preliminary techno-economic analysis

The 5G sales are picking up worldwide, fueled also by the need to work remotely due to the COVID-19 pandemic. It is notable that the market is driven by mobile broadband use case, so essentially the same functionality as in 4G. While the new functionalities brought in by 5G dedicated to new industries and verticals is not yet widely present in the market worldwide, questions can be raised about how future applications of edge computing will come about. Members of the project have reached out different partners active in other industries or verticals using different contact networks. This has further enriched the discussion about technical specifications in connection with the AI@EDGE architecture, since these partners are typically outside the telecommunication networks sector. We identified and grouped the following categories:

- E-health, with a focus on the demographic challenge and data-driven self-care for elderly chronically multi-diseased.
- Transportation, centering on a future rural autonomous electric transportation system to address environmental and social sustainability.
- European food supply, addressing both farming on land and in archipelagos.
- Europe's electric power system, discussing the trend towards a power grid with more heterogeneous local power generation and the stability challenges this incurs.
- European indigenous minorities, in particular the needs of the nomadic Sami people.

In the case descriptions below, we report on broader insights drawn from discussing each case even without drawing detailed technical conclusions. 5G edge computing needs to come in the form of a global mass-market platform that meets all, or at least most, of the societal needs. The new edge computing paradigm should not be specific to each use case, but rather each use case will push for and request the new capabilities edge computing should possess. Finally, Section 5.6 introduces the main drivers of the techno-economic analysis that will be detailed in future stages of the AI@EDGE project.

5.1 Dialogues about e-health

The demographic challenge was, until the COVID-19 pandemic, a predominant theme in discussions about the future of healthcare. The effects of an ageing population, the ever-increasing number of chronically multi-diseased elderly, threatened the very survival of the system and the security of many patients. The demographic challenge had become a daily challenge and locally the staggering prospects lead to increased staff turnover.

The envisioned relevant component in a solution seems to be self-care. This self-care would then be based on digital sensors connected to the Internet to facilitate individualized data analysis, something that could be described as a personal digital doctor or health assistant. The devices are often simple as a digital scale to monitor the weight in case of congestive cardiac failure (decompensatio cordis), blood sugar sensors for diabetes, blood pressure monitors, etc. What they have in common is that they are affordable and produce data specific to the individual patient. Thus, a data-driven technology-supported self-care will likely be a central part of future health care. It is often stressed that this would significantly increase the patients' quality of life, their empowerment and enable more active lifestyles.

The impact on edge computing is profound. Many argue that this will become the most important use case. The low latency of edge computing, which often is suggested as one of its most important qualities, plays

some role here, but not as much as in other use cases. Latency is mostly mentioned in connection with exoskeletons that could help people to stand up after having fallen down, or even prevent them from falling down. These exoskeletons are not like "Iron man"-style military battle suits, rather for example thin shorts that increase the wearer's leg strength by a few tens of per cent. One of the common reasons for leaving one's home to live in a public care facility is the inability to stand up after falling and the hope is that these devices would enable many to live in their own homes a few years longer. This would increase the quality of life of the patients and reduce costs for the public healthcare systems.

However, the most important quality of edge computing when it comes to e-health seems to be the locality itself. The various national legal frameworks for health data are considered complex and restrictive, with a great focus on preserving data integrity, and thus largely preventing cloud computing solutions. The learning part of any AI solution may possibly reside in the cloud but any inference needs to be local at the edge or in a user device.

The other main quality that is voiced is the ability to reduce battery consumption in wearables. Personalized data-driven e-health solutions will partly rely on wearable sensors and devices and these must not run out of battery. Edge computing offers the possibility to off-load wearables and move computing to the edge. This could be done flexibly depending on the state of the wearables and the edge network infrastructure.

5.2 Dialogues about the future of transportation

Transportation is another area on the brink of a large transformation. Electrification, digitization and automation point towards a future transportation system with autonomous electric vehicles that could reduce the cost of transportation and its environmental footprint to a fraction of the present system. An interesting side note is that this transportation system will probably be deployed and developed in the rural areas before the cities. Today's algorithms can handle the simpler road network and sparse traffic in the rural areas but are still not able to drive in the big cities.

The availability of affordable and sustainable transportation services is crucial for a society to function and a lack thereof often determines a person's quality of life and steers everyday life.

The availability of transportation services also affects economic growth and many economic activities. In a dialogue with owners of small rural businesses, one expressed, with notable frustration, that he had an unavoidable environmental footprint on their mechanical products because no fossil-free transportation was available. Another stated that they could not grow their business and export berries, fish and meat because there was no chain of temperature-controlled transportation services from their rural area. During a conversation with the National Veterinary Institute (SVA), one of Sweden's state agencies responsible for the national food supply and crisis management expressed their deep concern about the "urbanization of cows". Evidently, dairy farms are becoming concentrated around the main highways instead of being located where there are farmers and pastures. This has a negative effect on food production and land usage. These and many other examples illustrate the key role transportation plays in an economy.

Edge computing is seen as a crucial enabling component in a future autonomous transportation system. The vehicles, that includes both vehicles on roads and drones, will be both autonomous and remote-controlled as they will require human assistance to resolve difficult or unforeseen situations. The vehicles are number-crunching platforms and edge computing allows flexibility in distributing the calculations. This can save battery, which is particularly important for drones and for road vehicles when going into "sleep modes". However, flexibility is most important for safety as it gives an additional level of robustness against hardware or software failures. Here the low latency, high reliability, and redundancy of the edge compute fabric becomes essential. An autonomous transportation system does not only include fleets of vehicles but

encompasses a lot more infrastructure ranging from automatic charging stations, traffic lights, and remote-control rooms, to logistics centres that can redirect transportation capacity to handle national emergencies. The edge compute fabric will be central to this complex flow of information and data analysis. Edge computing is again seen as a crucial enabling technology.

5.3 Dialogues about European future food supply

Many of the forward-looking members of the food and farming community seem to agree that European large-scale farming, which provides us with the bulk of food at a low cost, is coming to an end. Soil depletion and climate change have set an agricultural doomsday clock and pollination puts additional stress on the system. It is also stressed that we have very little time to act if we are to achieve an agricultural paradigm shift or at least try to moderate the impact.

In the dialogues the project has had, two main lines of thinking and acting seem to solidify. One is that farming on land could be steered towards precision farming, a highly-automated way of farming using fleets of small autonomous mechanical units that tend to crops and weeds. Scale is then reached by deploying massive fleets while the use of pesticides and fertilizers is low per area unit. Technically this is similar to the autonomous transportation system envisioned in Section 5.2 and edge computing could again play a central role. Today's agricultural machinery focuses on very large heavy machines so that a few persons can farm large surfaces. Autonomous machines would break the dependency on an on-board driver. Large fleets of smaller units could be autonomous and remote-controlled by a few persons. The development towards autonomous machines is already well underway in the agricultural sector, but the machinery is still very large.

The other main line of thinking concerning the future of European food supply revolves around farming in coastal waters and archipelagos. Large scale farming of algae and clams could complement the fish farms that have already turned into a large and successful industry. There is speculation and hope that increased automation, again with fleets of small but now aquatic, vessels could reduce cost and show a path towards scaling up. If so, edge computing could become an enabler. When discussing with aquaculture thinkers and enthusiasts, they often stress that any infrastructure investments (e.g., in mobile coverage and edge computing) would also benefit other areas like tourism, shipping and the inclusion of people living on islands.

5.4 Europe's electric power system

Although deeply troubled by the contribution to climate change, strategists in the electric power area seem to have a generally positive outlook on the future, thanks to the green transformation of society paving a wave for green growth while allowing a better quality of life for many.

However, the electrical power system itself will also need to undergo a transformation, and this will have an impact on edge computing. There will be a transformation from a system with few sources generating electricity and a star network for distributing it, to a grid of innumerable sources that provide electricity (in the form of solar panels, home batteries, wind turbines, electric cars, etc.) and an interconnecting mesh-like electrical grid. While maintaining the stability of the electrical grid was fairly straightforward in the old structure, it becomes a significantly more complex task in the new. The sheer number of sources of electricity makes stability challenging as well as the time-varying characteristics of these sources. While the power generators of old plants, like coal-fueled electricity plants, nuclear power plants and hydro-power plants, had a stable output determined by the operator, the power output of many of the new types of sources

are determined by other factors such as the availability of sunshine, wind, and owners controlling their home battery or electric car.

In the old paradigm, the stability of the electric grid could be maintained by regulating a few large power plants. The new paradigm moves responsibility towards smaller plants and thus requires a distributed, coordinated control, a task that is well suited for edge computing. In fact, in a dialogue with the Swedish power company Vattenfall, they describe a newly developed, proprietary control system that, for all practical purposes, is edge computing. It should be noted that Vattenfall has a long tradition of developing technology. For instance, they developed "Internet", i.e., a packet-based, wireline communication system long before the real Internet and used it to control power plants from a thousand kilometers of distance, but never spread their technology outside the company. Nonetheless, they were thrilled about what edge computing could offer them and the electric power industry. The ability to synchronize power plants over long distances seems to be a special priority. Latency, availability, robustness and redundancy are key performance indicators for power control.

5.5 European indigenous minorities – the Sami

The rural populace of Europe in general, and European indigenous minorities in particular, are often low on access to infrastructure taken for granted in urban areas and important for the quality of life and safety. This infrastructure is important for value creation, as we foresee edge computing will become. In this context, we worried that rural inhabitants and minorities would be left out from the value created by edge computing. Through the project's network we thus sought the indigenous Sami people's perspective on mobile coverage, 5G and edge computing. The traditional Sami lifestyle in the northern parts of Scandinavia is nomadic, herding reindeer between winter and summer pastures. The various national laws in Scandinavia gives everybody the right to equal access to public services like healthcare, postal services, etc. In practice, reality often looks quite different as these areas are rural or extremely rural. If we take mobile coverage as an example, the difference between urban and rural becomes striking. The market for mobile broadband never reached as far out as many of the areas where the Sami live and work. In particular, the summer pastures are largely beyond the reach of mobile communication, or, if you so wish, beyond the market's reach. Various government initiatives to make the market forces extend further out have been ineffective. With mobile coverage comes the possibility to call blue-light services for help, inclusion in democratic processes, access to information, access to healthcare, etc. And without mobile coverage, there will be no edge computing either. When edge computing becomes an essential component in, e.g., the healthcare of the future the gap in services, quality-of-life and safety will increase further. European society needs to try harder to include rural areas and minorities in the sphere of digital services that are becoming increasingly important.

There are a number of viable technological solutions to address this problem. Vendors of mobile communications equipment, in Europe Ericsson and Nokia, have long-range, high-power, high-tower base stations often referred to as umbrella cells, that allow for cell sizes that are hundreds of times larger than the typical urban macro-cells. The idea has been around for some time, and, for instance, Ericsson sold the so-called boomer-base in Australia early in this millennium. The modern long-range systems give coverage over more square kilometers per invested Euro than the base stations designed for urban use, which focus on capacity and would thus enable the market to reach further out into the rural areas.

In dialogues about rural coverage, it is often pointed out that these long-range umbrella cells could be complemented with local hot-spot solutions that would then be back-hauled by the long-range base. This would enable low-power devices such as 5G IoT sensors to connect. Edge computing could then be installed both at the long-range base and in the local hot-spot. To exemplify, we include a photo of one of the most

extremely rural hot spots in Scandinavia, covering a Sami summer village and an important hiking trail (Kungsleden). The hot spot was installed as a part of Sweden's VINNOVA #fulltäckning project.

Satellites, especially low-orbit satellites, are also a promising technology when it comes to providing cellular coverage and mobile broadband in rural areas. However, things do not look as promising concerning edge computing. Although handover of a connection is doable from one satellite to another, moving edge computing tasks could be difficult because of the bandwidth it would consume. At least for now, mobile network infrastructure seems to be the safe bet for edge computing.

Figure 19 A rural hot-spot giving 2G and 4G coverage. Power comes from solar panels and fuel cells. Edge computing could be installed here. Photo: Mats Jonsson, the #fulltäckning project.

5.6 Drivers for Techno-economic Analysis

The goal of AI@EDGE is to create new and realistic opportunities for generating competitive advantages for the European ICT sector. The vision of innovative and demanding applications and services (like cooperative perception for connected cars, three-dimensional aerial photogrammetry, content curation, and IIoT) is set to transform the telecom industry that will benefit from the same level of agility as what is available today in the IT world: time to market for new innovative services will be significantly improved, and the overall Total Cost of Ownership (TCO) will be reduced.

TCO is the main key driver for a techno-economic analysis. A techno-economic analysis and work examines primary costs, benefits, risks, uncertainties, and timeframes to evaluate the attributes of technologies developed and produced in the project. The economic performance of the solutions will be calculated taking into account a life cycle perspective, which considers initial costs, operational costs, maintenances, substitution, etc.

A TCO model as well as revenue assumptions are used to judge the viability of the business cases. In addition, given the costs associated with different business models, performance-cost trade-offs can be identified, and their impact calculated. Finally, indirect benefits (i.e., non-monetary benefits for direct users or positive effects on the economy or society) should be included in the business case evaluation, especially for public stakeholders.

In order to evaluate the economic viability of the selected scenarios (use cases) a generic TCO is built. The model considers both the Capital Expenditures (CapEx) and the Operational Expenditures (OpEx) as well as overhead costs (e.g., marketing, helpdesk, etc.). Capital Expenditures contribute to the company's fixed infrastructure and are depreciated over time. For an operator, they include the purchase of land and buildings (e.g., to house the personnel), network infrastructure (e.g., IP routers) and software (e.g., the network management system). Note that buying equipment always contributes to CapEx, independent from the fact whether the payment is made in one time or spread over time. Operational Expenditures do not contribute to the infrastructure itself; they represent the cost to keep the company operational and include costs for technical and commercial operations, administration, etc. For an operator, OpEx is mainly constituted of rented and leased infrastructure (land, building, network equipment, fiber) and personnel wages. This classification is illustrated in Figure 20.

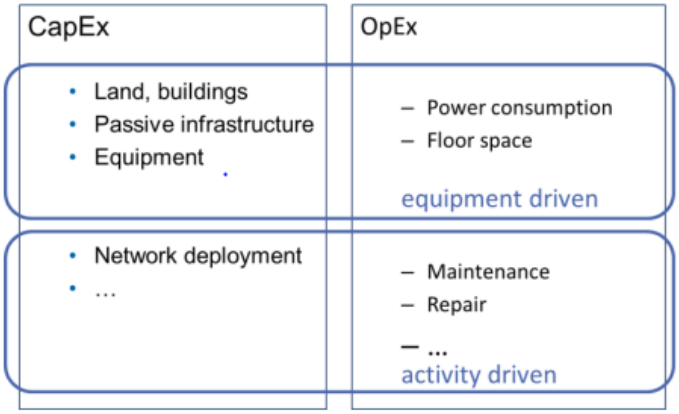

Figure 20 Cost classification.

Specific TCO models will be described and developed for each use case in D2.3 in parallel with the development and testing of the use cases (described in D5.1).

On the other hand, economic restrictions (costs) include: the available budget, the human resources and the expenses. Economic indicators for a techno-economic analysis are: Net Present Value (NPV), Internal Rate of Return (IRR), Return On Investment (ROI) and Dynamic Payback (DP). Moreover, it is important to identify the economic benefits and impacts of the outcomes of the project (for the whole economy).

Risks and uncertainties are related with the achieved values of the KPIs and how they close (or not) are with the target values. A risk mitigation analysis is needed to be carried out based on the findings and work done in the first twelve months. Finally, a timeframe and time plan are needed for the evaluation of the developed technologies.

One of the challenges would be to define a value proposition and identify where edge computing and AI are driving values for specific sectors and businesses. Different business models compared and concluded by a cost-benefit analysis for the most relevant use case should be investigated. The techno-economic

analysis should identify the main reasons why edge computing and AI are playing a vital role in computing and Industrial IoT markets, and analyze emerging architectures and edge platforms where industries need to agree on functions, interfaces, and technologies in order to realize digital products and services.

Environmental performance evaluation will also follow the life cycle approach, accounting for all products and flows through the whole lifetime of the system: equipment production and installation, operation including use, maintenance and replacement, and end of life.

6 Key Performance Indicators

This section summarizes the process of exploring the different components developed under the cloak of the AI@EDGE platform, in order to define a set of Key Performance Indicators (KPIs) within the project. The methodology will make an effort to define a common format, which encapsulates all aspects that are being tackled from a holistic point of view. Hence, to collect the first draft of possible KPI definitions, we followed an iterative procedure which involved getting input from every partner engaged in this process. Of course, the formation of well-defined KPIs is an endeavour that changes dynamically as the project moves on towards its completion. To that end, it is expected that both the format of the KPIs and their content per se will be modified during the whole evolution of the project. In order to ease this dynamic process, a KPI matrix was introduced, which will encapsulate all of the aforementioned characteristics.

6.1 Methodology

As stated, KPIs are by nature prone to change and evolution. But it is important to state a first version of them in order to serve as a baseline. At its core, a KPI is a quantifiable parameter associated with a metric, while it evaluates one critical parameter of a core component of the AI@EDGE project. Each of them should follow a set of properties, depending on the section of the project that they refer to. To achieve such a premise, a number of KPIs have been identified in four basic pillars-domains: Technical, Societal, Economic and Environmental. For the Societal and Environmental domains, a preliminary linkage with the relevant United Nations Sustainable Development Goals (UN SDGs) has been identified. It is important to note that Societal KPIs benefits are related primary to the academic (number of papers produced by the project), gender equality (proportion of women in administrative positions) and healthcare (death rate due to road traffic injuries).

This division serves as a method to capture the KPIs in a holistic manner. To further expand upon this approach, each of those domains can be separated into a set of groups, which will further elaborate on the domain properties. Note that each domain can be found with a different number of groups attached to it, due to the vast variety of subjects covered.

Each group then, shares a number of columns in the table that further define the KPIs involved:

- KPI name and description.
- The use case linked with each KPI (1 to 4) (or if the KPIs applies to all UCs or it is generic).
- The threshold value for each KPI: it expresses the limit we set in a KPI in order for the outcome to be acceptable and/or feasible.

In future revisions of the presented KPIs, we will extend this table with the following columns:

- The target value: value that at the beginning of the project is set as the desired outcome.
- The achieved value: value obtained during the evaluation of the KPIs.

6.2 KPIs Matrix

Bearing in mind the methodology stated in the previous subsection, we present the first version of the AI@EDGE Key Performance Indicators matrix:

Table 1 KPIs matrix (Techical)

Domain [ID]	Group [ID]	KPI Description [ID]	Use Case/ All / Generic	Threshold (Number / Qualitative Description)
Technical [T]	Networking [N]	Vehicle Density [**TN1**]	1	1200 vehicles/km2
		Drone Range [**TN3**]	3	> 20km
		Data Rate/Client for Streaming [**TN4**]	4	> 15 Mbps
		Aggregate In-Cabin Throughput Density [**TN4**]	4	≥ 20 Mbit/s/sqm
	Computing [C]	Latency V2V [**TC1**]	1	< 160 ms
		Latency V2N [**TC1**]	1	≤ 2000 ms
		Control Signal Latency [**TC3**]	3	≤ 50 ms
		Video Processing Latency [**TC3**]	3	≤ 100 ms
	AIF [A]	Robust AIFs [**TA2**]	2	< 5% detection rate decrease against adv. samples
		Fast Detection [**TA2**]	2	Local within 1s, global within 1m
		False Alarm Rate → possibly on AIF group [**TA2**]	2	Rate: < 0.1 %
		Known-Attack Detection → possibly on AIF group [**TA2**]	2	Detection Accuracy ≥ 97 %
	Reliability [R]	Service Deployment Time [**TR4**]	4	A few minutes
		Service Recovery Time [**TR4**]	4	≤ 180 s
		Curated Content Delivery Time [**TR4**]	4	≤ 180 s
		Content Curation Precision of Recommendation [**TR4**]	4	≥ 80 %
		Number Of Served Passengers [**TR4**]	4	12 for demonstration
		Communication Reliability [**TR1**]	1	99.9%
		Control Signal Packet Loss [**TR3**]	3	≤ 1 %

Table 2 KPIs matrix (Societal, Economic and Environmental)

Domain [ID]	Group [ID]	KPI Description [ID]	Use Case/ All / Generic	Threshold (Number / Qualitative Description)
Societal [S]	Academic [A]	Number of Papers produced by the project	All	
	Gender Equality [G] (SDG 5)	Proportion Of Women in Administrative Positions [**SG-A**]	All	
	Healthcare [H] (SDG 3)	Death Rate Due to Road Traffic Injuries [**SH1**]	1	
Economic [Ec]	Budget	Budget Variance	All	
		Return on Assets	All	
	Human Resources	Employee Satisfaction	All	
	Expenses	Payroll Headcount Ratio	All	
Environmental [En]	Atmosphere [A] (SDG 13)	Reduced Carbon Emissions [**EnA-G**]	Generic	X tCO_2e (metric tons of CO_2 equivalent)
		Reduced Ozone Depleting Substances [**EnA-G**]	Generic	
		Penetration of properly equipped vehicles	1	
	Oceans [O] (SDG 14)	Reduced Metal Emissions to Water [**EnO**]	TBC	
		Reduced Organic Pollutants to Water [**EnO**]	TBC	
	Energy [E] (SDG 7)	Renewable Energy Share in The Total Final Energy Consumption [**EnE-A**]	All	
		Proportion Of Population with Primary Reliance on Clean Fuels and Technology [**EnE-G**]	Generic	
	Land [L] (SDG 15)	Reduced metal emissions to land [**EL-G**]	Generic	
		Reduced acid and organic pollutants	Generic	
		Percentage of monitored areas	3	

7 Conclusions and next steps

This deliverable presents the first contributions of tasks T2.2 and T2.3: the intermediate system architecture and interfaces (Milestone MS8), the preliminary techno-economic analysis (Milestone MS2.4) and the draft KPIs (Milestone MS7). As such, it provides a first view on the envisioned architecture, interfaces and workflows to fulfill the technical challenges of the project, and on the techno-economic impact of the developed solutions and use cases.

Regarding the intermediate system architecture, the AI@EDGE project aims to develop and validate a network and service automation platform that leverages AI/ML based closed-loop automation solutions to enable the full potential of Multi-access Edge Computing in multi-tier multi-connectivity scenarios. In this sense, the presented architecture remarks the cross-platform (i.e., NSAP and CCP), cross-system (i.e., MEC and 5G Systems, including Wi-Fi RATs) and cross-tier (i.e., Cloud, Near and Far edge) interactions, which have an impact on the definition of the main components of the architecture and their placement, on the needed interfaces and on the envisioned workflows. The presented contributions at this stage are based on the initial inputs from WP3, WP4 and WP5, which will be further described in deliverables D3.1, D4.1 and D5.1, respectively, and on leveraging state-of-the-art standards, open-source solutions and outputs of other 5G-PPP projects. Future revisions of the architecture will go deep on the interaction between MEC and 5G System through NSAP and CCP platforms to enhance AIFs and Network automation operations, defining for instance the concrete functionalities of the Intelligent Orchestration Component.

Chapter 5 and 6 detailed the methodology that will drive the realization of the techno-economic analysis and the specification of the KPIs of the AI@EDGE project, respectively. At this stage of the project, the main efforts have been focused on the identification of the technologies and the innovations brought by the consortium, and the definition of the testbeds and use cases. Therefore, the preliminary techno-economic analysis presented in this deliverable gives a general overview of the possible impact of the AI@EDGE project in different application areas beyond the scope of the use cases. Future iterations of this analysis will assess the concrete impact of AI@EDGE use cases and innovations on AI, MEC and beyond 5G ecosystems, and on society as a whole. Regarding the KPIs, this deliverable groups and extends the UC-related KPIs introduced in deliverable D2.1, presenting a KPI matrix organized in four basic pillars-domains: Technical, Societal, Economic and Environmental. This matrix will evolve during the project according to the inputs received from the different partners, experimentations and use cases.

References

[1] ETSI GR MEC 017, “Mobile Edge Computing (MEC); Deployment of Mobile Edge Computing in an NFV environment v1.1.1”, February 2018.
[2] O-RAN Alliance, “O-RAN Architecture Description v5.0”, July 2021.
[3] O-RAN Alliance, “O-RAN AI/ML Workflow Description and Requirements v1.03”, July 2021
[4] ETSI GS MEC 00, “Mobile Edge Computing (MEC) Terminology v1.1.1”, March 2016.
[5] 3GPP, “Technical Specification Group Services and System Aspects; System architecture for the 5G System (5GS); Stage 2 (Release 15), 3GPP TS 23.501 V15.12.0,” 2020.
[6] ETSI GS MEC 003, "Multi-access Edge Computing (MEC); Framework and Reference Architecture v2.2.1", Dec. 2020.
[7] M. P. Mena, A. Papageorgiou, L. Ochoa-Aday, S. Siddiqui and G. Baldoni, "Enhancing the performance of 5G slicing operations via multi-tier orchestration," *2020 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, 2020, pp. 131-138, Doi: 10.1109/ICIN48450.2020.9059546.
[8] ETSI GR MEC 035, “Multi-access Edge Computing (MEC); Study on Inter-MEC systems and MEC-Cloud systems coordination v3.1.12”, June 2021.
[9] O-RAN Alliance, “O-RAN A1 interface: General Aspects and Principles v2.03”, July 2021.
[10] O-RAN Alliance, “O-RAN A1 interface: Application Protocol v3.01”, March 2021.
[11] O-RAN Alliance, “O-RAN A1 interface: Type Definitions v2.0”, July 2021.
[12] O-RAN Alliance, “O-RAN Near-Real-time RAN Intelligent Controller Architecture & E2 General Aspects and Principles v1.01”, July 2020.
[13] O-RAN Alliance, “O-RAN Non-RT RIC & A1 Interface: Use Cases and Requirements v4.00”, July 2021.
[14] Andrei Paleyes, Raoul-Gabriel Urma, and N. Lawrence. Challenges in deploying machine learning: a survey of case studies. ArXiv, abs/2011.09926, 2020.