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

D2.3 Consolidated 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

D2.3 Consolidated system architecture, interfaces specifications, and techno-economic analysis	
WP	WP2 – Use cases, requirements analysis, and system design
Responsible partner	Telecom Italia Mobile (TIM)
Version	1
Editor	Jovanka Adzic (TIM)
Authors	Miguel Catalan-Cid (i2CAT), Estefania Coronado Calero (i2CAT), Javier Palomares Torrecilla (i2CAT), Nicola di Pietro (ATH), Daniele Munaretto (ATH), Omar Anser (INRIA), Jonathan Proietto (INRIA), Flávio Brito (EAB), Neiva Linder (EAB), Josue Castaneda (EAB), Zere Ghebretensaé (EAB), Per Ödling (ULUND), Cristina Costa (FBK), Arfan Wahla (FBK), Raman Kazhamiakin (FBK), Javier Renart(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), Jovanka Adzic (TIM), Daniela Di Rienzo (TIM)
Reviewers	Neiva Linder (EAB), Roberto Riggio (UNIVPM), Stefano Secci (CNAM), Arfan Wahla (FBK)
Deliverable Type	R
Dissemination Level	PU
Due date of delivery	31/12/2022
Submission date	01/03/2023

Version History				
Version	**Date**	**Authors**	**Partners**	**Description**
0.1	14/09/2022	Jovanka Adzic	TIM	First ToC starting from D2.2
0.1.2	30/09/2022	Jovanka Adzic, Flavio Britto, ALL	TIM, EAB, ALL	Consolidated ToC, identified responsibilities for sections and reviewers
0.1.3	31/10/2022	Javier Palomares Estefanía Coronado	i2CAT	i2CAT integrated contributions from: 4.2.1: Multi-Tier Orchestrator, 4.3.1: MEC System Components and 5.1: AI@EDGE interfaces.
0.1.4	30/11/2022	Jovanka Adzic, Raman Kazhamiakin, Miguel Catalan-Cid	TIM, FBK, I2CAT	Integrated contributions for Rule-based AI and ML AI Updated AIF concept and AIF Descriptor MEC/Apps Orchestrator
0.1.5	05/12/2022	Miguel Catalan-Cid Nicola di Pietro, Daniele Munaretto Brendan McAuliffe	I2CAT ATHONET, SRS	Integrated contribution for Non-RT RIC interfaces and functionalities 5G Core updates 5G RAN updates
0.1.7	16/12/2022	Omar Anser Flávio Brito Javier Renart George Lentaris Arfan Harder Wahla Nicola di Pietro Raman Kazhamiakin, Flavio Brito Javier Palomares Estefanía Coronado Miguel Catalan-Cid George Avdikos Stelios Koumoutzelis	INRIA EAB ATOS ICCS FBK ATHONET FBK EAB I2CAT 8BELLS	Integrated contributions for AI/ML problems Data Pipeline IOC at NSAP - IOC at CCP IARM Near-RT RIC workfows Slice Manager CCP updates Model Manager workflows MEC Workflows udates Non-RT RIC workflows rApps Consumer/Producer Tecno-economic analysis KPIs

0.2	19/12/2022	Jovanka Adzic	TIM	Additional contributions, preparation for first internal review
0.3	09/01/2023	Jovanka Adzic Daniela Di Rienzo	TIM	Integration of some comments from first internal review. Additional review of chapters: Model Manager Workflows, Tecno-economic analysis and KPI.
0.3.1	20/01/2023	George Avdikos Per Ödling Flavio Brito Josue Castaneda Cisneros	8BELLS ULUND EAB	Integration of Tecno-economic Analysis Integration of Support for Machine Learning and Model Manager Workflows
0.4	24/01/2023	Jovanka Adzic Daniela Di Rienza	TIM	Integration of Model Manager Workflows
0.5	30/01/2023	Jovanka Adzic	TIM	Integration of some comments
0.6	11/02/2023	Jovanka Adzic Neiva Liner Miguel Catalan-Cid	TIM EAB I2CAT	Integratin of comments form second internal review
0.7	22/02/2023	Jovanka Adzic	TIM	Final draft
1	01/03/2023	Irene Facchin	FBK	Final review and submission

Executive Summary

This deliverable details the consolidated AI@EDGE system architecture, describing its components, interfaces and workflows (Milestone 2.6). Together with D3.2 and D4.2, it provides a complete view of the key technical challenges and contributions of AI@EDGE project and defines the scope of the software prototypes used in the trials. This deliverable also reports on the revised techno-economic analysis and (Milestone 2.7) and revised KPIs (Milestones 2.5), introduced in D2.1 and D2.2.

The first part of the deliverable, from chapter 2 to 4, comprises the final contributions of Task 2.2, proposing a functional architecture of the end-to-end AI@EDGE system architecture, its principles, technical assumptions, NSAP (Network and Services Automation Platform) and CCP (Connect Compute Platform) main platforms' components, the description of main communication interfaces and workflows between them. In summary, this deliverable reports the work completed (Milestone 2.6) by Task 2.2, which was 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.
- Analysing the contributions of AI@EDGE project partners defined within the scope of WP3 and WP4.
- Deciding technical choices, fundamental components and trade-offs regarding the system architecture, interfaces, and workflows.
- Designing the consolidated AI@EDGE system architecture by refining the preliminary architecture presented in D2.2 according to the inputs from the analyses and decisions.

In this sense, the achievements reported in this first part of the deliverable report the 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 MS2.6 (*Consolidated System Architecture and Interfaces available*).

The second part of the deliverable, as chapters 5 and 6, provide an update of the work progressing in 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 KPIs (Milestone 2.5), extending the KPIs presented in D2.2 and linking them to the different domains considered within the project: technical, societal, economic, and environmental. In addition, the deliverable revises the techno-economic analysis available in the project (Milestone 2.7), which is based on practical applications of edge computing with the objective to enrich the discussion about technical specifications in connection with the AI@EDGE architecture. This techno-economic analysis considers various categories of services (E-health, Transportation, European food supply, Europe's electric power system and European indigenous minorities), beyond the scope of the use cases of the project representing 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.*). The future revision of the socio-economic impact assessment and the techno-economic analysis, to be provisioned in D2.4, will be more focused on the specifics of each of the use cases. The D2.3 report is finalized by conclusions, giving a perspective, e.g., of how the system architecture functional components, interfaces and workflows are connected to the ongoing work in WP3, WP4 and WP5.

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
AF	Application Function
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
CNN	Convolutional Neural Networks
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

DME	Data Management and Exposure
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
GSMA	Global System for Mobile Communications
HIL	Hardware In the Loop
HITL	Human-in-the-loop
IARM	Intelligent Acceleration Resources Manager
ICS	Information Coordination System
ICT	Information and Communication Technology
IFE	In-Flight Entertainment
IIoT	Industrial Internet of Things
IOC	Intelligent Orchestration Component
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
NEF	Network Exposure Function
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
ONAP	Open Network Automation Platform
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
SME	Service Management and Exposure
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

1 Introduction

The deliverable details the consolidated AI@EDGE system architecture, describing its components, interfaces, and workflows (Milestone 2.6). Together with D3.2 and D4.2, it provides a complete view of the key technical challenges and contributions of AI@EDGE project and defines the scope of the software prototypes used in the trials. This deliverable also reports on the revised techno-economic analysis and (Milestone 2.7) and revised KPIs (Milestone 2.5), introduced in D2.1 and D2.2. AI@EDGE Project Structure and WP2 work dependencies.

Following the AI@EDGE project structure, Figure 1, WP2 has the responsibility to define challenging use cases and first analyse requirements (T2,1 work concluded by M06, report in D2.1), coordinate work with all consortium partners and across WPs towards a consolidated system architecture functional design (T2.2 work M05-M24, reported in D2.1, D2.2 and final report here in D2.3), and continuously work on KPIs definition/revision, socio-economic impact assessment and techno-economic analysis (T2.3 work M05-M36, reported in D2.1, D2.2, D2.3 and final report expected in D2.4 by M36) until the conclusion of the project.

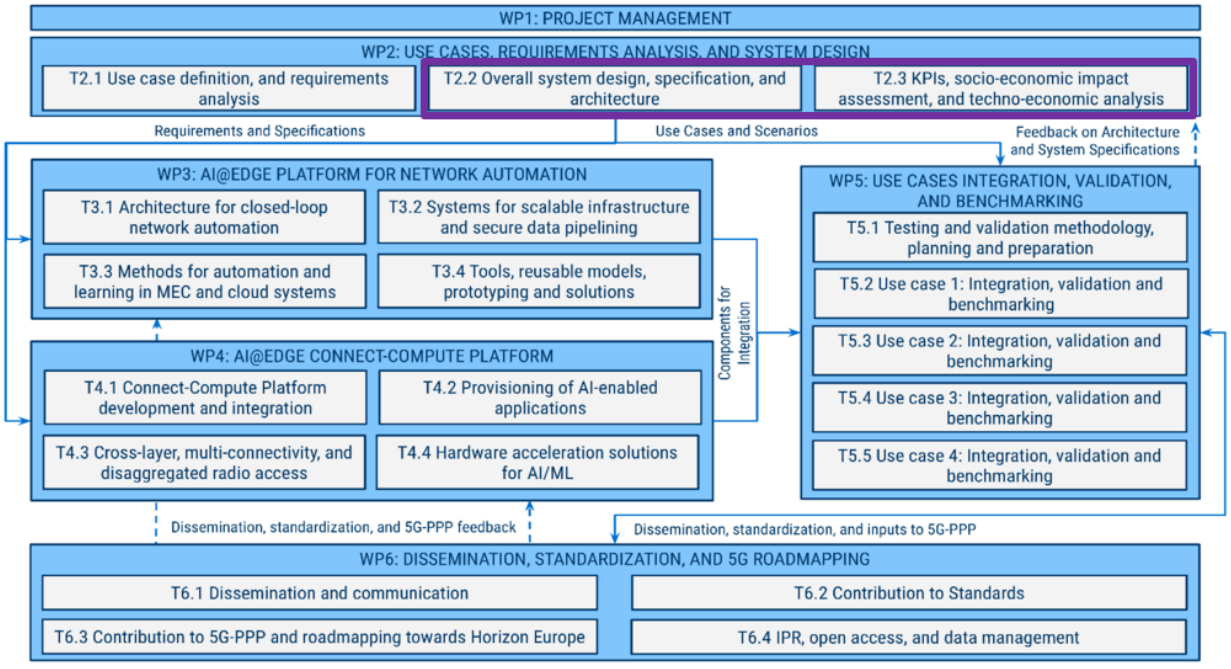


Figure 1.AI@EDGE Project Structure: D2.3 report reflects the work in T2.2 and T2.3

The work presented in this deliverable reflects the conclusion of the work in T2.2 (by M24), regarding the overall end-to-end system design, and the progress of the work carried in T2.3, still to be concluded by M36. From Figure 1, it is worth to highlight that the consolidated AI@EDGE System Architecture described in this D2.3 deliverable is a result of continuous interactions with WP3, WP4 and WP5 in order to reach a consistent functional architecture, definition of main components, interfaces, and workflows.

1.1 Structure of the document

The first part of the deliverable, from chapter 2 to 4, comprises the final contributions of Task 2.2, proposing a functional architecture of the end-to-end AI@EDGE system architecture, its principles, technical assumptions, NSAP (Network and Services Automation Platform) and CCP (Connect Compute Platform) main platforms' components, the description of main communication interfaces and workflows for design between them.

The second part of the deliverable, as chapters 5 and 6, provide an update of the work progressing in 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 KPIs, extending the KPIs presented in D2.2 and linking them to the different domains considered within the project: technical, societal, economic and environmental. In addition, the deliverable revises the techno-economic analysis available in the project, which is based on practical applications of edge computing with the objective to enrich the discussion about technical specifications in connection with the AI@EDGE architecture. This techno-economic analysis considers various categories of services (E-health, Transportation, European food supply, Europe's electric power system and European indigenous minorities), beyond the scope of the use cases of the project.

The D2.3 report is finalized by conclusions and next steps, giving a perspective, e.g., of how the system architecture functional components, interfaces and workflows are connected to the ongoing work in WP3, WP4 and WP5.

2 AI@EDGE System Architecture: Principles and Assumptions

This chapter will provide a description of the AI@EDGE system architecture main principles and assumptions. It will start with the description of AI@EDGE system Architecture wanted position, following with Edge Intelligence and AI general aspects. The progress toward a consolidated functional architecture and standardization brief overview will be also described.

2.1 AI@EDGE System Architecture Wanted Position

The AI@EDGE System Architecture wants to provide capabilities that go beyond the specific requirements of AI@EDGE use cases, thanks to the support for

- *artificial intelligence "in-platform" inclusion (*Network and Service automation intelligence) enabling better usage of infrastructure resources, i.e., high performance converged computing and communication platform;
- *artificial intelligence "on-platform" inclusion (*End-user Application intelligence) enabling better end-user services quality of experience in various application domains (verticals such as automotive), easy and flexible deployment of the third-party applications;
- *data pipeline and data governance framework*, being the data-driven platform, while preserving privacy and security of data in multi-stakeholder environment;
- *E2E overall system orchestration and management*, with orchestration and management for AIFs workflows, seamlessly integrated with network and resource orchestration and management, enabling open, collaborative, distributed ecosystems.

AI@EDGE platform considers 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 observe that AI/ML based closed-loop automation solutions are playing 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 lay the foundation toward providing a fully autonomous zero-touch network and service management platform. Thus, in the remainder of this chapter, different principles and assumptions considered to structure the AI@EDGE system Architecture are discussed.

2.2 AI@EDGE Edge Intelligence related general aspects

The demand for low latency, privacy, security and context-awareness are pushing intelligence to the edge. The project AI@EDGE objectives are exactly that, to bring AI to the Edge of the network, providing the platform with two layers: Connect-Compute Platform (CCP) and Network and Service Automation Platform (NSAP).

The "6G White Paper on Edge Intelligence" [1] provides an overview of the core research questions that will be tackled in the development of 6G edge intelligence. The project AI@EDGE already addressed and provided approaches for some of those identified as core research questions:

- **Edge infrastructure** – The AI@EDGE Connect-Compute Platform (CCP) - distributed connect-compute platform with Standalone (SA) and Non-Standalone (NSA) deployment options, that combines cloud and edge computing, virtualization, 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. CCP aims at facilitating the AI functionality extending the MEC application deployment and orchestration to the AI-related aspects in the life-cycle management process.

- **Data and network management** – AI@EDGE Data Pipeline @NSAP and @CCP level offers a high amount of data with the desired granularity for AI models, training and inferencing, at the Cloud, Near and Far Edge, aiming at being "Native-AI" network platform. The Data Pipeline supports different types of machine learning (supervised, reinforcement, federated) and provides a pipeline for both data and models used by Artificial Intelligence Function (AIF) to train, retrain, and update models according to the AIF descriptor. The components of the pipeline are Data Collector, Data Processor, Model Repository, Data Repository, and Model Manager.

- **Real-time requirements and online learning** – AI@EDGE Data Pipeline provides Model Manager responsible for the monitoring of the evaluation of online model performance, including the detection of the situation where a model needs to be retrained or replaced due to some type of drift (e.g., data drift) or due to periodic/trigger changes according to the AIF descriptor. Model Manager ensures that model training can be quickly adapted along the network. AI@EDGE Data Pipeline considers three domains capable of deploying AIFs with reinforcement learning: NSAP, Near Edge, and Far Edge. For the NSAP and Near Edge two modes of reinforcement learning are possible, namely online and mirror, while Far Edge, since resources are scarce, only the online reinforcement method is supported.

- **Distributed training, algorithmic design, and deployment** – AI@EDGE Data pipeline supports distributed model training and deployment, ensuring that data under the AI@EDGE architecture remains local and private. For federated learning, data is local and only parts of the model are sent to a centralized location. Both the NSAP and the Near Edge domains support federated learning. Furthermore, other learning types such as assisted learning, where data and models remain local, are also supported. Moreover, CCP aims at facilitating the AI functionality extending the MEC application deployment and orchestration to the AI-related aspects in the life-cycle management process, such as appropriate placement of AIFs, due to their specific requirements such as hardware acceleration, or even migration of an AIF from one location (Cloud, Near or Far Edge) to another.

A MEC App/AIFs Edge Orchestrator (MEO) can be deployed as a **Standalone (SA)** and **Non-Standalone (NSA)** module. The NSA mode will be utilized when the MEC orchestrator is part of the full, multi-tier architecture and will act as a second-tier orchestrator. For the scenario where an Edge System is isolated and has no connectivity with the NSAP, the SA Edge Orchestrator will act as the entry point of the system

and be the main manager and controller. This feature allows the architecture to be highly distributed, when operating with the MTO and a decentralized system of autonomous MEOs, when the MTO is not deployed.

The AI@EDGE architecture can adapt to **various types of AI algorithms and models** such as Rule-based AI, ML AI and/or combination of both. Therefore, the AI@EDGE architecture inherently supports AI needs and requirements in terms of technology and business ecosystem to truly enable the exploitation of the artificial intelligence at EDGE with ecosystem that include all key players, such as 5G Vendors, IT Integrators, including providers of End-user Application AIFs.

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.

2.3 AI@EDGE AI related general aspects

In this section we introduce AI@EDGE AI-related general aspects to provide the reader with a brief context on various types of AI, requirements, and how the work in the project addresses the latter.

2.3.1 *Rule-based and Machine Learning Artificial Intelligence (AI)*

To obtain actionable items from data using Artificial Intelligence (AI), one of the following approaches could be adopted: Rule-based AI or Machine Learning AI or the combination of both approaches. It is crucial to identify the best approach and the right "AI solution" for the specific problem domain.

The AI@EDGE platform supports the applications implementing Rule-based AI or ML AI or the combination of both that apply "reasoning" to the environment, i.e., a protocol, a network function, a whole node, a network of nodes and systems. The ability of AI@EDGE platform to efficiently implement Rule-based AI and ML AI applications is key for network and service automation, reactivity, adaptability, and reusability. The AI@EDGE project particularly studied and experimented with ML AI, i.e., the most promising and the most complex types of intelligence, including Federated Learning at the edge of the network.

Rule-based AI

Rule-based AI systems are generally reusable solutions in response to specific conditions, known and already identified problems, that include predefined logics and/or constraints [2]. Rule-based AI systems are probably the most widely used AI systems, often consisting of a set of rules that are used to make inferences. It is mainly based on known and successfully experimented observe-and-react schemes. However, the learning dimension is essential: those schemes need to be continuously verified and possibly improved via the discovery, experimenting and performance evaluation of new behavioral rules. Expected rule's performances indicators could be continuously monitored during operations, i.e., live inferencing.

Rule-based AI solutions are, in various forms, already adopted in existing networks and in particular for automation loops requiring quasi real-time responses and operating on network assets with limited

processing and storage resources. Examples of this type of intelligence are traffic engineering equations, self-healing protection configurations, real-time adaptive bitrate media streaming [3].

Rule-based AI is commonly used in so-called Expert Systems, being one of the attempts to emulate the human intelligence and decision-making process, by inferring from existing knowledge and dealing, better than humans, with vast amounts of complex rules. Rule-based AI generally has a low level of learning abilities involved, but still the equations knobs as well as the input variables can be adjusted, or the thresholds can be re-tuned, based on new findings/experiences. Modern Expert Systems can more easily incorporate new knowledge, i.e., new rules, and thus update themselves. The use of these systems is of huge interest for network and service management, for the automation of complex decision-making processes (e.g., network configuration systems and tools), for the application of complex interpretation rules (for example inferring environmental conditions from sensor data), for prediction (inferring likely consequences of a given situation) or for diagnosis (inferring causes of malfunctions and correlated repair actions).

ML AI

Machine Learning (ML) AI techniques are very sophisticated decision-making algorithms supported by powerful learning capabilities and modeling techniques [4]. This type of intelligences allows to improve the understanding and to continuously gain knowledge about the environment, and hence support the achievement of the goals such as optimized resources/network utilization and agility, being able to detect and respond to real-time changes in the environment [5]. Machine Learning AI applications are based on trained models (supervised or unsupervised or reinforcement) rather than explicitly programmed rules and models. The main component of ML AI process and techniques is a model for the environment, which is created and trained via an appropriate data set. The trained model for the environment is tested on different data sets and evaluated against expected model performances indicators, (reliability, convergence, false negatives, and false positives thresholds, etc.) before deployment in the live network and subject to continuously monitoring during operations, i.e., live inferencing.

Supervised, Unsupervised or Reinforcement ML, described in literature, enable systems to gain knowledge from data without necessarily being explicitly programmed and driven only by already available knowledge. Supervised Learning aims to learn mapping function from given input, labeled training data set, to the output [6]. Unsupervised Learning aims to learn a function that describes a hidden structure, characteristics from unlabeled training data set [7]. Reinforcement Learning (RL) is a goal-oriented learning process, based on learning by interactions with the environment, possibly simulated for some initial steps [8]. The RL agent aims to optimize an objective by interacting with the environment based on a direct trial-and-error process in live network. This is the so-called “learning by doing” or “self-learning”.

2.3.2 *Artificial Intelligence Functions - AIF Conceptual Model*

The view of AI@EDGE is that edge computing applications can leverage on the Connect-Compute fabric to run artificial intelligence algorithms instrumental for the use case applications. The AI logic is therefore meant to be distributed in the network to run distributed AI algorithms. AI@EDGE aims at provisioning AI-enabled applications over a distributed Connect-Compute platform where applications and services are dynamically orchestrated.

In AI@EDGE, we introduce a conceptual model for AIFs capable of capturing and representing in generic terms the common aspects of a component of an AI-enabled application. Figure 2 depicts such a model from the point of view of the interfaces exposed externally. The conceptual model of AIFs could be adopted as a reference model for Rule-based AI and ML AI, being deployed over MEC System as MEC Apps, or over 5G System as rApps/xApps.

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

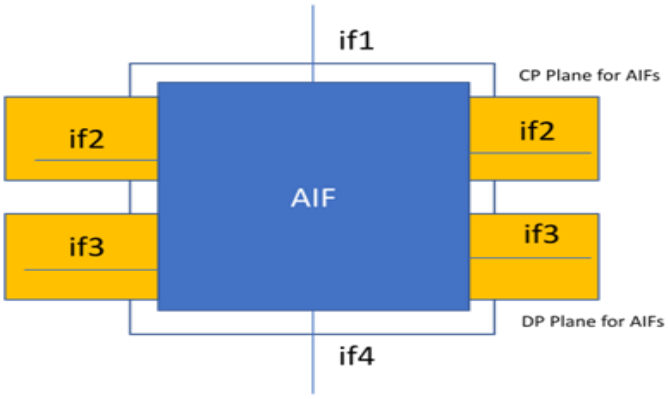


Figure 2 The AIF reference model

The AIF's interfaces are the following:

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

A complex AI-enabled application can be composed starting from the composition and configuration of AIFs including also complex ML AI models and other components that can result in reusable and trustworthy AI/ML pipelines.

The security and privacy aspects related with AI techniques are addressed with specific attention to the distribution of models and sharing of models' parameters in such a way to preserve the confidentiality of the data used during the local training phase. Adversarial ML attacks are used to inject malicious inputs crafted to fool the AIFs. Then, to make AIFs more resilient to such attacks, vulnerabilities in the model are identified and countermeasures devised (e.g., outlier detection filtering).

In the frame of WP3, WP4 and WP5 activities have been undertaken to determine how the four identified AIFs for the four use cases (Traffic Controller AIF, Security AIF, Monitoring AIF, and Content Curation AIF) should be mapped to specific AIF architecture challenges (WP3, Task 3.1) and algorithmic challenges

(WP3, Task 3.3), computing system and networking challenges (WP4, Task 4.2) and use case testbed experimentations (WP5).

2.3.3 *AIF Descriptor*

The above concept of AIF and its characteristics has been captured with the AIF Descriptor specification, represented, and detailed in D4.2. in relation to the AIF orchestration and life-cycle management. Besides addressing the aspects related to the deployment and management of the MEC applications in general, the AIF Descriptor specification aims at targeting the functional and non-functional requirements related to the AI and ML Ops, distribution, data management performance, and HW acceleration. More specifically, the AIF Descriptor specification targets

- HW/SW resource profiles and acceleration to address computation-intensive and AI-intensive workloads;
- Data operated, generated, and processed by the AIFs used to produce the necessary AI models;
- ML Ops life-cycle policies related to model performance, their evolution, replacement, and distribution.

The AIF concept could be elaborated and customized for the xApps/rApps, considering the related ongoing work of O-RAN specifications. One possible approach for xApps containing AIFs could be to add the AIF Descriptor in the "Controls" section of xApps descriptor defined in O-RAN specifications [9]. There are also current discussions on adopting or adapting ONAP's Application Service Descriptor for rApps/xApps.

2.3.4 *AI/ML problems addressed*

The AI@EDGE project investigated and developed several methods, mainly based on ML AI, for automation and learning in MEC and cloud systems, implemented as AIFs, addressing problems arising in the AI@EDGE fabric, and how AI algorithms and ML frameworks such as FL can help in meeting the stringent use case requirements. These methods may directly support the deployment of services in the Connect-Compute platform and/or may provide additional services to improve security or performance. Predictive resource monitoring for service placement predicts resource availability in the platform in order to deploy or migrate services accordingly. Among the currently addressed ML AI problems we can mention the following groups of methods:

- *Intelligent Placement of services at the edge*: MEC (Multi-Access EDGE computing) is a keystone for massive applications and critical services that require high bandwidth and low delays for 6G Edge network applications. With a higher services concentration in small areas and limited resources, it is challenging to develop a system manager to control the optimal location, dynamic resources distribution management and end-to-end service applications maintenance. Based on that, one of the AI@EDGE contributions is the development of ML mechanisms for the optimal and dynamic placement of MECs, and VNFs. We propose a DQN-Based Intelligent Controller at NSAP to deploy virtual function services that are requested among several MECs. For more details, please check D3.2.
- *Application Placement and Active Node Minimization*: we propose a ML, based on network telemetry information, that would suggest the most efficient placement to fulfil the application requirements of storage, CPU and RAM optimally, minimizing the number of active nodes. The ML method uses Reinforcement Learning (RL). If the agent chooses to place an application in a domain that fulfils the requirements, it will receive a high reward value. However, if a placement decision is made that not fulfil the application requirements, the agent is penalized with a negative reward. We are working to

implement this method in the AI@EDGE architecture by putting the ML actors in each MEC orchestrator. The MTO receives the application requests including the CPU, RAM, and storage requirements. Then, the specific MEC system is chosen, and the RL algorithm is responsible to decide the best app placement inside the correspondent MEC. For more details, please check D3.2.

- *Forecasting of measurable performance KPIs*, capturing contextual RAN low-level and network layer data at the edge for user mobility: in this approach, an ML algorithm was developed to predict which base station a user is associated with based on their mobility patterns. The algorithm was trained using data from 12 base stations, collected using the NS3 application. The ML algorithm used was a Multi-Layer Perceptron with 1024 hidden neurons. For more details, please check D3.2.
- *Anomalous Event Detection*: how to detect anomalies is part of AI@EDGE security domain. Being able to identify if anomalous traffic is occurring is important as it is a first step to protect the AI@EDGE platform from attacks. To address this issue, we propose to use multi-layer LSTM neural network to operate on a time-series data. Furthermore, the method uses encoders and decoders to form a deep autoencoders system to train a set of metrics. For more details, please check D3.2.
- *Support for AI-enabled applications*: this group of methods provides advanced support for AI-enabled applications, for example, detection of anomalous events, methods involving federated learning, distributed and collaborative service placement, and data augmentation to increase robustness. Data augmentation: ML-based network traffic classification has been widely studied and has become the most widely used method. One of the characteristics of Internet traffic is its naturally uneven distribution. Knowing that ML algorithms generally aim to achieve the highest overall accuracy without considering class imbalance, this leads to a sharp degradation in the performance of existing ML-based systems when faced with imbalance scenarios. One of the solutions to combat this is data augmentation which consists of extending artificially the set of input data for learning in a realistic way. In a nutshell the method is based on three steps: (1) Definition of transformations to be applied on the dataset, (2) Definition of the policy and (3) Validation of the data augmentation policy. For more details, please check D3.2.

A complete set of AI/ML problems addressed by the project, their formal definition, modelling, and resolution approaches will be documented in D3.2.

2.4 AI@EDGE System Architecture progress toward a consolidated functional architecture

The initial AI@EDGE baseline system architecture, Figure 3, was considered as composed of two layers:

- The **AI@EDGE Network and Service Automation Platform (NSAP)** 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 (CCP)** brings distributed computation over the cloud, far edge and near edge. The architecture is aligned with ETSI/MEC reference architecture [11] and 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).

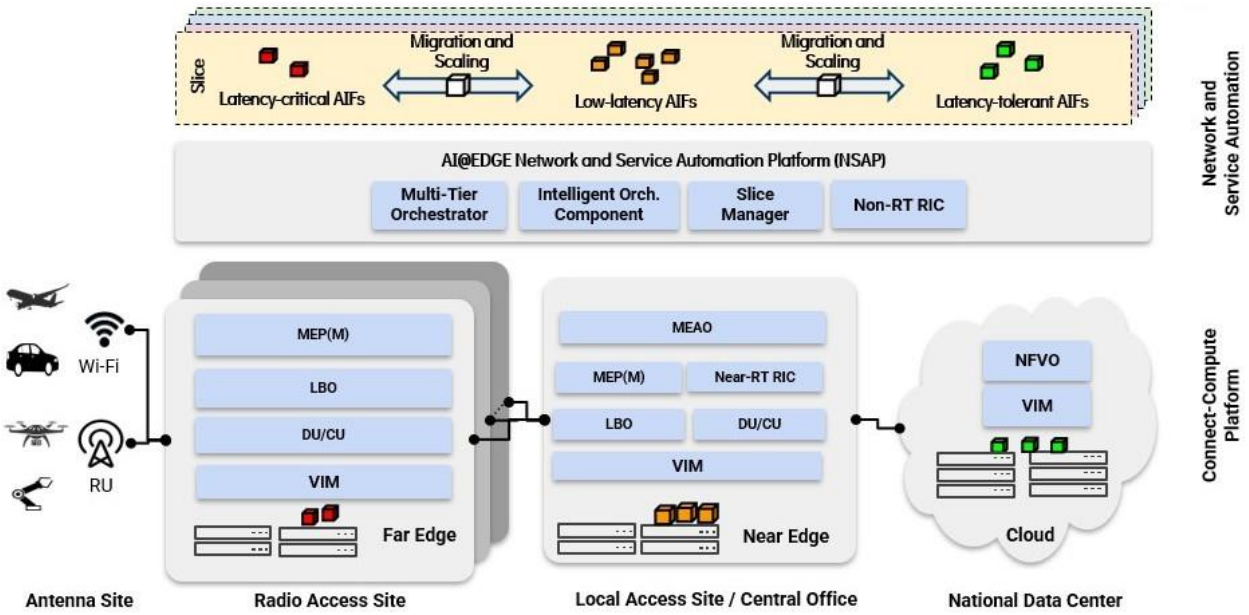


Figure 3 AI@EDGE system architecture

From the beginning, it was important to consider closed loops as an important enabler for automation in the AI@EDGE system architecture. Figure 4 represents the main closed loops considered as part of the system. In this case, there are three types of closed loops that should be enabled 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 4).
- NSAP Closed loop (represented by the blue closed loop in Figure 4). 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 4). 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.

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, 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 into unified multi-tier slices.

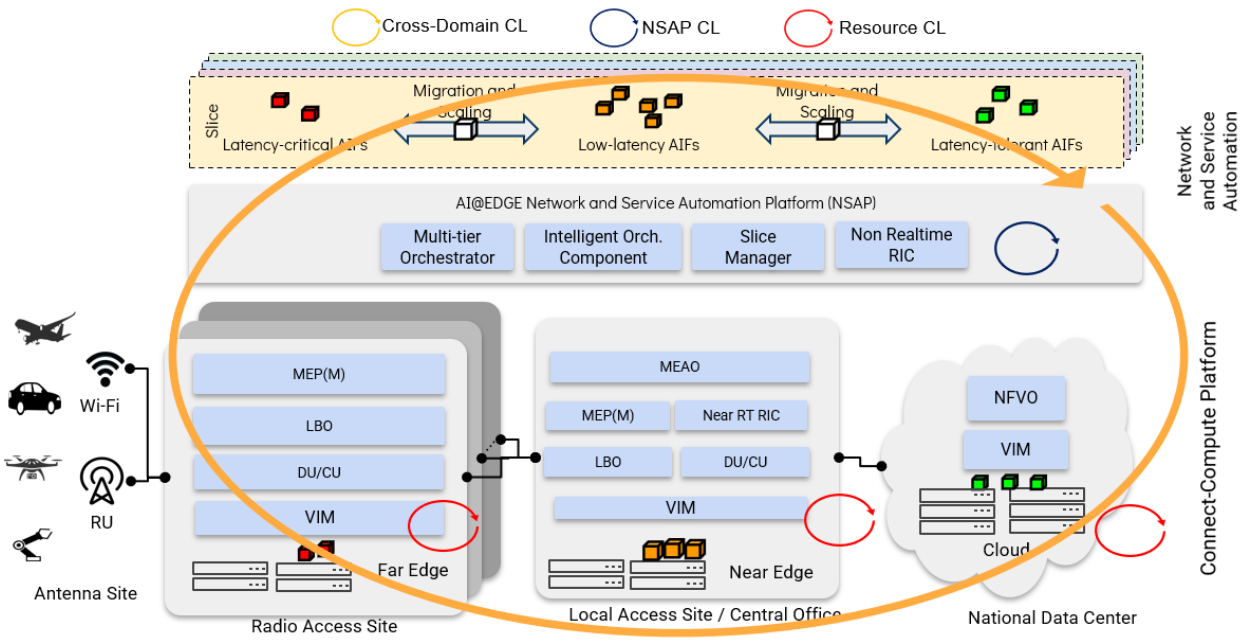


Figure 4 AI@EDGE system architecture including Closed Loops

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 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.

In the Figure 5, the AI@EDGE Consolidated System Architecture is presented, shown in more details components and interfaces, e.g., the data driven architecture and intelligent components in NSAP like the Non-RT RIC, and the AI/ML data pipeline blocks: AI/ML Model Manager, Data Processor and Data Collector, which may or not be implemented as part of the Non-RT RIC implementation solution.

It is understood that the AI/ML Data Pipeline and overall, a data driven infrastructure, is available and distributed over the E2E architecture. This implies that, e.g., data storage requirements set by AIFs or AI/ML algorithms can be addressed by placing components like Data Collectors, Data Processors and Databases in the Near/Far Edge on demand.

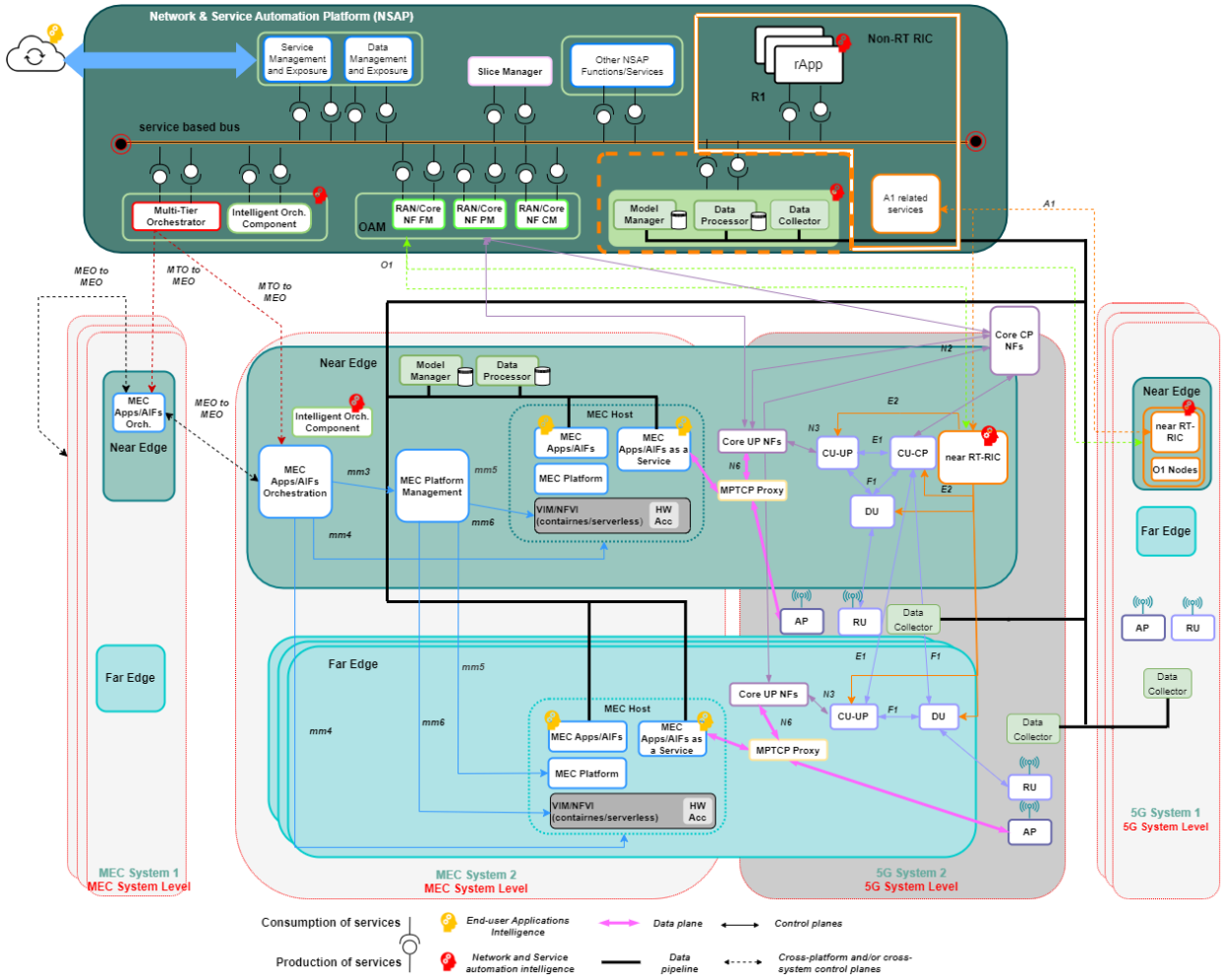


Figure 5 AI@EDGE Consolidated System Architecture

2.5 Autonomous Networks - Standardization Brief Overview

Academia and industry have witnessed new concepts trends influencing and impacting the standardization fora open-source groups and vice-versa. It's no difference when it comes to AI/ML applied "intelligence" modifying the way network architecture components and interfaces are being integrated in existing well proof architectures, or to evolve current architectures toward truly AI native. All of this is motivated by the promise and high expectations, e.g., to bring near zero-touch automation solution to telecom network operations, simplifying processes, lowering complexity of operations, facilitating introductions of new business and services, etc. In this section we briefly indicate some of the standardization bodies that has acknowledge the importance to consider AI/ML applied networks towards a fully autonomous network.

O-RAN

During 2022, O-RAN alliance decide to make it public all completed specifications from now on [12]. The system architecture work in AI@EDGE considered different technical assumptions and trends under discussion in O-RAN and specifically in the O-RAN next generation research group (nGRG). We foresee that the work toward a networks architecture fully supporting AI/ML based network operations and services for more autonomous networks it's still in the beginning for standardization and still many research questions remain to be addressed in upcoming years. The authors in [13] offer a quite comprehensive overview of the current O-RAN architectural aspects and research challenges.

As an example of architecture principle adopted by AI@EDGE, while still being discussed in O-RAN Architecture work (WG1/WG2), is the choice of a service-based architecture for the realization of the AI@EDGE NSAP, which can be considered as an equivalent of the Service Management and Orchestration (SMO) in O-RAN architecture.

3GPP

The 5G System has initiated to provide a network design suitable for analytics services by introducing the Network Data Analytics Function (NWDAF) [14] to the 5G CN to implement network automation and optimize the related network functions, based on collection of data from other 5G network elements to analyze data, possibly with AI models, that can be used by other network services. Nevertheless, 5G only focuses on how to use data analytics and AI to optimize the network itself and the emergence of NWDAF is only the starting point to support AI. The intelligent features could be utilized by the end-user applications, which however is out of the scope of 5G architecture design.

ITU-T - FG-ML5G

In 2018 the International Telecommunication Union (ITU) initialized the work on how to integrate AI framework within the communication systems. The ITU-T Focus Group [15] on Machine Learning for Future Networks provided technical specifications starting from architectural framework for machine learning in future networks including IMT-2020, that introduce the concept of a Machine Learning based pipeline and its functional components with ML Orchestrator [16] ITU-T Y.3172, June 2019.

3 AI@EDGE Consolidated System Architecture: Platform and Components

This chapter will provide a description of the AI@EDGE consolidated system architecture with a focus on its components. It will start with the description of each component of the AI@EDGE Consolidated System Architecture diagram represented in Figure 6. This figure also details the main interfaces among each component of the AI@EDGE architecture which will be described in Chapter 0. Besides the grouping based on NSAP and CCP platforms, Figure 6 also differentiates between MEC System and 5G System at the CCP level, 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 fulfil AI@EDGE objectives.

This chapter is organized as follows: in Section 3.1, the NSAP main blocks are presented: the Multi-Tier Orchestrator, the Intelligent Orchestration 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 Orchestration 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 D3.2, and similarly, read further about the Connect Compute Platform in the deliverable D4.1 and D4.2.

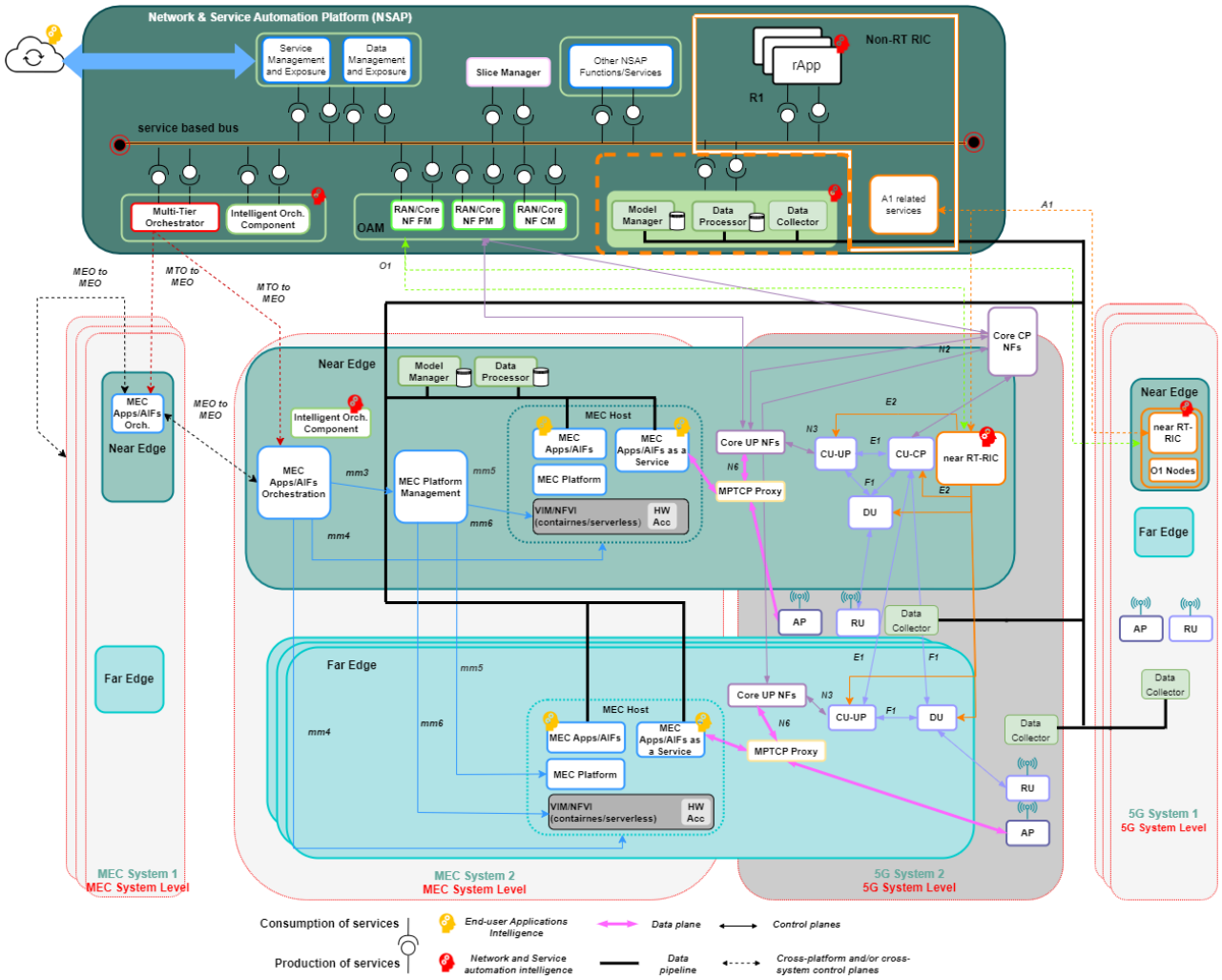


Figure 6 AI@EDGE Consolidated System Architecture.

3.1 *Network and Service Automation Platform (NSAP)*

We group under the NSAP the components 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 Multi-Tier Orchestrator (MTO), together with the Intelligent Orchestration Component (IOC), 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. 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.

In Figure 7 we can see NSAP as a service-based architecture with a common service bus enabling the production and consumption of different services provided by its components.

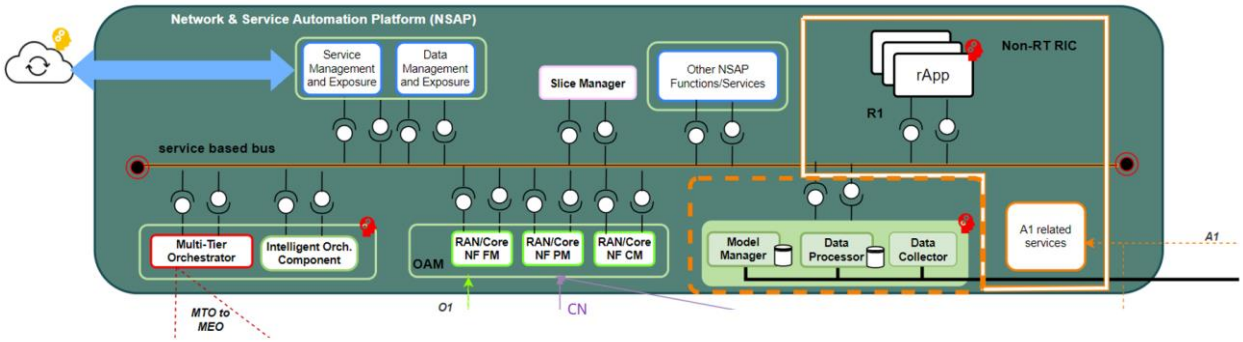


Figure 7 NSAP - Network & Service Automation Platform

The following architectural aspects of the AI@EDGE NSAP platform are to be noted:

- **Service-based architecture (SBA)** is the approach for interfacing components themselves and may be used by the MTO to orchestrate and manage the AIF deployment and management.
- **Data handling reference points** are defined between logical components in a way that any impact to the live network is localized to the source of data, so the extensions of the existing protocols may be used to minimize the impact to the live networks.
- **Domain federation** - the deployment of AIFs may span different domains, e.g., distribute the components across different domains (e.g., Core, RAN and Edge).
- **Third-party integration -** enables the third-party solution providers to integrate with AI@EDGE platform providing AIF application, the complete pipeline or some of the components of the pipelines, such as Model component.

3.1.1 *Multi-Tier Orchestrator*

The Multi-Tier Orchestrator (MTO) is the Network and Service Automation Platform (NSAP) entry point for the operations related to the lifecycle management of MEC AIFs/Apps. The MTO enables communication with different types of orchestrators across multiple southbound clients, such as: (i) the

MEC Orchestrators (MEOs) located at the Near Edge of each MEC System; and (ii) the cloud-based NFV Orchestrators. Depending on the incoming request, the MTO can issue different orchestration operations (e.g., deployment, migration, termination, etc.) to nodes located on the near edge or far edge of the network.

When the MTO receives an instantiation/migration request, the requirements of the specific applications are shared with the Intelligent Orchestration Component (IOC) to seek the most suitable solution to the orchestration request received. Notice that this procedure is possible thanks to the metric aggregation performed at the NSAP, which congregates at the NSAP level the data information and infrastructure telemetry data coming from various MEOs. The infrastructure telemetry data is a sub-functionality of the data pipeline related to infrastructure data and provides an overall view of the status of the various resources available at the various MEC systems. The requirements of the application being managed are specified on the AIF descriptor (described in detail initially in D4.1, and refined in D4.2 document, which includes among others, application id, application name, application provider, the URL pointing to the helm-chart containing the application, specific ML dependencies, and a list of computational requirements (RAM, CPU, disk, hardware acceleration, among others). Moreover, information regarding input sources may be also provided to be used for orchestration purposes, including data dependencies and data source ID. Moreover, information related to possible data locality required, can be also included. The decision of selecting the best candidate for allocating the AIF/app is computed by the IOC, which belongs to the decision-making module of the network automation closed-loop stated in D3.1., and it is described in more detail in the next subsection. Once the decision is provided by the IOC, the MTO proceeds with the deployment (or migration) process to the selected destination communicating to the specific MEC Orchestrator required.

3.1.2 *Intelligent Orchestration Component (IOC)*

The Intelligent Orchestration Component (IOC) will leverage functionalities of fault, security, and resource management for AIFs in runtime. At NSAP level, the IOC will be part of the MTO, providing intelligence to the decisions taken on AIF placement and migration.

The IOC will receive two types of data to take its decision. On one hand, it will know the minimum requirements to run an AIF which will be included in the AIF descriptor. On the other hand, it will receive specific metrics from AIFs, nodes and MEC system, which will be aggregated by the Data pipeline components and will aid the different AI/ML models inside the component take optimum decisions.

For the placement of AIFs which require HW acceleration, the IOC will have access to metrics like CPU, RAM or Disk, as well as the existence of HW acceleration components in specific nodes. This information on HW acceleration will help the IOC decide on a MEC system to deploy the AIF, delegating the final decision on the node to the IOC at CCP level, which will have further information on real availability of HW acceleration components.

Furthermore, the IOC will continuously monitor the different MEC systems, looking for anomalies or faults in specific AIFs and nodes. In case any anomaly or fault is detected in a system, the IOC could decide to migrate certain AIFs to balance the load between different nodes, always ensuring the minimum requirements of the AIFs specified in the descriptor are met. It is expected that the IOC includes several AI/ML models developed in the project to solve placement or runtime issues in AIFs. These models could include Anomaly Detection FL model and AIF placement models which will be further detailed in D3.2.

3.1.3 *Slice Manager*

The Slice Manager provides control over the lifecycle of the slices in the AI@EDGE platform. Starting from a slice template, it can create E2E slice instances and trigger their deployment over multiple MEC Systems and 5G Systems. It enables the Create, Read, Update, and Delete (CRUD) operations over slice instances and it oversees creating a mapping between SLA requirements and a logical slice network. To do so, the Slice Manager interacts with the MTO to allocate the computational resources needed for the slice at MEC Level, and with the Core and RAN controllers to enable the allocation of the needed network resources to a particular slice.

Currently GSMA's NEST defines a generic slicing descriptor from a business perspective for Network as a Service (NaaS) [17] while 3GPP SA5 standard defines NSI and NSSI [18]. These standards consider slicing at RAN, Core and Transport levels. There is currently no standard approach for E2E slicing which includes MEC-level resource reservation or data isolation. Nonetheless, quoting [19] is important to highlight that the 5G core network's User Plane Functions (UPFs) "can be seen as a distributed and configurable data plane from the MEC system perspective" Therefore, at the data plane level, 3GPP-defined network slicing and the consequent resource reservation and isolation are naturally mapped onto the MEC framework. Furthermore, again according to [19], a MEC platform, specific MEC services within it, or even a MEO may all be deployed as 5G Application Functions (AFs). As such, they can request the implementation of specific traffic rules, the enforcement of policies, and the control over PDU sessions, quality of service, or other 5G features, via the interaction with 5G core network functions like the Network Exposure Function (NEF) or the Policy Control Function (PCF) - cf. Section 3.2.2.2. Hence, a slice manager in control of the 5G slice configuration and assignment of 5G core network functions to specific slices, can also control and authorize the interaction of MEC-level functionalities seen as AFs with such network functions by specifying to which slices each AF has access (or not). This can be effectively leveraged to "partition" MEC systems and associate them with different 5G network slices. Then, a slice manager that also controls how computing and storage resources are "sliced" at the edge-cloud-computing level, has in practice the logical capability to govern end-to-end slices that combine 5G network and MEC resources into a joint management framework. Implementation-wise, this requires a certain degree of unification in the representation of 5G and MEC resources internally to the slice manager, such as a unified slice indexing system.

3.1.4 *Non-Real-Time RAN Intelligent Controller*

At the NSAP level, as shown in Figure 6, intelligent network automation related to the 5G System is done by the rAPPs, which are managed by the non-RT RIC, and which have access to the SMO/non-RT RIC functionalities through the R1 interface. The R1 interface, which is still under definition in O-RAN [20], is a service-based interface with Service Management and Exposure (SME) and Data Management and Exposure (DME) capabilities.

In the current architecture of O-RAN's non-RT RIC [21][22][22], Data Management and Exposure services are performed by the Information Coordination System (ICS) element. In AI@EDGE, the non-RT RIC leverages the ICS implementation from O-RAN (release F) to coordinate the data exchange among the rAPPs, which can act as data producers and/or data consumers. This way, producer rAPPs can collect, process and expose internal (i.e., from the RAN) or external data (e.g., from Core network NFs or application servers), while consumer rAPPS can use this data to intelligently orchestrate the RAN through the O-RAN interfaces exposed by the non-RT RIC API (e.g., O1, A1). This model also allows a chaining approach, where consumer rAPPs can produce and expose new data (e.g., AI-driven predictions). Thus,

this can be considered an implementation of the basic subset of R1 interface functionalities needed to create to intelligent closed-loop automations

The implementation of the non-RT RIC and the developed rAPPs in the scope of the multi-connectivity radio access will be further described in deliverable D4.2.

3.1.5 *Data Pipeline*

As stated in previous deliverables, the AI@EDGE Data pipeline offers a high amount of data with the desired granularity for AI models training and inferencing at the Cloud, Near and Far Edge. However, data is only half of the equation for a successful "Native-AI" network platform. The Data Pipeline supports different types of machine learning, and so the AI@EDGE consolidated architecture provides a pipeline for both data and models used by AIFs to train, retrain, and update models according to the AIF descriptor.

The components of the pipeline are Data Collector, Data Processor, Model Repository, Data Repository, and Model Manager, that we briefly describe here. For a more detailed description, readers are welcome to review Section 3 from Deliverable 3.2.

The *Data Collector* receives data from multiple data sources required to train models to be deployed in the AIFs. It also offers authentication functionality to prevent spurious data.

The *Data Processor* cleans, filters, and prepares data to be used during the training for a model. It can also augment data if an AIF descriptor calls for it. Additionally, it can store data in the Data Repository along with meta-data, enabling reusability of data by many models. Moreover, the models can be stored in the Model Repository to be used later or re-deployed afterwards.

The *Model Manager* is responsible for the evaluation of performance for one or multiple models, once a model has been instantiated due to an Inference AIF it is registered to the Model Manager. Some of the responsibilities of the Model Manager include detecting when a model needs to be retrained or replaced by a type of drift (e.g., data drift) or due to periodic/trigger changes according to the AIF descriptor. This information is sent as output of the Inference AIF to the Model Manager. More details about the pipeline and AIF interaction are given in Section 4.2.1.

The Data Pipeline should be able to handle security aspects such as authorized data access and data delivery to the correct user, avoiding data corruption and poisoning. Furthermore, as AI@EDGE Platform supports Federated Learning (FL) environment, the Data Pipeline should be able to ensure a secure environment for FL detecting, for example, model poisoning attack. These are some examples of fields that AI@EDGE can explore in the upcoming deliverables such as D3.2.

3.2 Connect-Compute Platform (CCP)

The AI@EDGE Connect-Compute Platform (CCP) combines Cloud computing and virtualization, 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. CCP aims at facilitating the AI functionality extending the MEC application deployment and orchestration to the AI-related aspects in the life-cycle management process.

More specifically, the management and orchestration of AIFs is determined by the requirements included in the AIF descriptor which, besides more traditional MEC application properties, covers the essential characteristics of the AI and ML operations. This includes, in particular:

- Different deployment profiles suitable for different target resources in order to address computation-intensive and AI-intensive workloads;
- Information about the data operated by AIFs in order to create and update the relevant AI models;
- ML life-cycle policies in order to monitor the performance of the underlying models, their evolution, deployment, and replacement.

Such a definition allows the CCP components to operate in an integrated manner also with the Data Pipeline and with the underlying infrastructure in order to ensure the necessary life-cycle operations. Further details about the AIF descriptor and its semantics are represented in deliverable D4.2.

As was shown in Figure 6, we can group the functions provided by the CCP in MEC System Components and 5G System. MEC System is a collection of MEC hosts and the MEC management necessary to run MEC applications. As per [23] 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 Orchestrator

The MEC Apps/AIFs Orchestrator (MEO) element corresponds to the MEO in the ETSI MEC architecture [23] and maintains an overview of the complete MEC system. It has the following main features and design choices:

- Communication capability in a bi-directional manner with the MTO, located at the NSAP, which manages several types of orchestrators at both the edge and cloud. This communication takes place over the defined MTO-to-MEO interface through a messaging system (broker) to automatically detect a new MEC System that joins the network, to deploy an AIF/app or to migrate it because its requirements are no longer met.
- Incorporation of a file storage system of descriptor files received from the MTO for the onboarding of AIFs and MEC applications. Therefore, the MEO maintains a registry of any AIF or MEC application running in the MEC system.
- Capability to deploy AIFs and MEC applications on specific near-edges and far-edges. In the scenario of a deployment needing to be deployed at a specific location, the orchestrator can receive these requirements and execute the deployment at that location.
- Working capability to be deployed as a Standalone (SA) and non-Standalone (NSA) module. The NSA mode will be utilized when the MEC orchestrator is part of the full architecture and will act as a second-tier orchestrator. For the scenario where a MEC System is isolated and has no connectivity with the

NSAP, the SA MEO will act as the entrance of the system and be the main managing and controlling the network. This feature allows the architecture to be highly distributed, when operating with the MTO, while in the SA mode, it enables multiple entry points to the platform that do not depend on the central NSAP. Additional information on the SA and NSA functionalities of the MEO will be found in deliverable D4.2.

- Distributed and bi-directional communication with other MEOs to migrate AIFs and MEC applications between different MEC systems. This communication occurs through the defined MEO-to-MEO interface through a messaging system (broker) to migrate an application because its requirements are no longer met. The workflow associated with this interface is depicted in Section 4.2.2.2. Moreover, technical details of this interface are provided in more detail in deliverable D4.2.
- Encapsulation of the module logic in a docker image that can be deployed as a container, making it lightweight, scalable, secure, and portable. This gives the MEO the characteristics and benefits of containerization.
- Access to the metric aggregator and data management monitoring system, which records real-time metrics in a time series database that can be accessed by the Orchestrator and the IOC to make decisions for the placement and migration scenarios.

3.2.1.2. Intelligent Orchestration Component (IOC)

The Intelligent Orchestration Component (IOC) is a plug in of the MEC Apps/AIFs orchestration module. AI/ML methods could be applied to provide intelligent orchestration of the MEC Apps and AIFs inside the MEC system [44]. At CCP level, the IOC will guide the corresponding MEC Apps/AIFs orchestration module to scale or migrate at a local level, taking its decisions ingesting platform data coming from the Data Pipeline, and considering AIF requisites and specific performance metrics. In case resources for scaling out are not available, the IOC could trigger a migration through the NSAP IOC.

The IOC will include a specific case/function called *Intelligent Acceleration Resources Manager (IARM)*. IARM is concerned with the management of acceleration resources in the MEC system. IARM will input the AIF-descriptor of an incoming AIF and will output a suggestion to the MECPM regarding the placement of the new AIF on the most suitable HW node of the MEC. IARM will read the key HW requirements of the AIF, i.e., the AIF's available executable versions with their HW architecture and memory requirements, as well as the state of the MEC nodes, i.e., the availability of HW accelerators at a given point in time. Subsequently, IARM will match the computational characteristics of the AIF to the available HW and propose the most suitable placement, e.g., to assign an incoming medium-workload CNN to a specific idle near-edge GPU. IARM can also monitor the MEC's HW state during AIF execution for proposing migration actions to the MECPM towards improving certain performance metrics (e.g., to move a medium-workload CNN to a specific near-edge GPU when it becomes available towards improving the AIF's throughput).

IARM can be instantiated in the CCP either as a function inside the IOC container or as an individual container. Its location in the CCP can be either at the NSAP tier or at a lower level, inside the MEC, with a distinct IARM per MEC cluster. The former case has the advantage of global knowledge and optimization, i.e., the ability to manage all CCP resources with fine-grained decisions. However, the heterogeneity of the HW across all MECs and the requirement of collecting regularly all HW states at a centralized entity would impede CCP scalability. The later approach, i.e., one IARM per MEC for local optimization, will improve CCP scalability and MEC autonomy/dependability (HW resources are not managed by a remote entity), however, at the cost of potentially sub-optimal placement solutions (e.g., in cases where the nearby MEC would have temporarily slightly better HW available for an incoming AIF). In a two-tier orchestration

approach, the MEC will expose coarse HW metrics to the central IOC (e.g., accumulated acceleration utilization), which will select the most suitable MEC based on coarse knowledge of HW availability in MECs and additional HW-independent criteria (e.g., data locality).

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. 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 [23]. 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 has been providing a 5G RAN stack (open source srsRAN project). The srsRAN software suite currently supports both 4G LTE and 5G NR. For 5G NR both 5G NSA and SA are supported. srsRAN applications are implemented in efficient and portable C/C++ supporting a wide range of baseband hardware platforms, including x86, ARM and PowerPC.

For AI@EDGE, the srsENB stack is being modified to add support for both slicing and interfacing with the near real time RIC via the E2 interface as defined by the O-RAN. The E2 interface and communication with the near-RT RIC will allow AI@EDGE to exploit AIFs for the management of the RAN and the prediction of events and anomalies across multiple domains. Slicing will allow for QoS assurances within the RAN, dependent on end-user requirements. These slices will be deployed at run-time based on the needs of users as reported by MEC apps, with the RAN then provisioning the appropriate resources.

The E2 interface is responsible for all communication between the RIC and the RAN stack. To fully support this communication, several protocols are needed, specifically:

- **E2AP:** this manages the setup and maintenance of the communication link between the RIC and the RAN stack.
- **E2SM**: this manages the specific services that are provided by E2 interface, specifically the KPM (metrics reporting) the RC (the control of the RAN) and the NI (network interface message passing).

To support this, ASN1 packing and unpacking were added to srsENB for each protocol. This packing/unpacking functionality is then used to create a wide range of E2 specific procedures (E2 Setup, E2 reset, RIC subscription, RIC indication etc) These procedures are used to create the interface and use it to communicate metrics & commands between the RAN and the RIC.

NSSAI-based slicing support required modifications to both srsUE and srsENB. A slice with a bespoke QoS configuration is achieved by assigning and creating a dedicated PDU session between the UE and the AMF. To enable this the UE must have the ability to communicate its desired NSSAI to both the gNodeB and the AMF. Support for this message extension (including ASN1 packing & unpacking) was added to the RRC and the NAS in the UE. Additionally, the gNodeB must have the capability to both parse NSSAI messages coming from the UE in the RRC and communicate supported NSSAIs to the AMF using the NGAP interface, this support was also added.

3.2.2.2. 5G Core Relevant features and work items

The 5G Core network architecture is continuously evolving based on the progress of the 3GPP standardization groups. We report in this section a selection of relevant internal work items on which Athonet, the 5G core network provider of AI@EDGE, is progressing and that will drive the research and innovation efforts of the project in the near future.

In the current Athonet 5G Core SW release, a first implementation of the N5 standard interface is realized to allow application-level session information exchanges between AFs and the PCF, thus allowing the allocation and management of guaranteed QoS policies for any 5G service or application with specific requirements. In particular, the current version focuses on the support and integration of the MCPTT applications, which address the large market sector of public safety and mission critical applications. Thanks to this interface, the MCPTT application can instruct the network about application-related session information and push the 5G QoS Indicators (5QIs) to guarantee the QoS and retention priority defined in 3GPP for such critical services.

This new feature was showcased and tested in the ETSI MCX Plugtest, held at the University of Malaga, Spain, in November 2022. Note that Athonet is the only 5G Core network provider exposing the N5 interface to support the MCPTT 5QI data flows considering all MCPTT vendors participating to the event. The architecture supporting such feature has been enhanced in all the core network components affected by it for propagating the 5QI information also on the radio side, i.e., Access and mobility Management Function (AMF), Session Management Function (SMF), and UPF.

Beyond MCPTT applications, these QoS management and control mechanisms can be leveraged to enable guaranteed-performance operations of other kinds of services, including AI functions and edge-deployed applications. In particular, the same N5 interface and PCF can also support Voice over New Radio (VoNR) calls in 5G, even though the maturity level of the market allows us to test only a small set of mobile phones fully supporting this service.

Another new work item under study is the NEF (Network Exposure Function). For an AF (Application Function), the NEF is "the entry function" towards network assets in the 5G Core. NEF receives information from NFs based on exposed capabilities and maintains the list of services provided/exposed by the network. Since it authorizes applications to access exposed network services, the NEF is mainly utilized in public and/or national networks rather than private mobile networks, where the services and applications in use are likely to be trusted and local for the final private customer, and do not need NEF mediation.

The main reference 3GPP specifications are TS 23.501[24], TS 23.502 [25] TS 29.522 [26], TS 29.122 [27].

Linked to the NEF functionality, a rising topic is the need to design a proper analytics framework to automate the mobile network from a management point view, and to make available to external parties' data and analytics obtained or elaborated by the network. The latest 3GPP specifications suggest to use the NEF and NWDAF components to allow third-party AI and/or ML based algorithms to perform computational exercises on the metrics retrieved from the network, with the possible outcome of pushing back to the network configuration modifications to outperform the current network behavior or, e.g., reduce the energy consumption, thus meeting the most critical requirements of the 5G Advanced standard specifications from a Core network point of view.

Last but not least, recently 3GPP has enriched its specifications dedicated to the support of edge computing (cf TS 23.548 [28]), introducing support to Edge Hosting Environments (EHE) that host Edge Application Servers (EAS), and identifying as enablers of edge computing a series of 5G features: local routing, traffic steering, session and service continuity, AF influenced traffic routing, etc. This work complements the activities of the ETSI MEC standardization group. The 5G Core network deployments of AI@EDGE include such MEC-based networking solutions for offloading the traffic at the edge of the network, mainly upon deployment of dedicated UPFs which turns out to be either a dedicated network slice or shared by multiple network slices.

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 based on the policies configured by the Non-RT RIC through the A1 interface, still through the E2 interface. More information regarding the implementation of Near-RT RIC and development of xApps in the scope of this project will be outlined in Deliverable 4.2.

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 interfaces and workflows

This Chapter introduces the interfaces and workflows of the AI@EDGE architecture that have been identified in the project. The interfaces and workflows comprehend cross-platform interactions between NSAP and CCP components, internal interactions between CCP components and between NSAP components. The project reuses interfaces introduced by 3GPP, O-RAN, ETSI MEC. It reuses some interfaces introduced on the ETSI MEC architecture, but introduces others, such as the MEO-to-MEO interface, as a novelty.

Section 4.1 describes AI@EDGE interfaces and Section 4.2 describes AI@EDGE workflows.

4.1 AI@EDGE interfaces

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

4.1.1 *AI@EDGE MEC System related interfaces*

Figure 8 displays the interfaces interconnecting MEC Systems: (i) the Multi-Tier Orchestrator and the MEC System Orchestrator; (ii) different MEC Systems; and (iii) the modules within each MEC System.

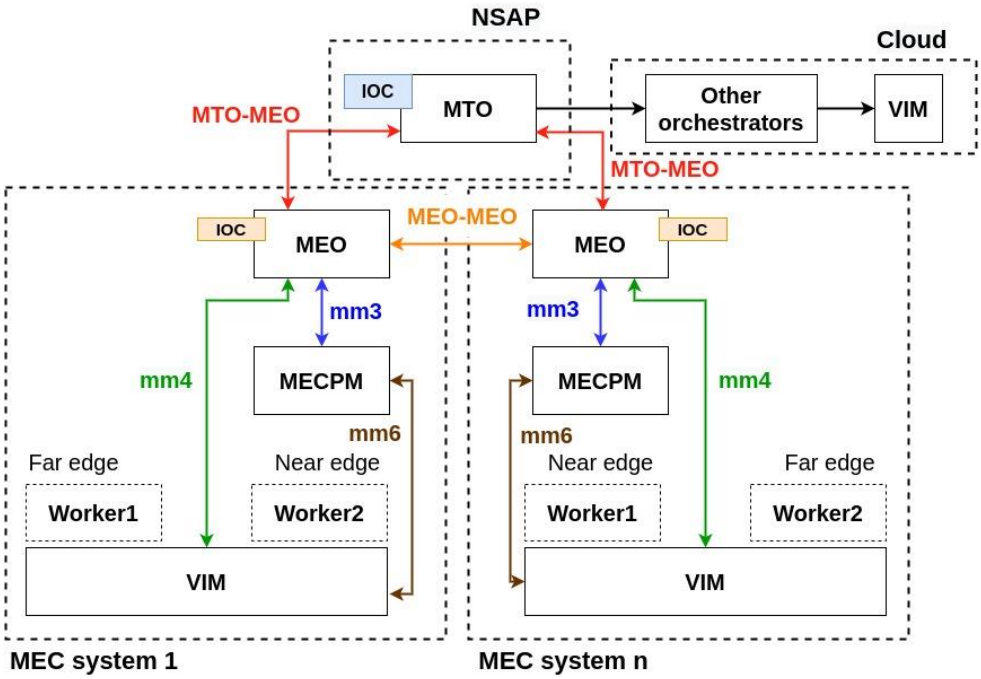


Figure 8 AI@EDGE MEC system architecture interfaces

4.1.1.1. MTO-to-MEO

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 [29], 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

through a messaging system (broker), which supports two types of communication to cover a set of required functionalities and use cases:

- Broadcast: used to detect and identify additional MEC Orchestrators that have joined the network as well as to examine the availability of the present MEOs when needed. Because there is no guarantee that messages will be received, particularly in distributed systems, these queues feature a timeout limit to mitigate this issue.
- Direct (RPC) communication: used as an imperative command to deploy, migrate or delete applications on a particular MEC system.

4.1.1.2. MEO-to-MEO

This interface is designed and implemented within the AI@EDGE project to communicate two MEC systems in a distributed way even if the Multi-Tier Orchestrator suffers any kind of failure. In particular, this control 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, or to any kind of federated communication that may be required. This communication occurs through a messaging system (broker), which provides two types of communication to cover the following functionalities and use cases:

- Broadcast: each MEO can utilize a communication queue to evaluate if the other MEC systems have the capacity to migrate applications, for example, if an SLA breach is imminent. Because there is no guarantee that messages will be received, particularly in distributed systems, these queues feature a timeout limit to mitigate this issue.
- Direct (RPC) communication: it is required when the IOC at MEO level selects a specific MEC system to migrate an application because its requirements are no longer met.

4.1.1.3. 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.

4.1.2 *Near-RT RIC to 5G RAN interfaces*

The ORAN WG3 in [30][31][32] 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 [30], 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 9 Relationship between Near-RT RIC and E2 Node [30] illustrates these interfaces and components.

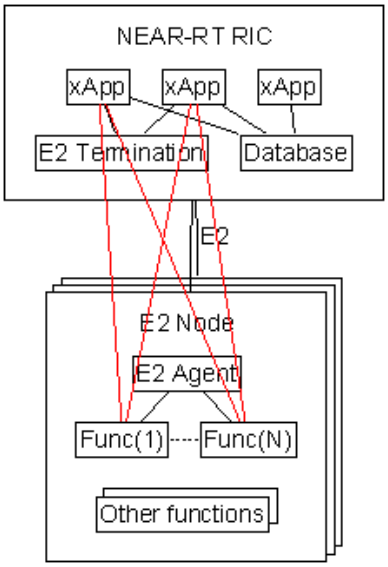


Figure 9 Relationship between Near-RT RIC and E2 Node [30]

The E2 functions are grouped into two categories:

- 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).

4.1.3 *5G RAN interfaces*

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 10:

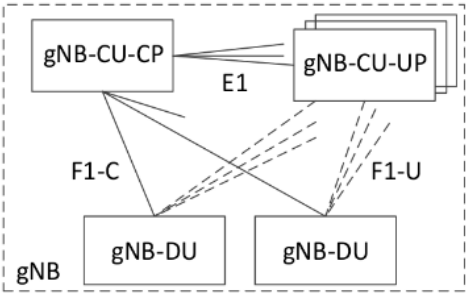


Figure 10 Disaggregated 5G RAN – Components and interfaces

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.

4.1.4 *5G Core Network interfaces*

Figure 11 is an updated version of the corresponding figure reported in D2.2. It depicts the 5GC, the other elements of a 5G system, and the reference points among these elements. The following list is slightly updated compared to that of D2.2:

- **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 signaling 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.

- **N5:** reference point between the PCF and the AF. The interface is used for Application-level session information exchanges between AF and PCF. This is the interface to provide policy rules to Control Plane functions via the PCF to enforce them in the network.
- **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.

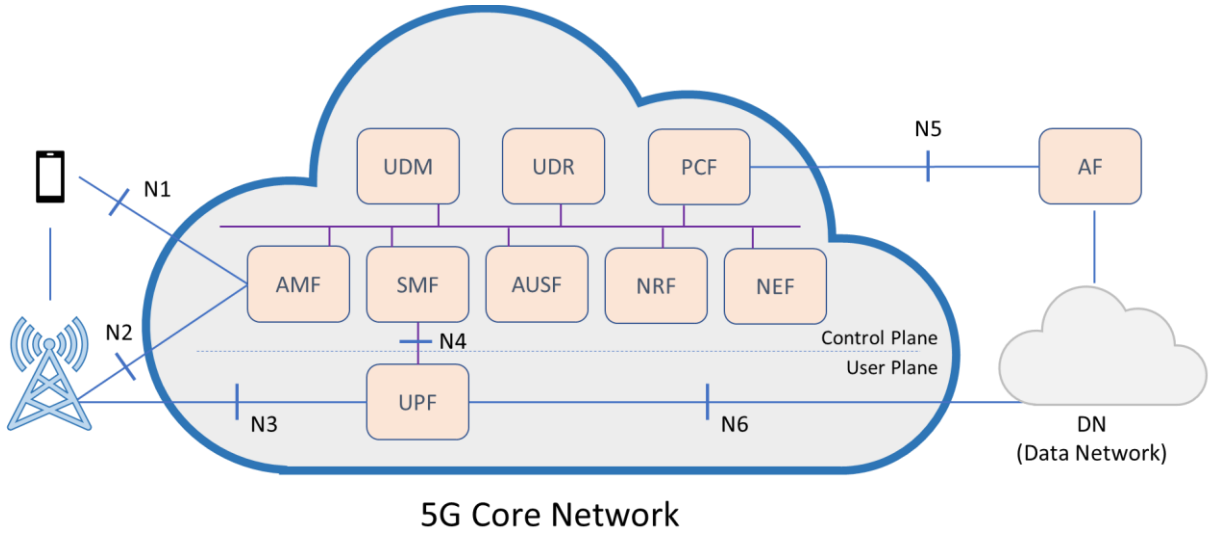


Figure 11 The reference points between the 5G Core Network and the other elements of a 5G system

4.2 AI@EDGE Workflows

This Section details the main workflows defined at the moment of writing this deliverable D2.3. Section 4.2.1 introduces Model Manager workflows, Section 4.2.2 introduces MEC workflows and Section 4.2.3 introduces Non-RT RIC workflowss.

4.2.1 *Model Manager Workflows*

In the AI@EDGE architecture, the AIFs will play a central role regarding the AI-based solutions. The AI-based solutions may be wrapped into one AIF or multiple AI-based solutions may be deployed inside the same AIF. It is important to introduce the concept of *Training AIF* and *Inference AIF*, that can be deployed in the same AIF or in separated AIFs. *Training AIFs* are the AIFs that can only handle training of AI/ML models that are inside them. *Inference AIFs* are the AIFs that can only handle the AI/ML model inference to support a given application. Such models need to be managed to handle changes in the environment where the models execute. In the AI@EDGE architecture, the Model Manager will be responsible for managing the model life stages (e.g., retrain, migrate) of AIFs, as shown in Figure 5.

The Model Manager is currently an independent component of the pipeline described in the AI@EDGE consolidated architecture (see page 29). In the future, it could be a plugin for the MTO/MEO. This would allow faster lifecycle management at the cost of re-usability of models. However, currently the Model

Manager detects if a model update for an *Inference AIF* is necessary. Such updates include the retraining and replacement of the model running inside the AIF. Some of the options to trigger a model update are:

- by monitoring the model performance and comparing it with the desired performance described in the AIF descriptor.
- by detecting data drift which is the change in the probability distribution of the input data; sometimes a model that was trained using a specific data statistic could not work as expected, given the update in the input data; this concept is called data-drift.
- by retrain a model periodically and this periodic time is described in the AIF descriptor.
- on-demand approach in which the AIF contacts the Model Manager asking the model retraining.
- on-demand approach in which the Model Manager contacts the AIF asking the model retraining.

Once the Model Manager decides to update the model running in the Inference AIF, the workflow will vary if the AIFs are independent or have some dependencies from other AIFs. An independent AIF can run by itself and does not depend on others AIF input or output to process its own computation. However, a different scenario could happen which is the one that can have multiple AIFs running according to their dependency graph. The underlying structure of a dependency graph could be seen as a directed graph where each node points to the node on which it depends. This dependency graph illustrates the hierarchy between AIFs and one AIF could not run without following this order and hierarchy. If a model inside AIF needs to be retrained and the AIF follows a dependency graph, then the Model Manager should be able to manage the model retraining without affecting the dependency hierarchy. We describe how to update the models running in the AIFs in different scenarios, both independent and dependent ones, in the next section.

4.2.1.1. Ways to update a model

According to the AIF Descriptor (AIFD), an Inference AIF can be updated in multiple ways, either by retraining or by replacing. Both ways of updating the model are supported by the AI@EDGE platform. By retraining a model, we mean retraining with new data via a Training AIF. By replacing a model, we mean taking a pre-trained model and deploying it on an Inference AIF.

Figure 12 shows how the platform can update the models inside the AIFs, and following we describe in detail each of the scenarios with their variants. It is important to note that every time a Training AIF executes a training procedure it will send the new model to the Inference AIF. The Training AIF will also register the new model to the Model Manager so it will be possible for this last one to handle the lifecycle of such new model.

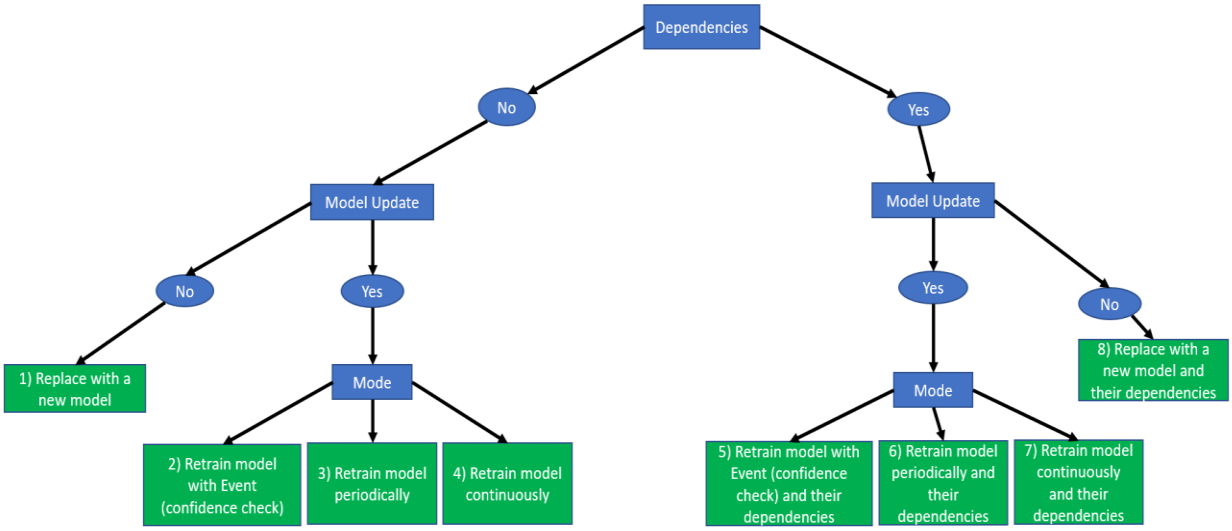


Figure 12 Model updates possible scenarios

1.A Replace old model with a new model (compatible model is available)

In this scenario, the Model Manager collects the model metadata to check the model performance by comparing the CONFIDENCE and PREDICTION output values with the model metadata information. This collection is triggered by any MODE (e.g., EVENT, TIME, PERMANENT), as specified in the AIF descriptor.

The evaluation done by the Model Manager depends on the values defined in the AIF Descriptor and the presence of a machine learning drift, such as concept or data drift. For example, in a supervised machine learning scenario a threshold for performance can be established; thus, by comparing the PREDICTION and CONFIDENCE the Model Manager can decide to update the model. Another reason to update a model would be detecting the presence of concept drift with the new data.

Once the Model Manager establishes the need for a model update, it will read the AIF descriptor to verify the presence of the model Update boolean flag; if the flag is unset, then the Model Manager will replace the model altogether. If the Model Manager decides to replace the model, then it checks in the Model & Data database if there is a compatible model to replace the old AIF inference one. In this case, the compatible model is available, and the Model & Data database returns the compatible model to the Model Manager which sends the new model to the correspondent Inference AIF. The new model with associated metadata is registered to ensure proper lifecycle management, such as updating the model due to a new concept drift.

The execution of the scenario is shown in Figure 13.

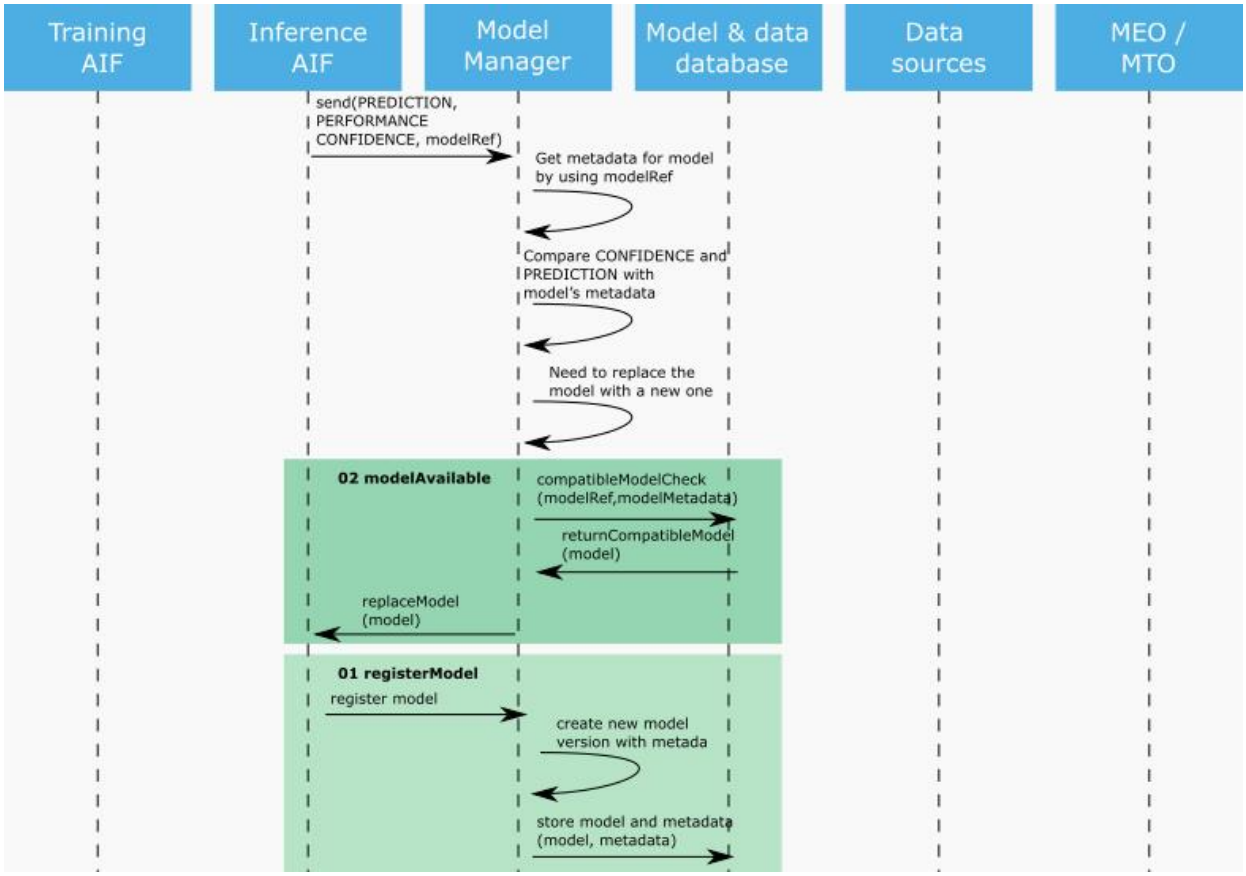


Figure 13 Replacing with a new model scenario

1.B Replace old model with a new model (compatible model is unavailable)

This scenario is similar to the previous one, but the compatible model is unavailable in the Model & Data database. To overcome this issue, the Model Manager checks for compatible dataset using the input source as parameter. If the compatible dataset is found, the Model Manager requests the MEO /MTO to instantiate a Training AIF using the compatible dataset to train the new model. After this training, the new model is delivered to Inference AIF and registered in the Model & Data database entity for the Model Manager to keep track of this new model, as shown in Figure 14.
However, if there is no compatible dataset, the Model Manager requests the MEO/MTO to instantiate a Training AIF to train the new model, but the training AIF will request for collecting the new data until a training dataset is created. After this the trained model is delivered to Inference AIF and registered in the Model & Data database to be tracked by the Model Manager, as described at the beginning of the section. For all subsequent workflows of the model manager, we consider that the Training AIF sends the new train model to the Inference AIF after finishing training. The execution of this scenario is shown in Figure 15.

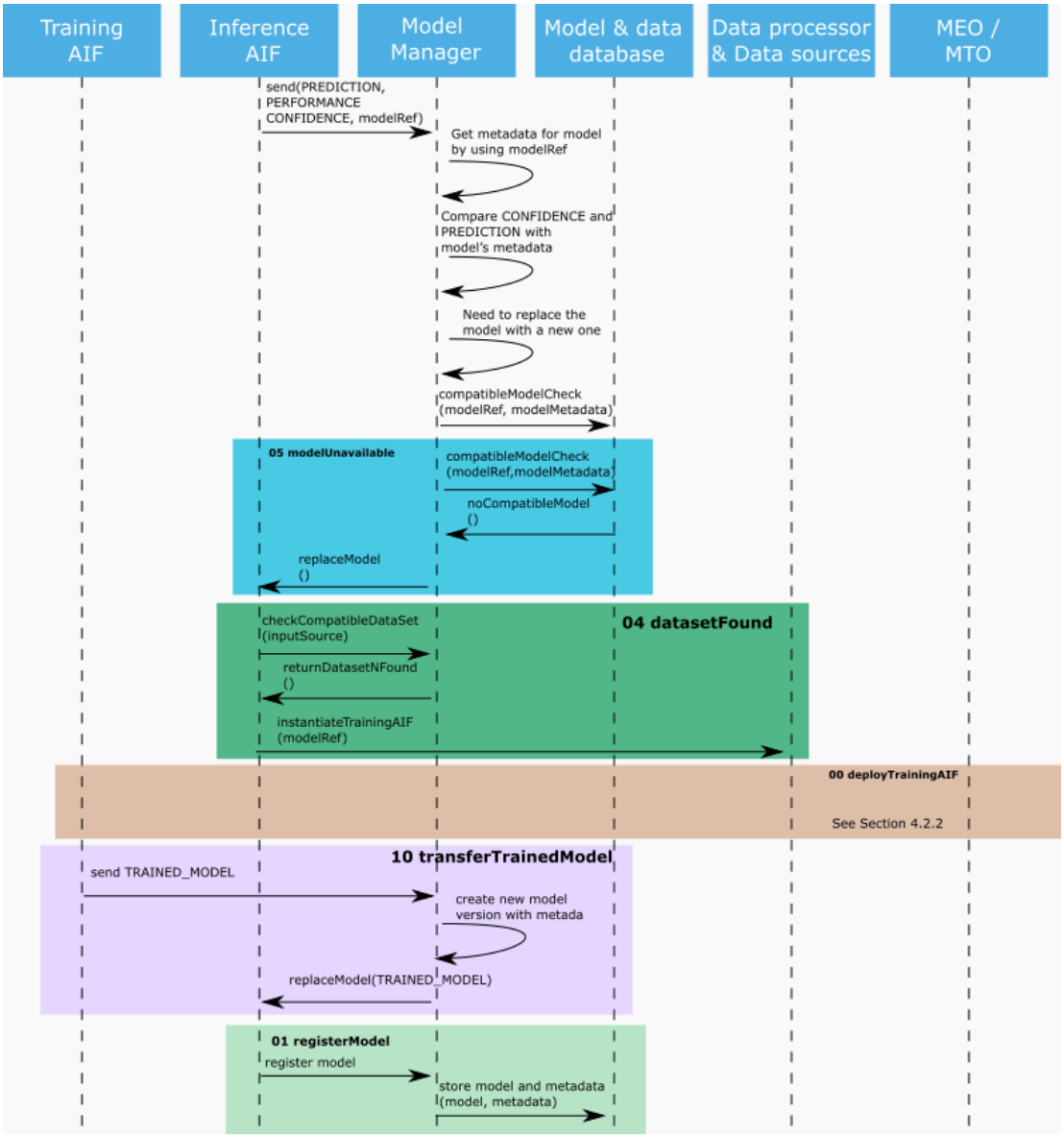


Figure 14 Replacing with a new model (compatible model unavailable). Dataset found

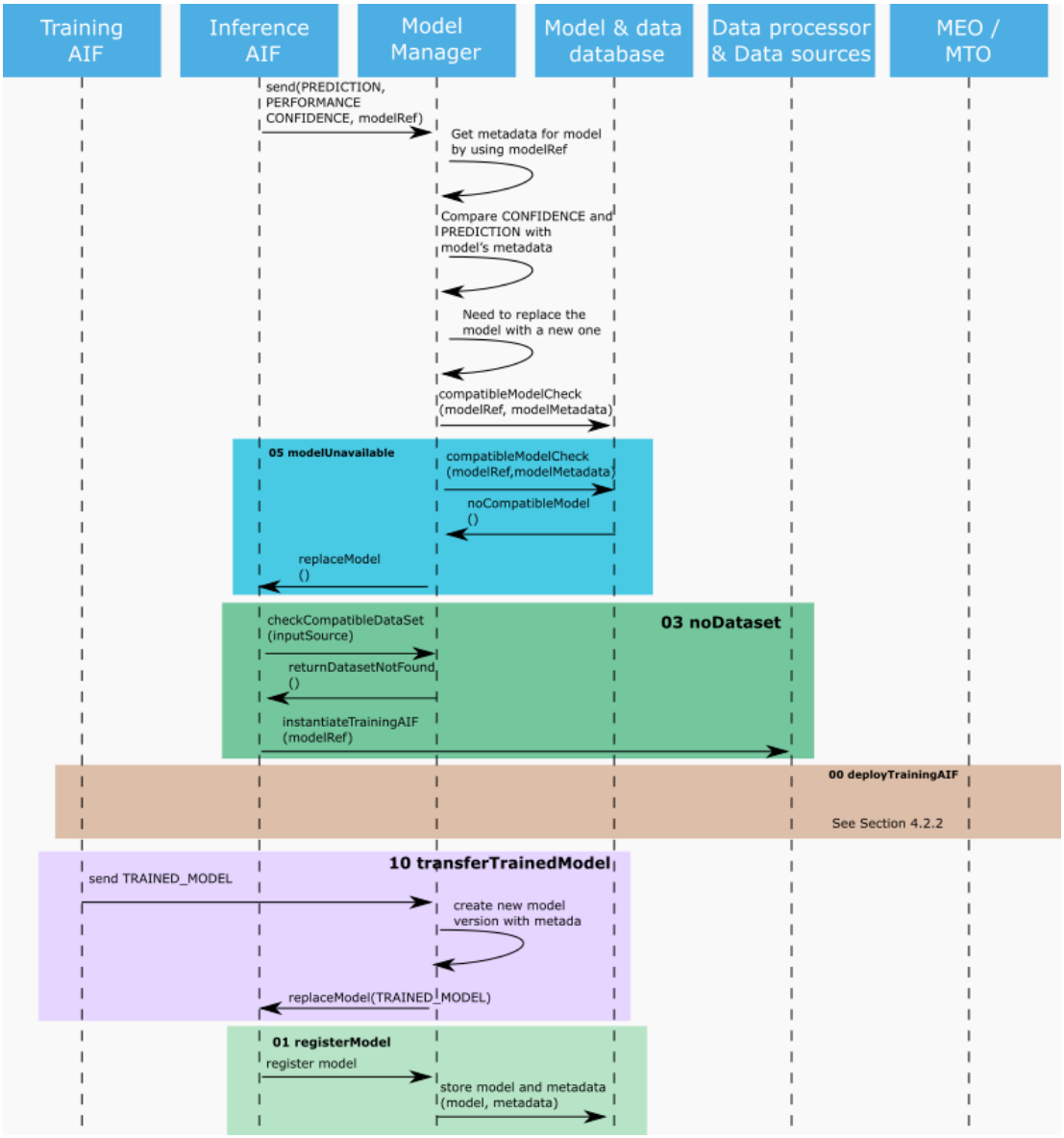


Figure 15 Replacing with a new model (compatible model unavailable). Dataset not found

2.A Retrain model with EVENT (Confidence check)

In this scenario, the AIF is "activated" by the Model Manager via an event that could be an API-call request, a message publication on a channel AIF is subscribed to, and an activation on deploy. After the Inference AIF operation, the Model Manager collects the output of the AIF, such as PREDICTION, PERFORMANCE, CONFIDENCE, and modelRef parameters, as defined by the AIF descriptor. Then, using the modelRef, the Model Manager collects the model's metadata and compares the current PERFORMANCE and CONFIDENCE values with the ones described in the metadata. In this case, the

Model Manager decides that the model needs to be re-trained due to the low performance. The Model Manager asks the Training AIF to re-train the new model.
If no compatible data is found in the database, the Training AIF needs to get new data from sources. After the Training AIF finishes the training, the updated model is registered in the Model & Data database to be tracked by the Model Manage, as shown in Figure 16.
If dataset is found in the database the process is simplified and the dataset is sent to the Training AIF, as shown in Figure 17.

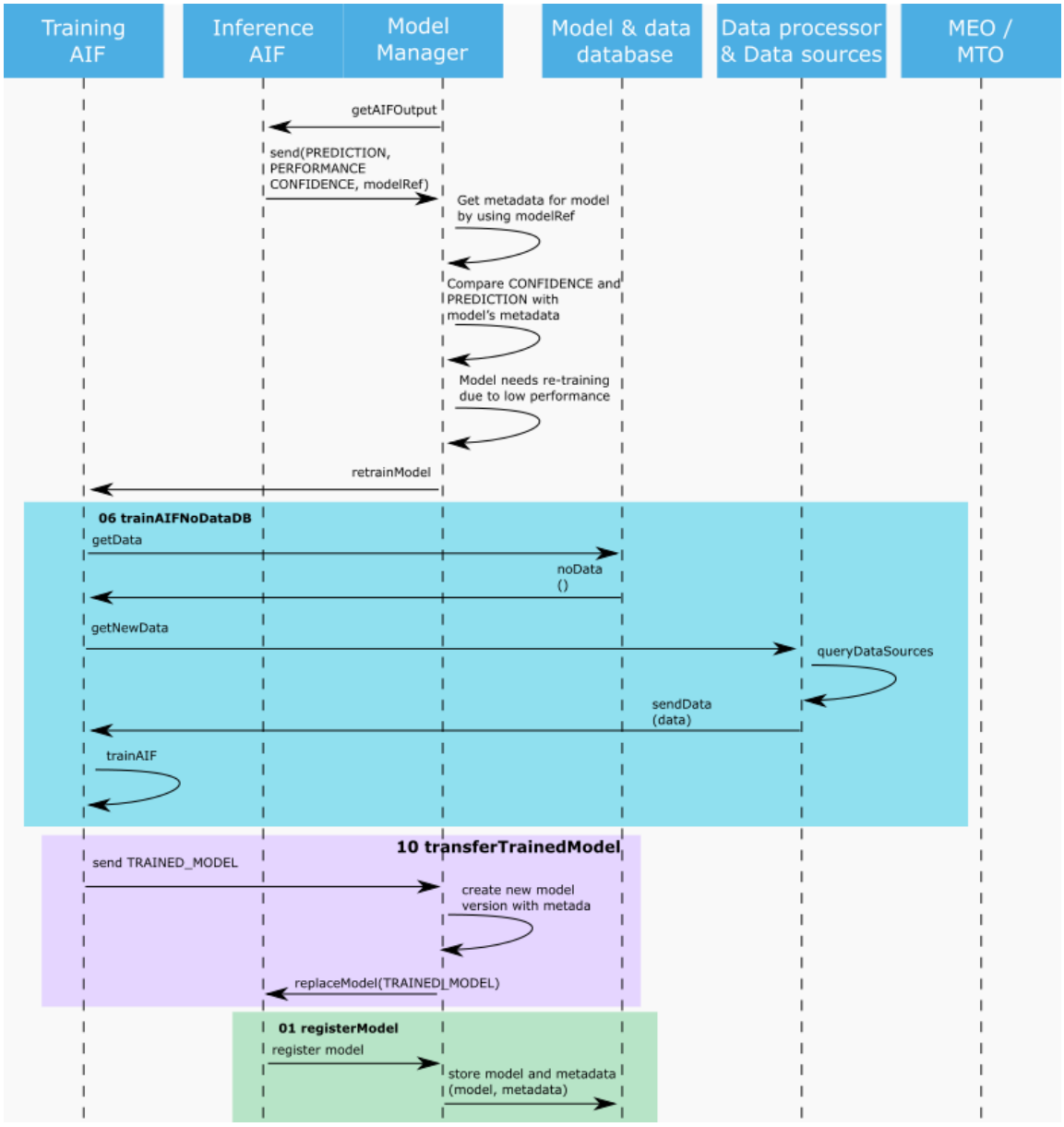


Figure 16 Retrain model with EVENT (Confidence check). No dataset in database

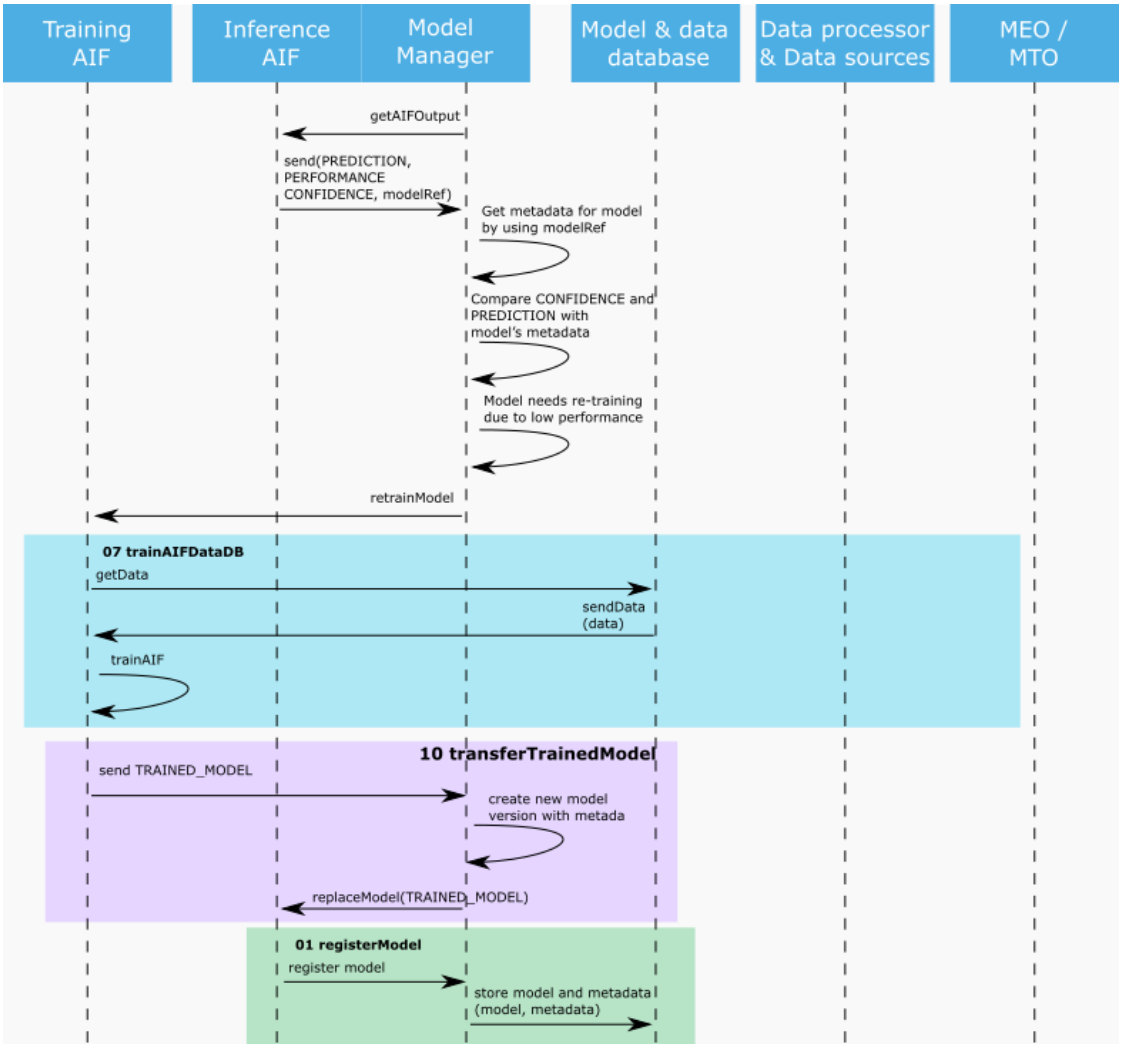


Figure 17 Retrain model with EVENT (Confidence check). Dataset in database

2.B Retraining model with EVENT (confidence check). Training AIF is not available

This scenario starts like the previous one, but the difference begins when the Model Manager decides that the current model needs to be re-trained, due to the low performance, but the Training AIF is not available. The Model Manager requests the MEO/MTO to instantiate a Training AIF to train the current model.
If the data is not present in the database, the Training AIF needs new data, as shown in Figure 18, otherwise, the Training AIF just fetches the data from the database, as shown Figure 19 .
After this training, the updated model is delivered to the Inference AIF and registered in the Model & Data database.

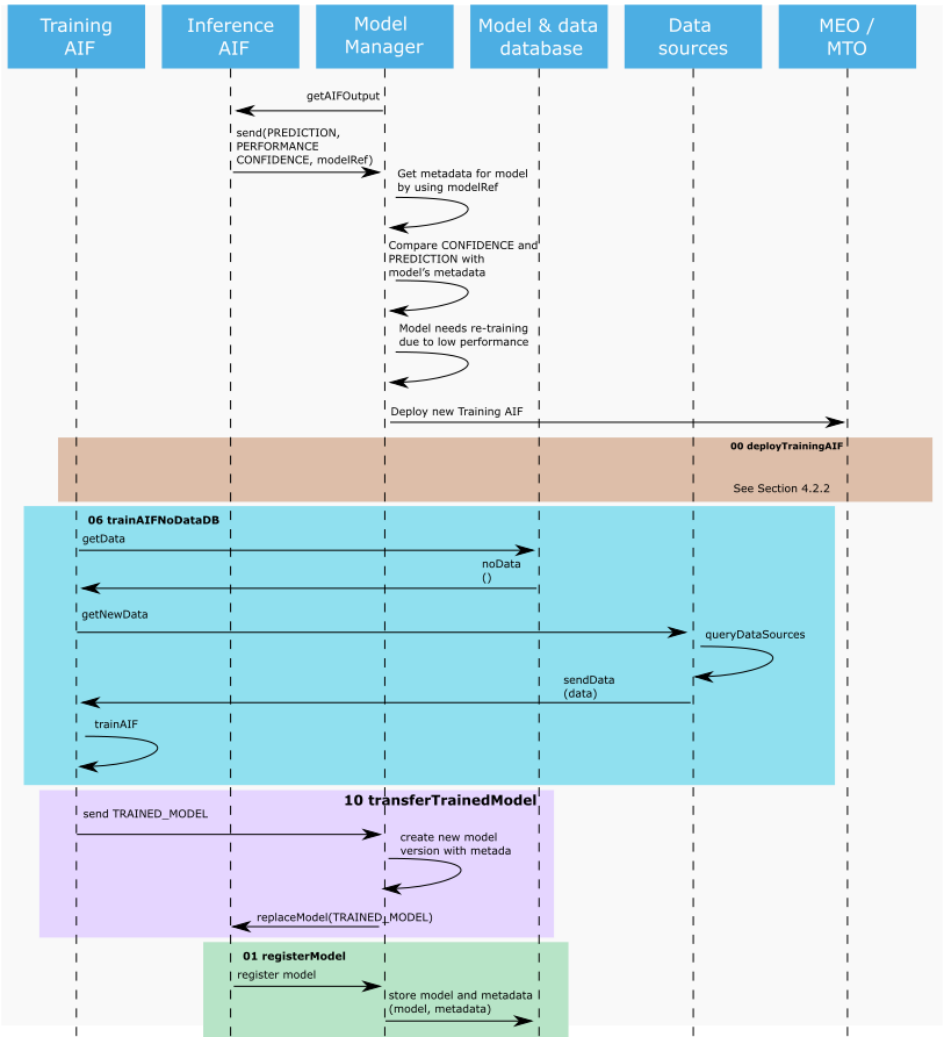


Figure 18 Retraining model with EVENT (confidence check). Training AIF is unavailable and no data in database

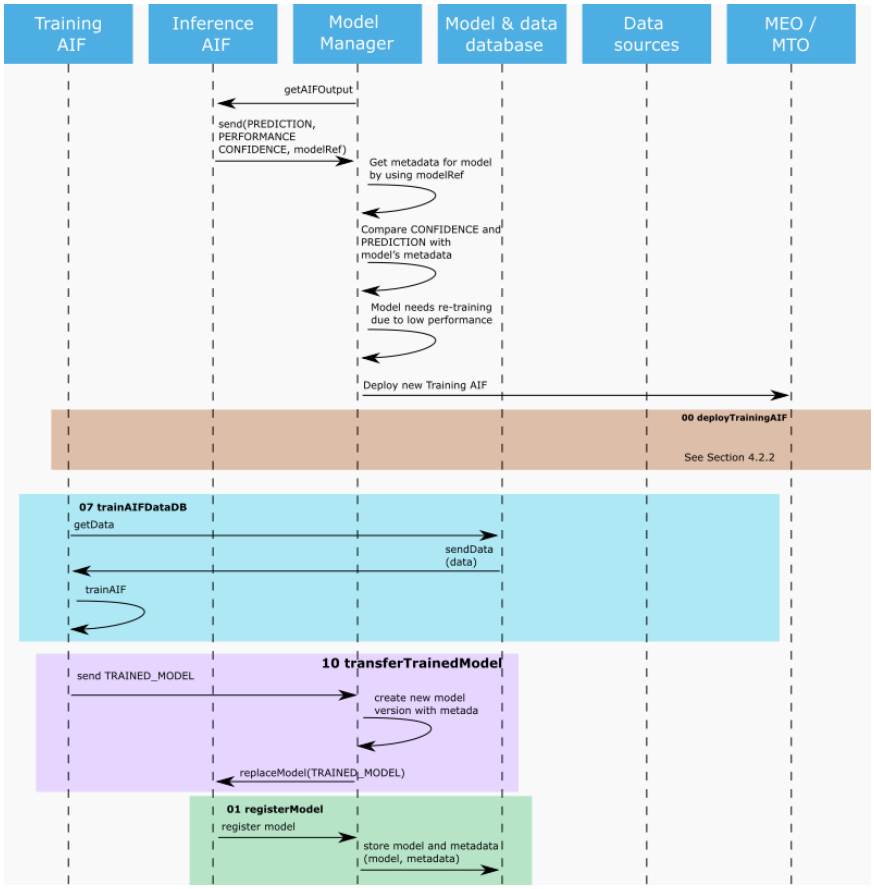


Figure 19 Retraining model with EVENT (confidence check). Training AIF is unavailable and data in database

3.A Retraining model periodically

This scenario starts with the Model Manager for models registered in the Model Manager as those needing periodical update. After the specific time period has expired, the Model Manager triggers the Training AIF, that initiates the training procedure. If data is unavailable in the database, the Training AIF will obtain new data from data sources, as shown in Figure 20. If data is found in the database, the Training AIF receives the data without needing to request new data from data sources, as shown in Figure 21. After it finishes training, it sends a model as an output to Model Manager that registers it in the Model & Data database.

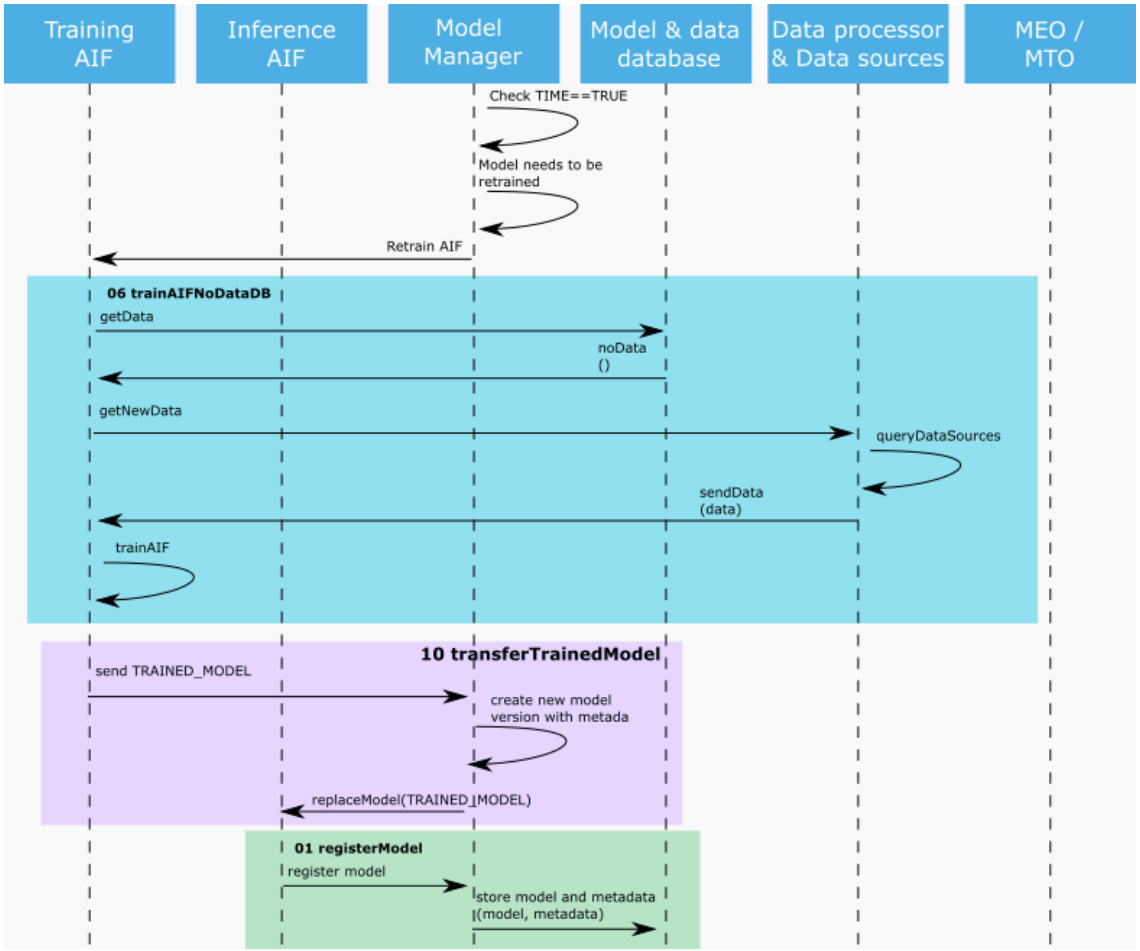


Figure 20 Retraining model periodically. No data in database

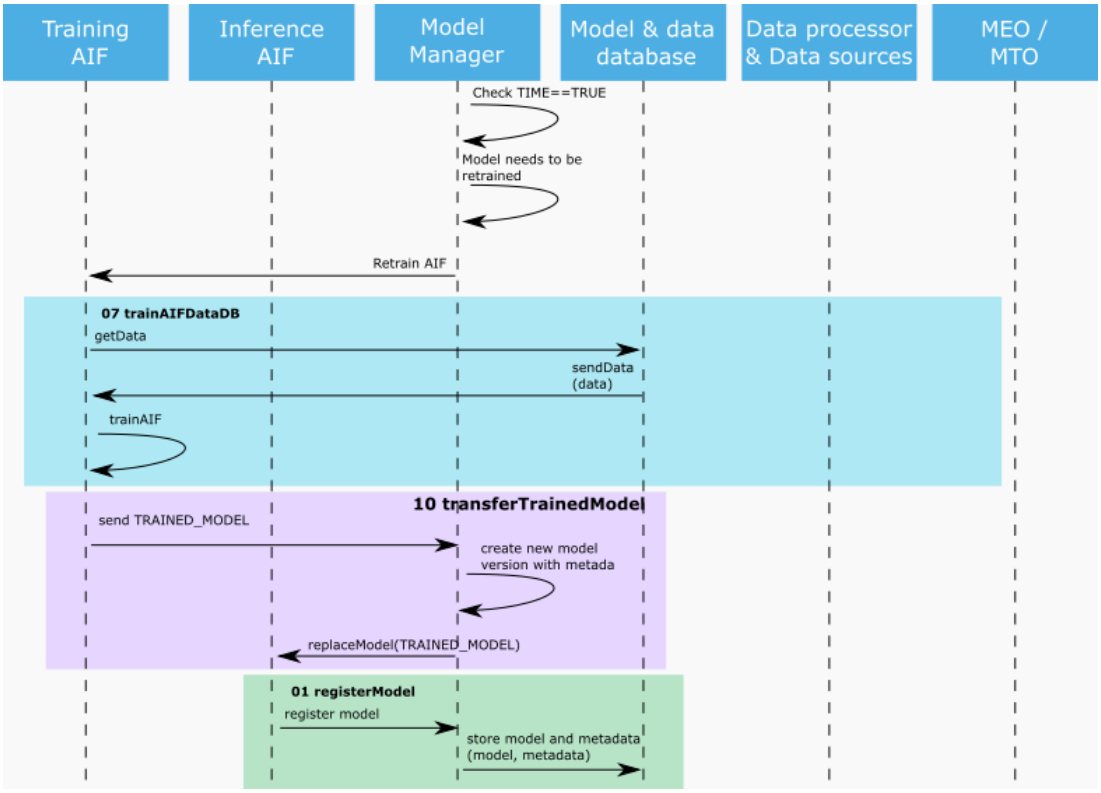


Figure 21 Retraining model periodically. Data in database

3.B Retraining model periodically. Training AIF is not available

This scenario starts with the Model Manager, for models registered in the Model Manager as models needing periodical update. After the specific time period has expired, the Model Manager sends the request to the MEO/MTO to deploy a Training AIF.
If data is unavailable in the database, the Training AIF will obtain new data from data sources, as shown in Figure 22.
If data is found in the database, the Training AIF receives the data without needing to request new data from data sources, as shown in Figure 23.
After successfully deploying the Training AIF, and the model has been trained, the Training AIF sends the updated model to the Inference AIF. The Inference AIF registers the model to the Model Manager and Model and Data database.

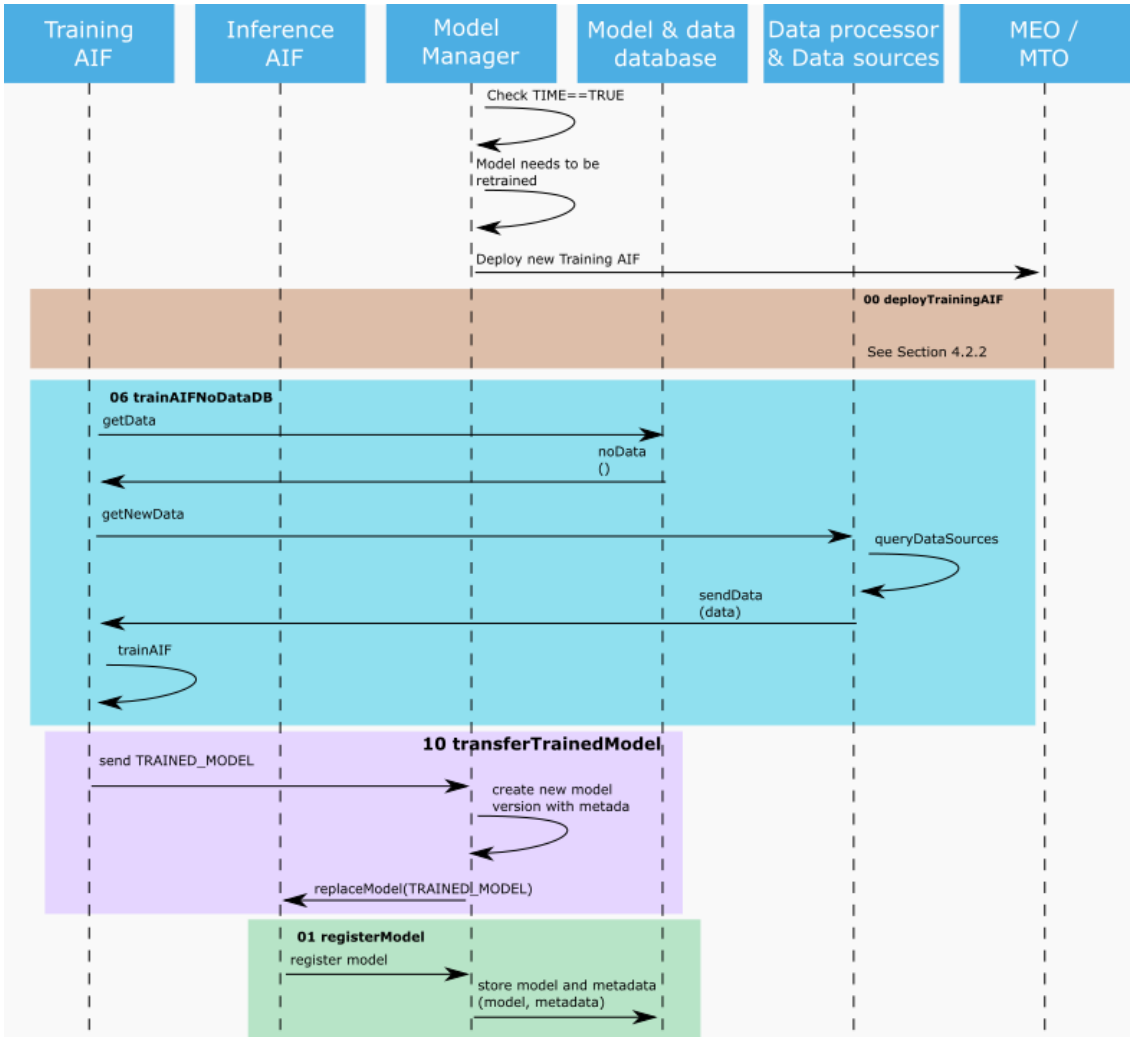


Figure 22 Retraining model periodically. Training AIF is not available, and data is unavailable in database

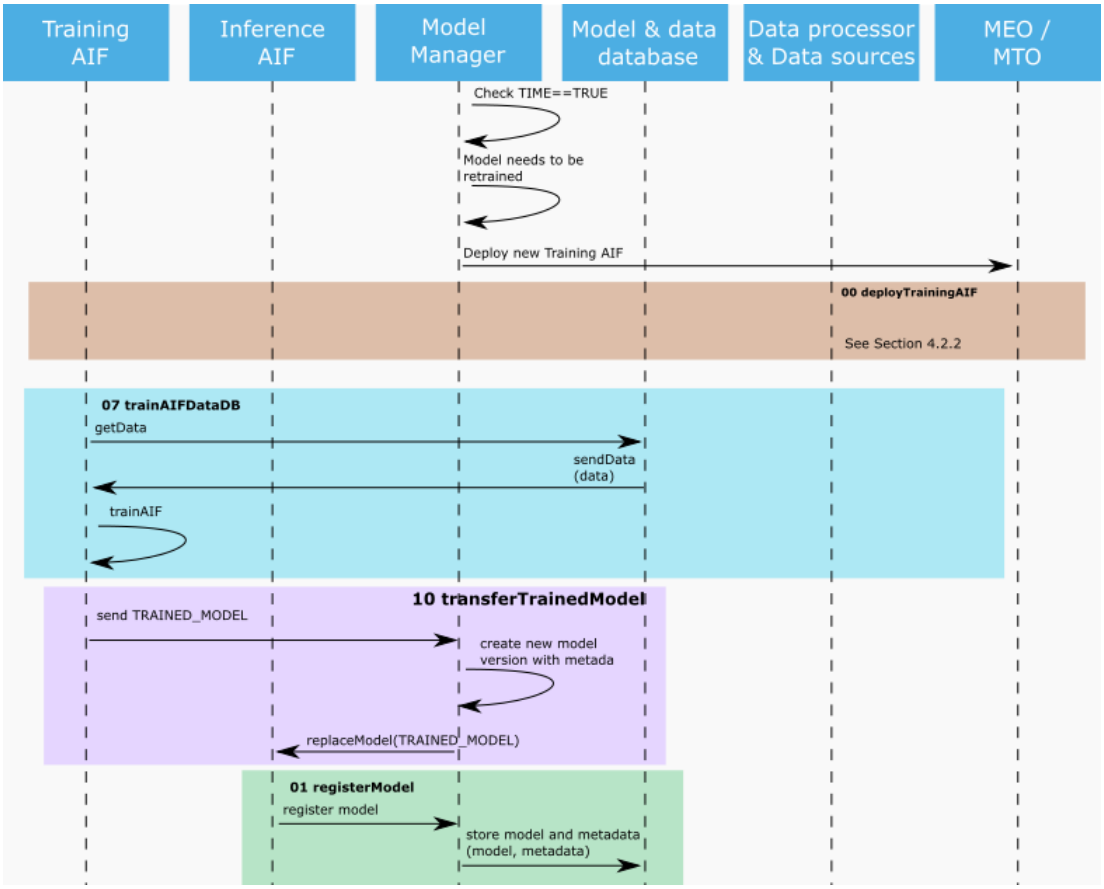


Figure 23 Retraining model periodically. Training AIF is not available, and data is available in database

4.A Retraining model continuously

This scenario starts in the AIF, due to the continuous training as described by Boolean flag in the AIF descriptor. In this scenario, the AIF continuously receives new data from either data sources or the database if compatible data exists in the database. As new data are received, the Training AIF trains on new data received.

If data is unavailable in the database, the Training AIF will obtain new data from data sources, as shown in Figure 24.

If data is found in the database, the Training AIF receives the data without needing to request new data from data sources, as shown in Figure 25.

To ensure proper model performance in the presence of a type of drift, the Inference AIF output (e.g., PREDICTION) is also sent to the Model Manager to track the performance of the model. In both Figures, the two coloured circles, red and green, represent the concurrent execution of the retrain phase and the evaluation of performance, respectively. This scenario illustrates the case of using reinforcement learning in a Training and Inference AIF, where a model is constantly getting new data to improve performance and to be trained.

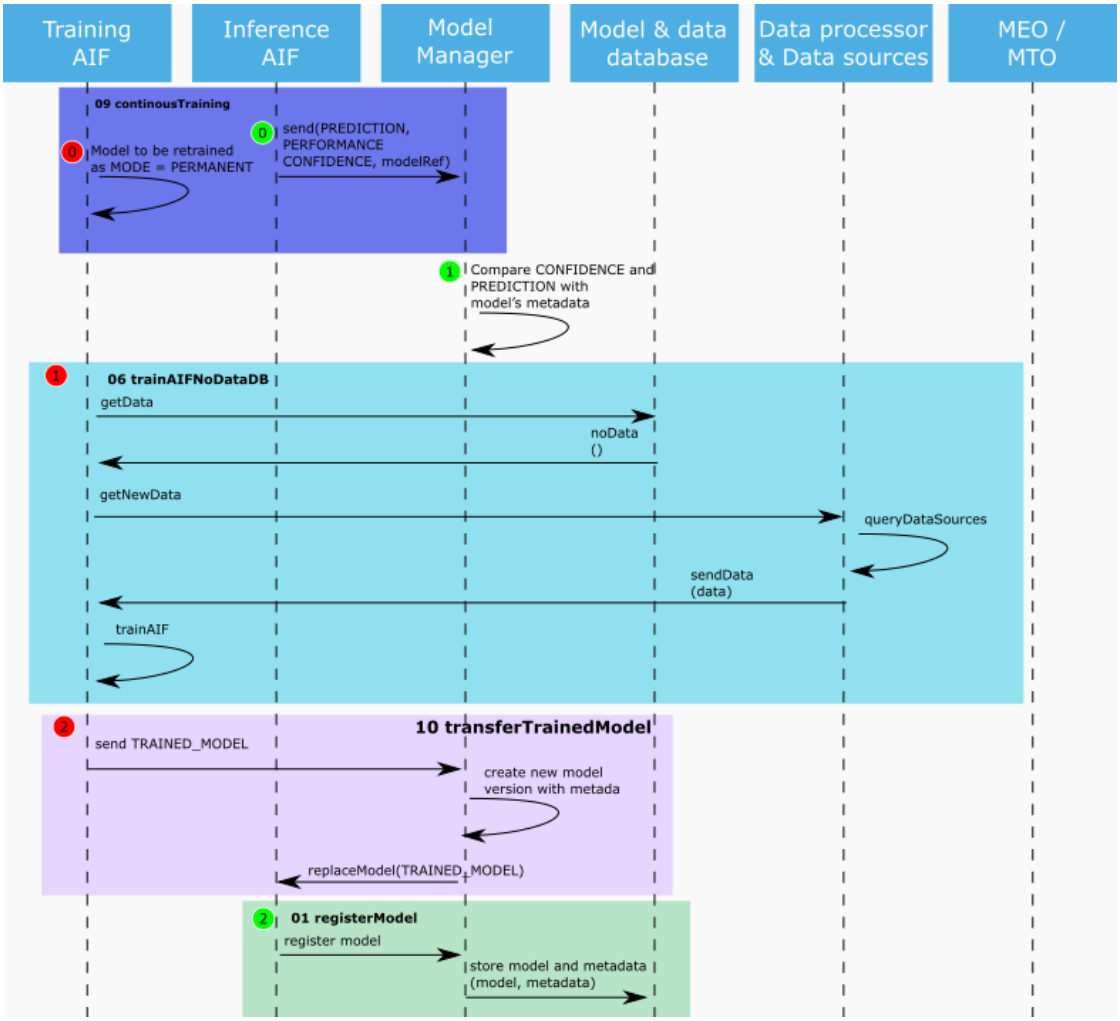


Figure 24 Retraining model continuously. The data is unavailable in the database

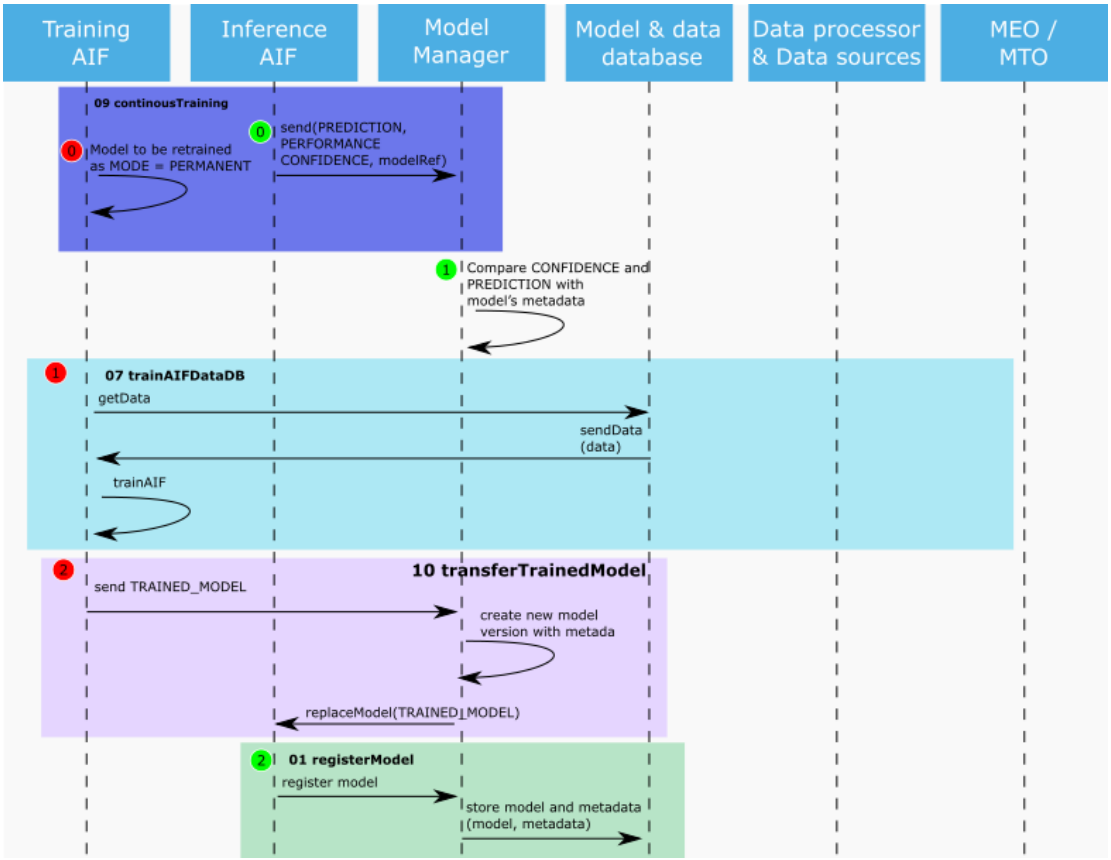


Figure 25 Retraining model continuously. The data is available in the database

5.A Retraining model with event (confidence check) and their dependencies

This scenario is similar to the one in 2.A, but the AIF has multiple dependencies. A dependency could be a data source or another service (e.g., another AIF). Such dependencies introduce overhead in the lifecycle management of the AIF. For example, if an AIF needs to be updated, its dependencies need to be notified about the change to ensure proper service execution. Otherwise, an AIF might change, and the input data no longer be valid.

To ensure the safe update (either by re-training or replacing the model), the Model Manager needs to trigger a synchronization process with the MEO/MTO, that has the information and can communicate outside the specific domain to enable E2E services composed by AIFs. Thus, when the Model Manager decides to update the model, it first sends a notification of change to the MEO/MTO, as shown in procedure 08 of Figure 26. This triggers a synchronization between the MEO/MTO and the dependencies (which could be out of the domain).

The synchronization mechanism involves a Model Manager, the new updates, the dependency's metadata, and the dependency's AIF descriptor. If the new updates do not trigger an impact, then the Model Manager can reply with a flag to signal safe update. For example, if the AIF descriptor establishes a threshold for performance which is impacted by the update, it will signal there is a problem for the update. It is important to note that this synchronization mechanism is recursive. In other words, the original AIF dependencies can have their dependencies, triggering a new synchronization mechanism.

After the synchronization mechanism, the MEO/MTO sends a notification to the Model Manager. As the current AIF that contains the model is the Training AIF, the Model Manager triggers the AIF to re-train the new model.
If data is unavailable in the database, the Training AIF will obtain new data from data sources, as shown in Figure 26.
If data is found in the database, the Training AIF receives the data without needing to request new data from data sources, as shown in Figure 27.
After this training, the updated model is registered with the model metadata and it is stored in the Model & data database to be tracked by the Model Manager.

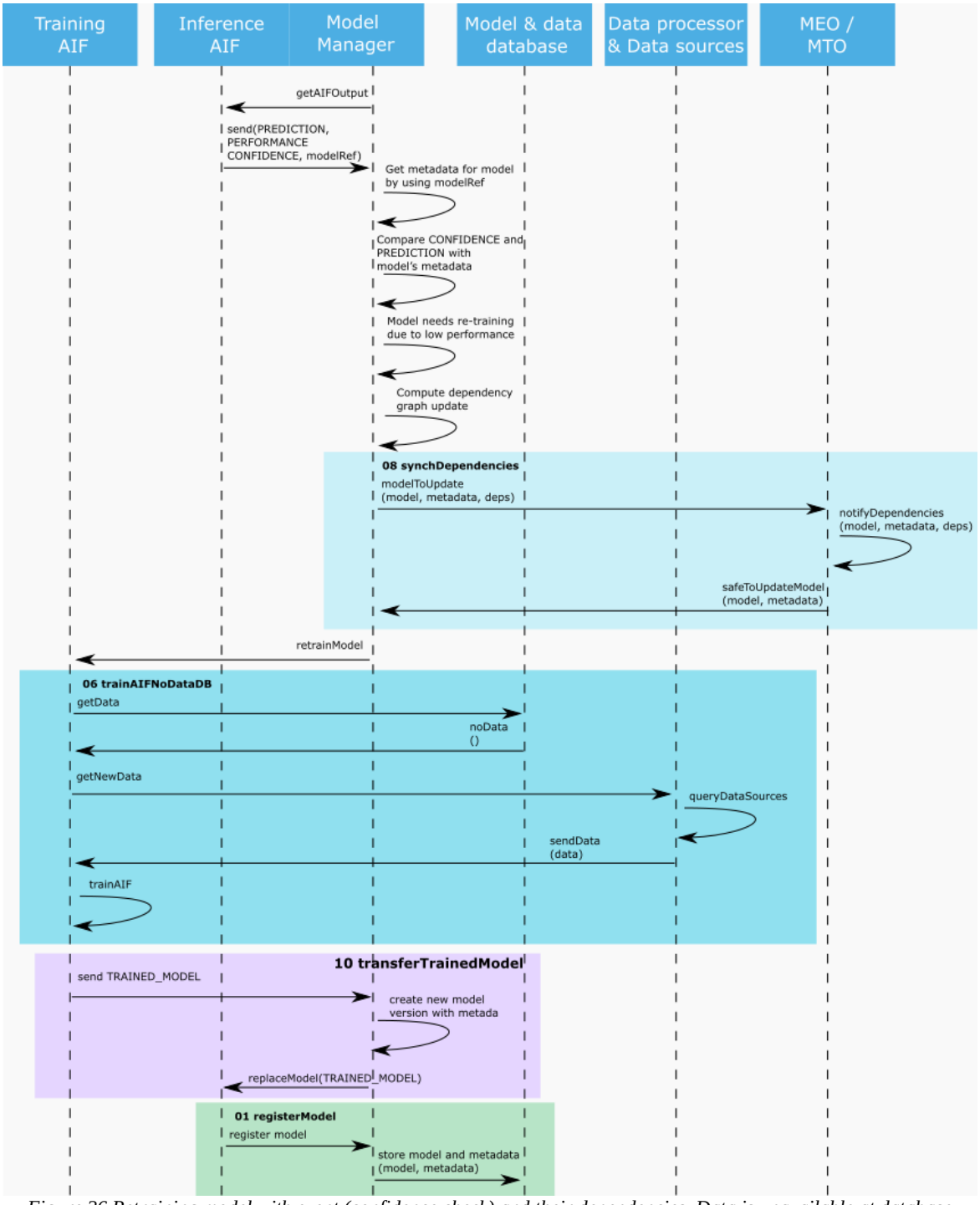


Figure 26 Retraining model with event (confidence check) and their dependencies. Data is unavailable at database

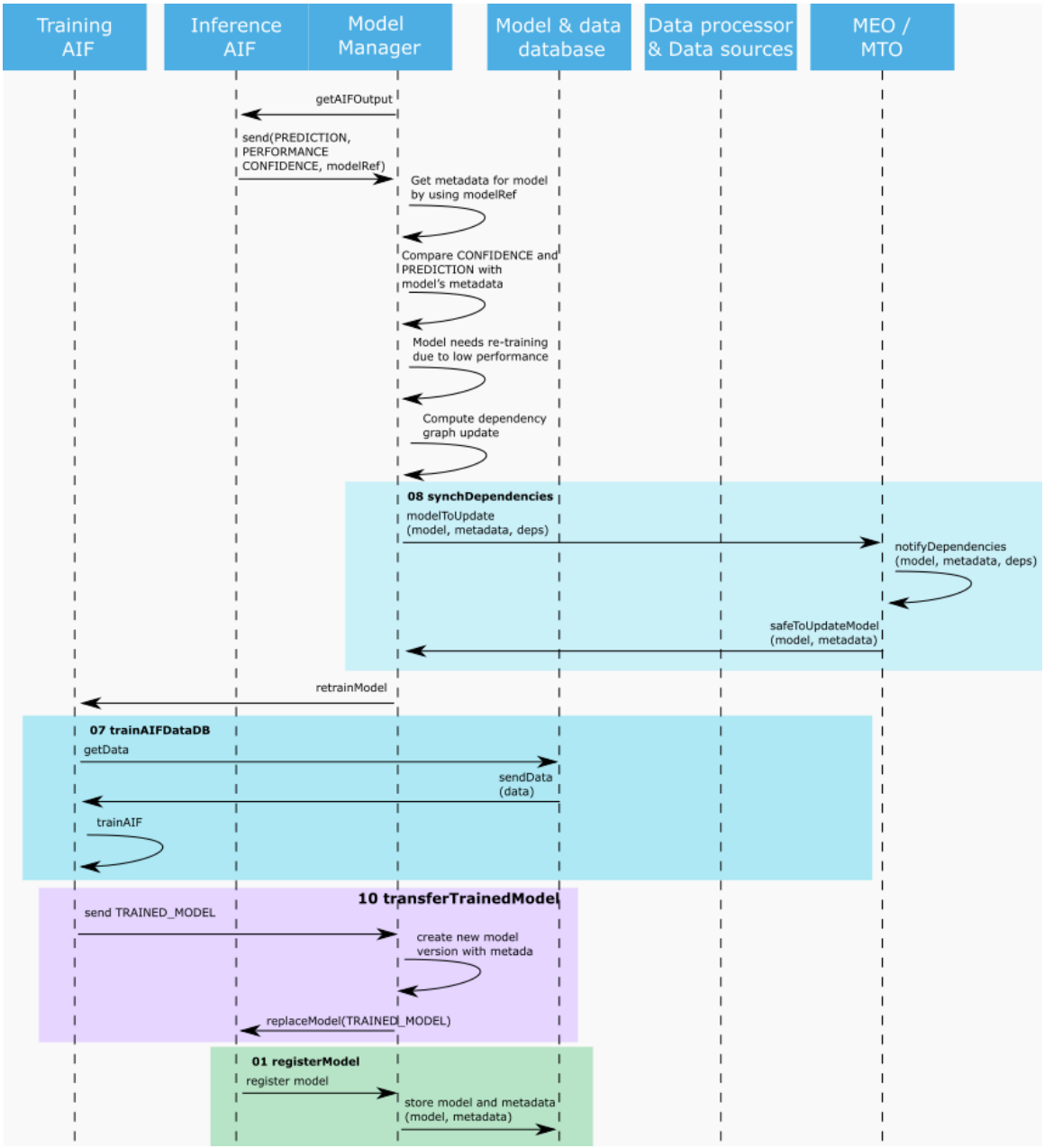


Figure 27 Retraining model with EVENT (confidence check) and their dependencies. Data is available at database

5.C Retraining model with event (confidence check) and their dependencies. Training AIF is not available

This scenario starts like the one shown in 5.B, but the difference begins when the Model Manager decides that the current model needs to be re-training due to the low performance. However, in this scenario there is no Training AIF and, to ensure the safe update (either by re-training or replacing the model), the Model Manager needs to trigger a synchronization process with the MEO/MTO. After this process, the Model Manager indicates to the MEO/MTO to instantiate a Training AIF. Then, the Model Manager checks if there is a available dataset that can be used to re-train the model, as described in Figure 28, otherwise, the data is collected until there is a dataset that can be used for the training as described in Figure 29.

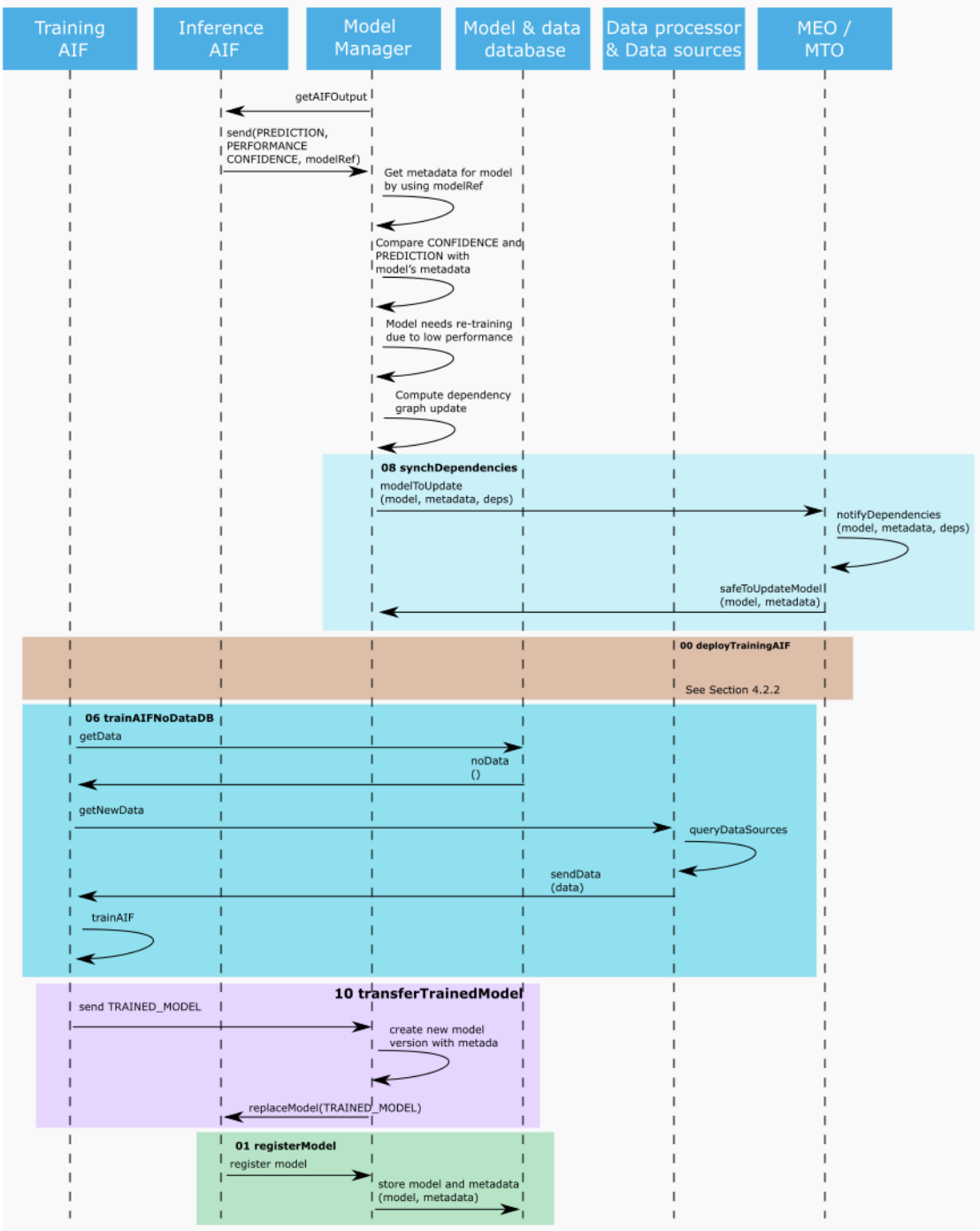


Figure 28 Retraining model with EVENT (confidence check) and their dependencies. Training AIF is not available, and data is unavailable at database.

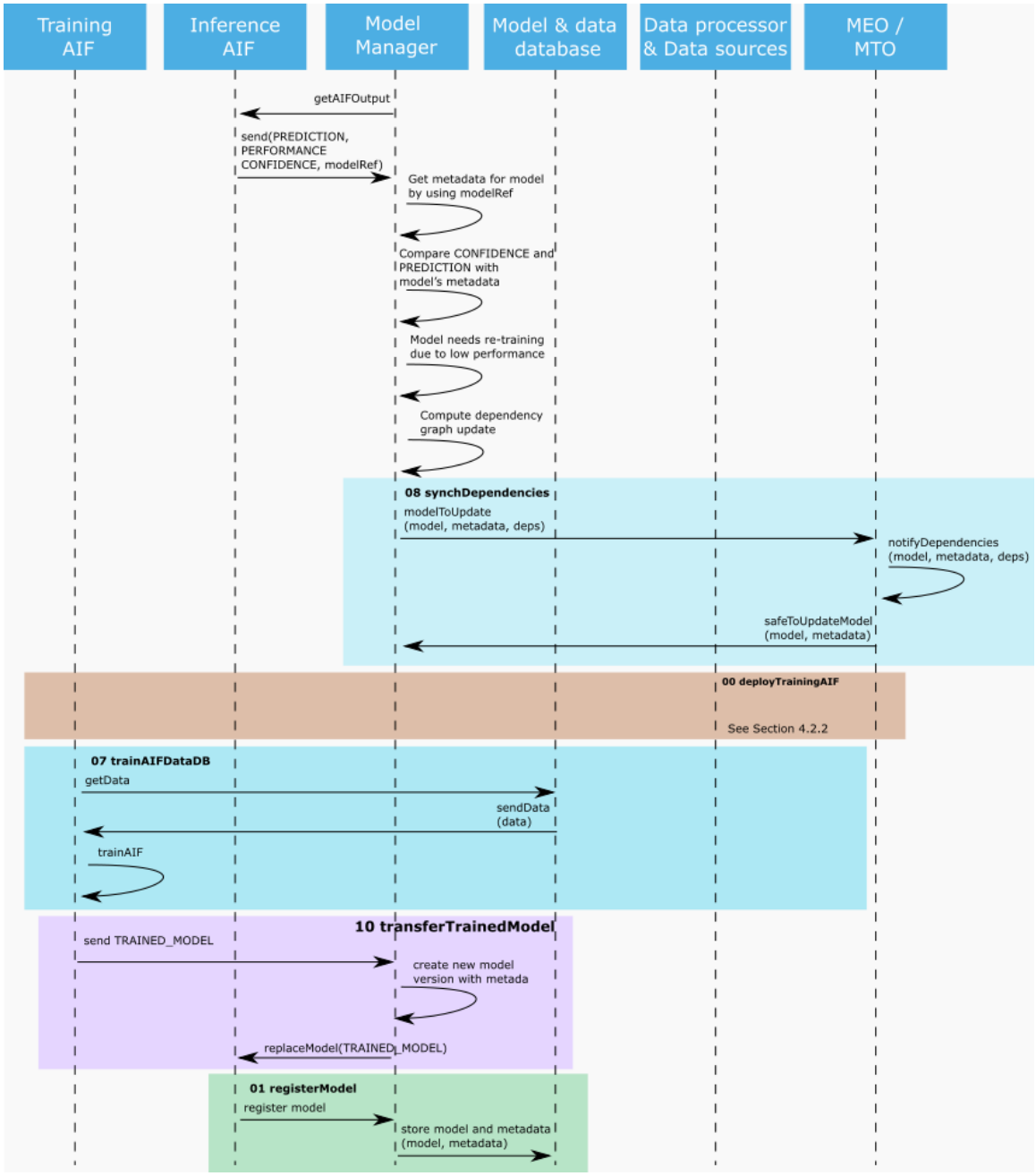


Figure 29 Retraining model with EVENT (confidence check) and their dependencies. Training AIF is not available, and data is available at database

6.A Retraining model periodically and their dependencies

This scenario is like the one described in 3.A, but the AIF has dependencies. At the beginning the Model Manager checks the TIME variable equals to TRUE to decide that it is time to retrain the model. Therefore, the Model Manager sends the MEO/MTO a notification to update the AIF dependencies as there will be an

update in the corresponding model. After this notification, the step called “08 synchDependencies” starts to make sure that, after this training, the models' dependencies will be still valid. Then the AIF retrains the model.
If data is unavailable in the database, the Training AIF will obtain new data from data sources, as shown in Figure 30. If data is found in the database, the Training AIF receives the data without needing to request new data from data sources, as shown in Figure 31. After this retraining, the Model Manager will Register the model for tracking purposes and will save it in the Model & data database. The last step is to notify the MEO about the new dependencies' configurations after this retraining.

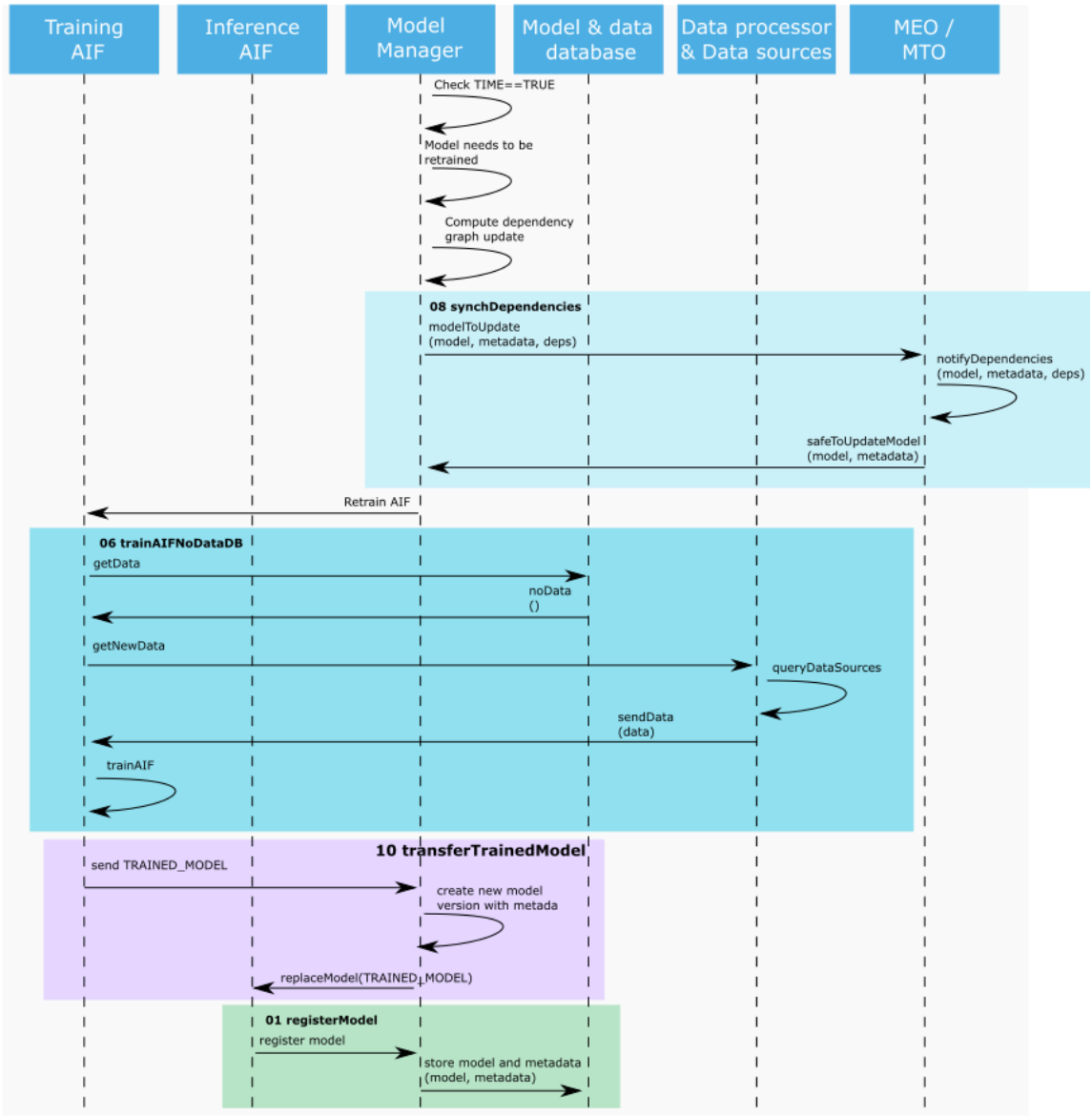


Figure 30 Retraining model periodically and their dependencies scenario. Data is unavailable at database

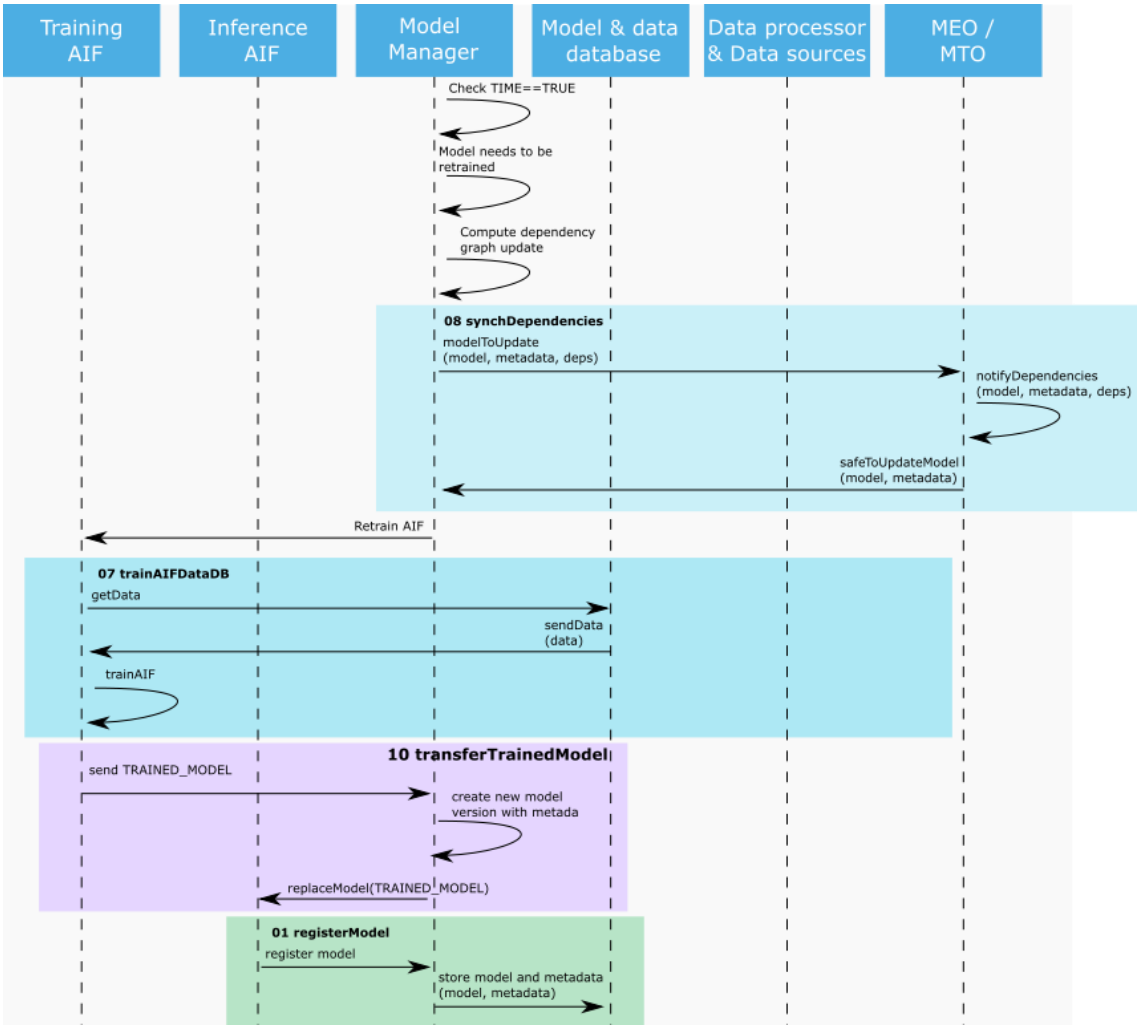


Figure 31 Retraining model periodically and their dependencies scenario. Data is available at database

6.C Retraining model periodically and their dependencies. Training AIF is not available

This scenario starts like the previous one, but the difference is that the Training AIF is not available. It is the role of the Model Manager to ask the MEO/MTO to instantiate a new Training AIF to retrain the model. After this retraining, the output of the AIF will be a training model. After this step, the trained model will be registered and saved in the Model and data database.

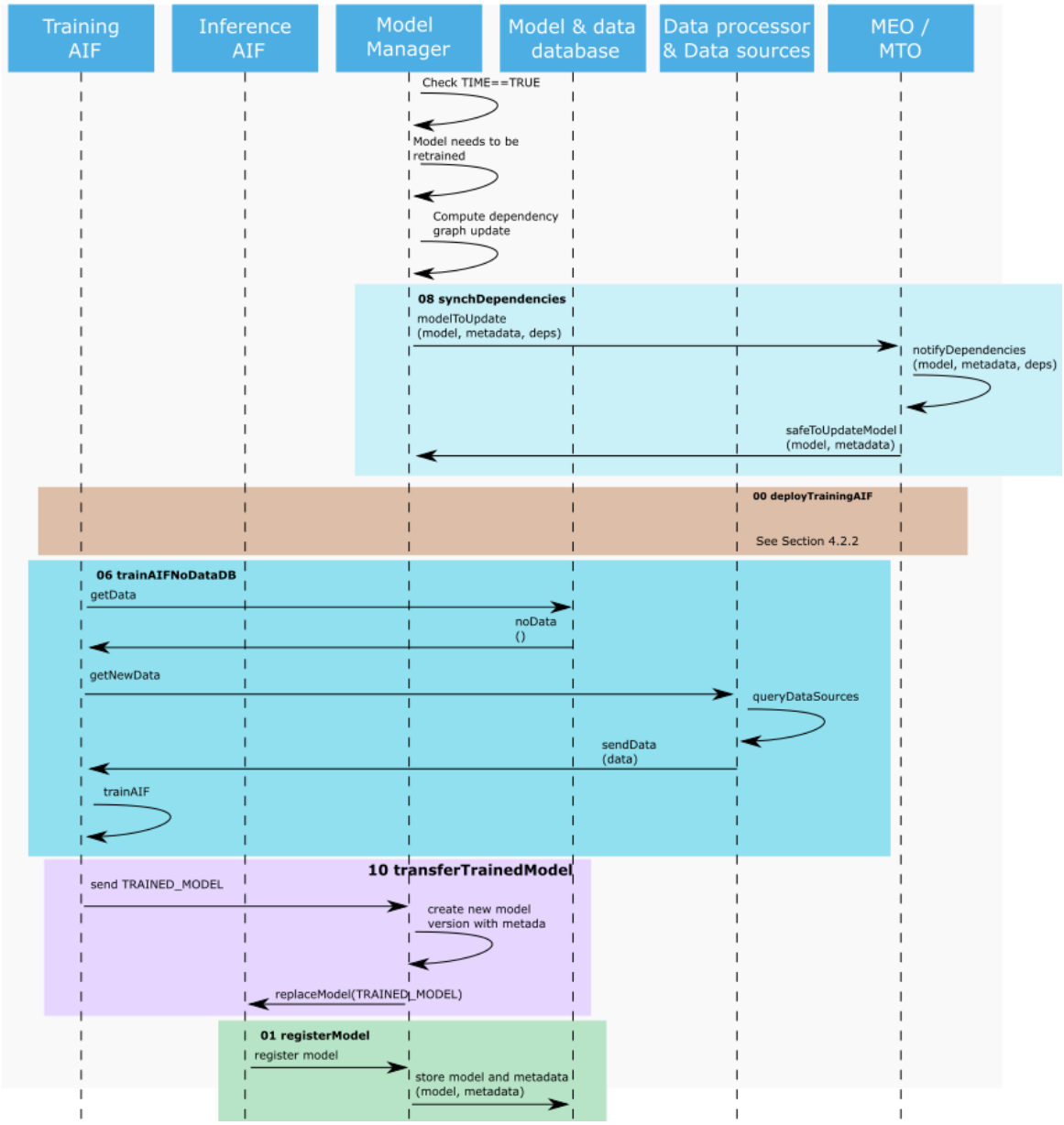


Figure 32 Retraining model periodically and their dependencies. Training AIF is unavailable, and data is unavailable

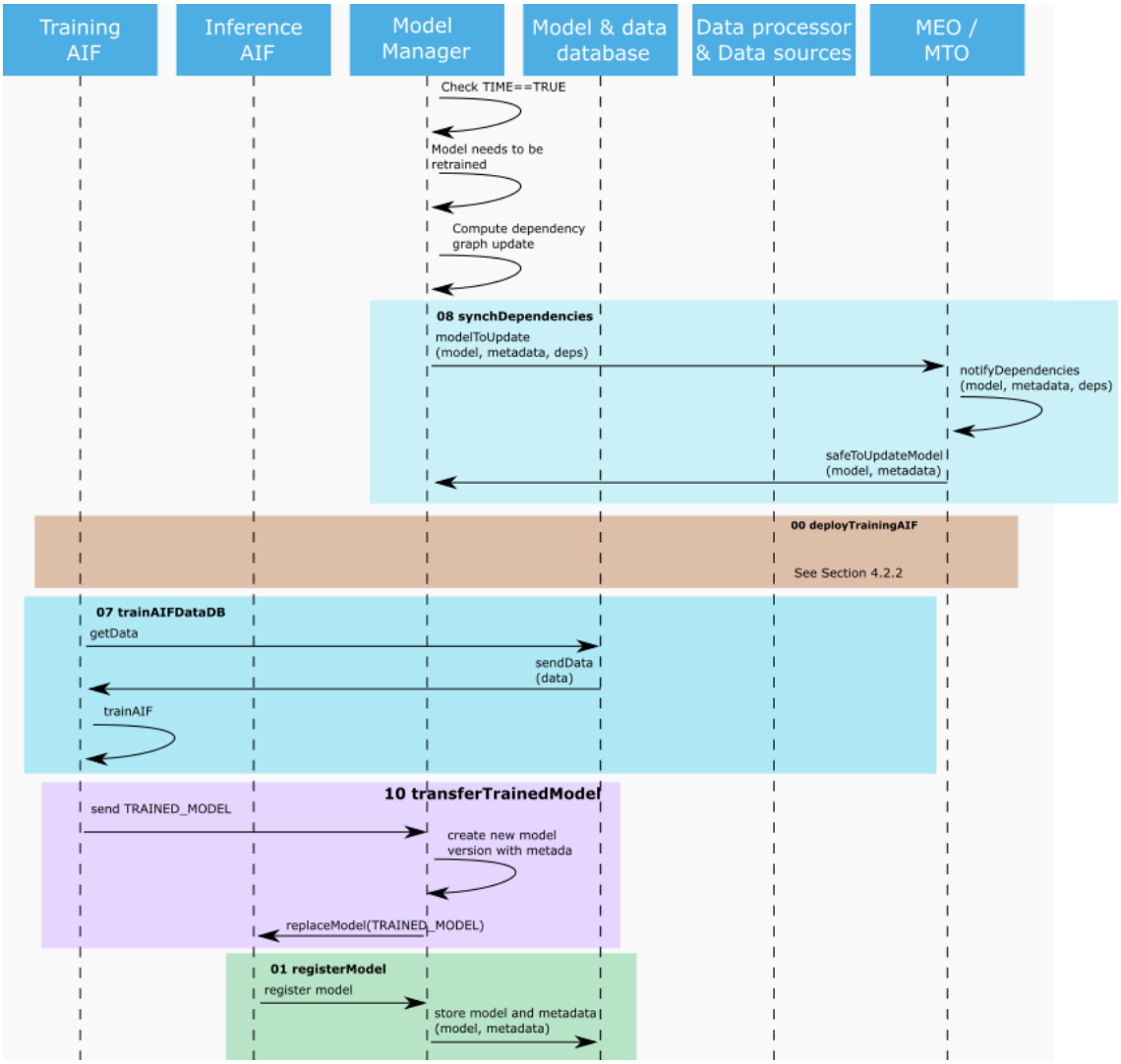


Figure 33 Retraining model periodically and their dependencies. Training AIF is unavailable, and data is available

7.A Retraining model continuously and their dependencies

This scenario is like the one described in section 4.A, but due to continuous training, the AIF will trigger simultaneously the Training AIF and send the prediction, performance, and confidence to the Model Manager. The main difference here is that the Model Manager will continuously synchronize the dependencies with the MEO/MTO at every iteration of the continuous training.

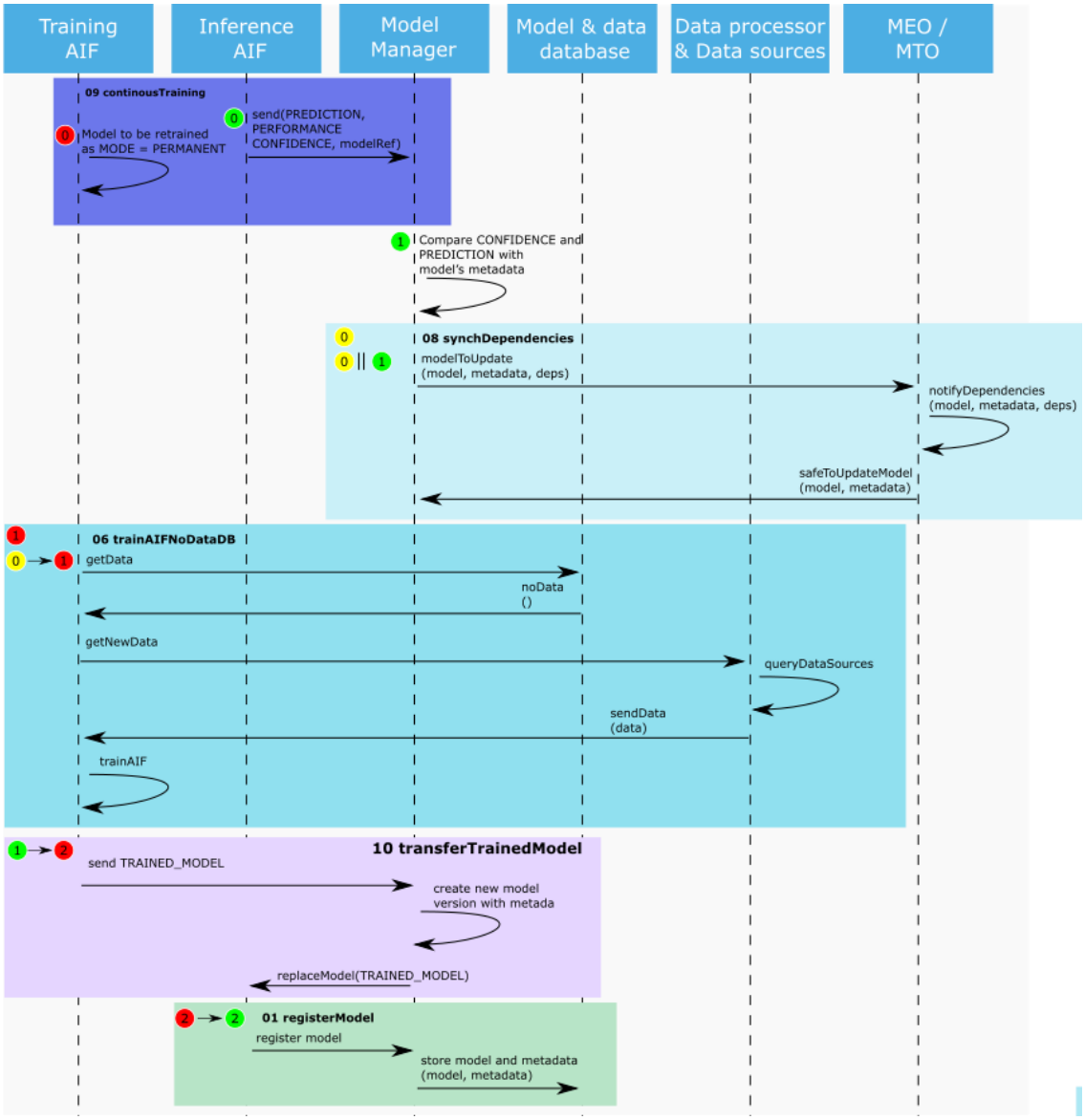


Figure 34 Retraining model continuously and their dependencies workflow. Data is unavailable in database. The synchronization of dependencies must happen before the training of the AIF.

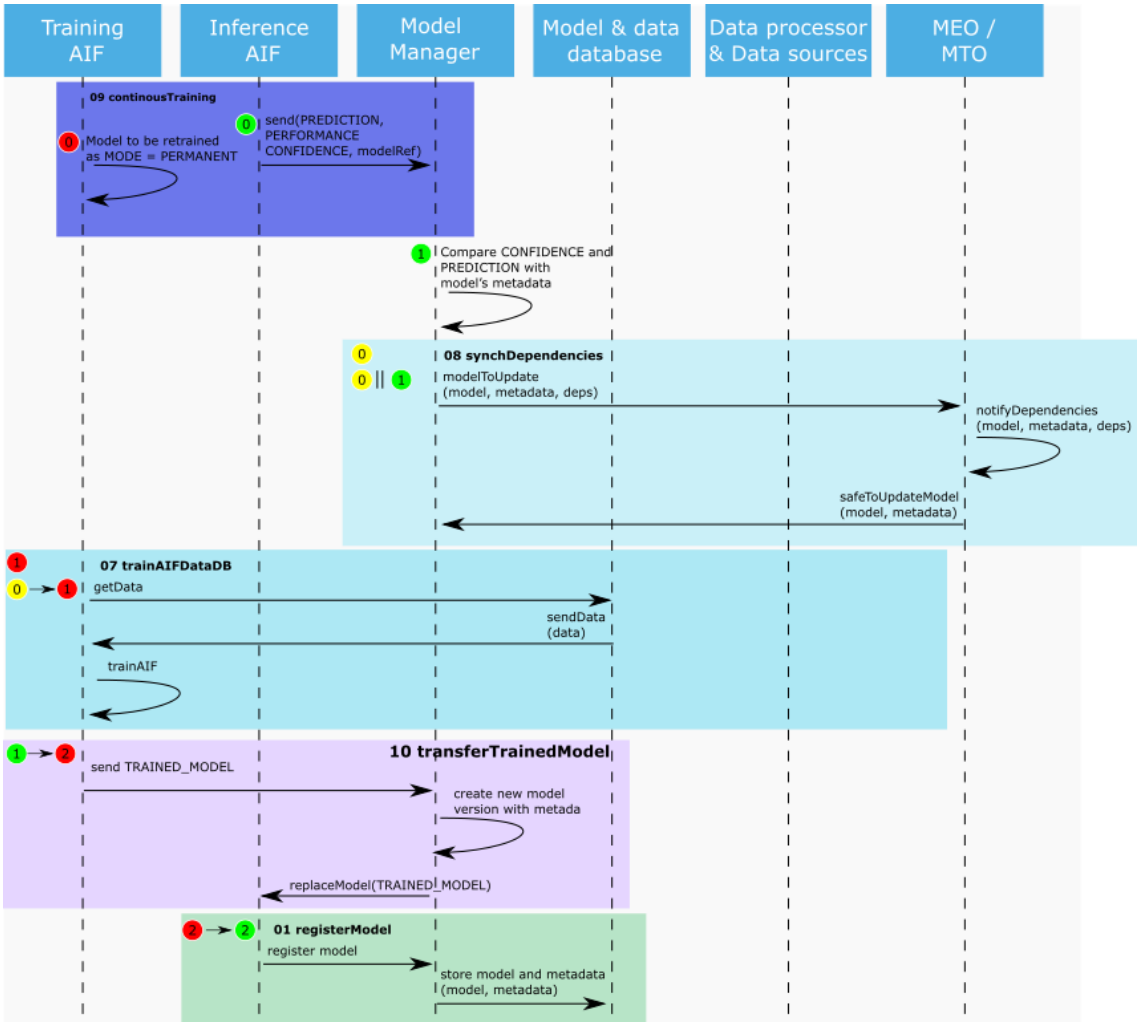


Figure 35 Retraining model continuously and their dependencies workflow. Data is available in database. The synchronization of dependencies must happen before the training of the AIF.

8.A Replace with a new model and their dependencies (compatible model is available)

In this scenario, the Model Manager collects the model metadata to check the model performance by comparing the CONFIDENCE and PREDICTION output values with the model metadata information. This collection is triggered by any MODE (e.g., EVENT, TIME, PERMANENT), as specified in the AIF descriptor. The evaluation done by the Model Manager depends on the values defined in the AIF Descriptor and the presence of a machine learning drift, such as concept or data drift. For example, in a supervised machine learning scenario a threshold for performance can be established; thus, by comparing the PREDICTION and CONFIDENCE the Model Manager can decide to update the model. Another reason to update a model would be detecting the presence of concept drift with the new data.
Once the Model Manager establishes the need for a model update, it will revise the AIF descriptor to verify the presence of the model update boolean flag and if the flag is unset, then the Model Manager will replace the model altogether. If the Model Manager decides to replace the model, then it will send a notification to the MEO/MTO about a model replacement for the Inference AIF. This enables a safe replacement that

will allow continuous procedure with their dependencies. After the synchronization process, the MEO/MTO sends a notification to the Model Manager. Then, the Model Manager checks in the Model & data database if there is a compatible model for replacement. In this case, the compatible model is available, and the Model & data database returns the compatible model to the Model Manager which sends the new model to the correspondent Inference AIF. The execution of the scenario is shown in Figure 36.

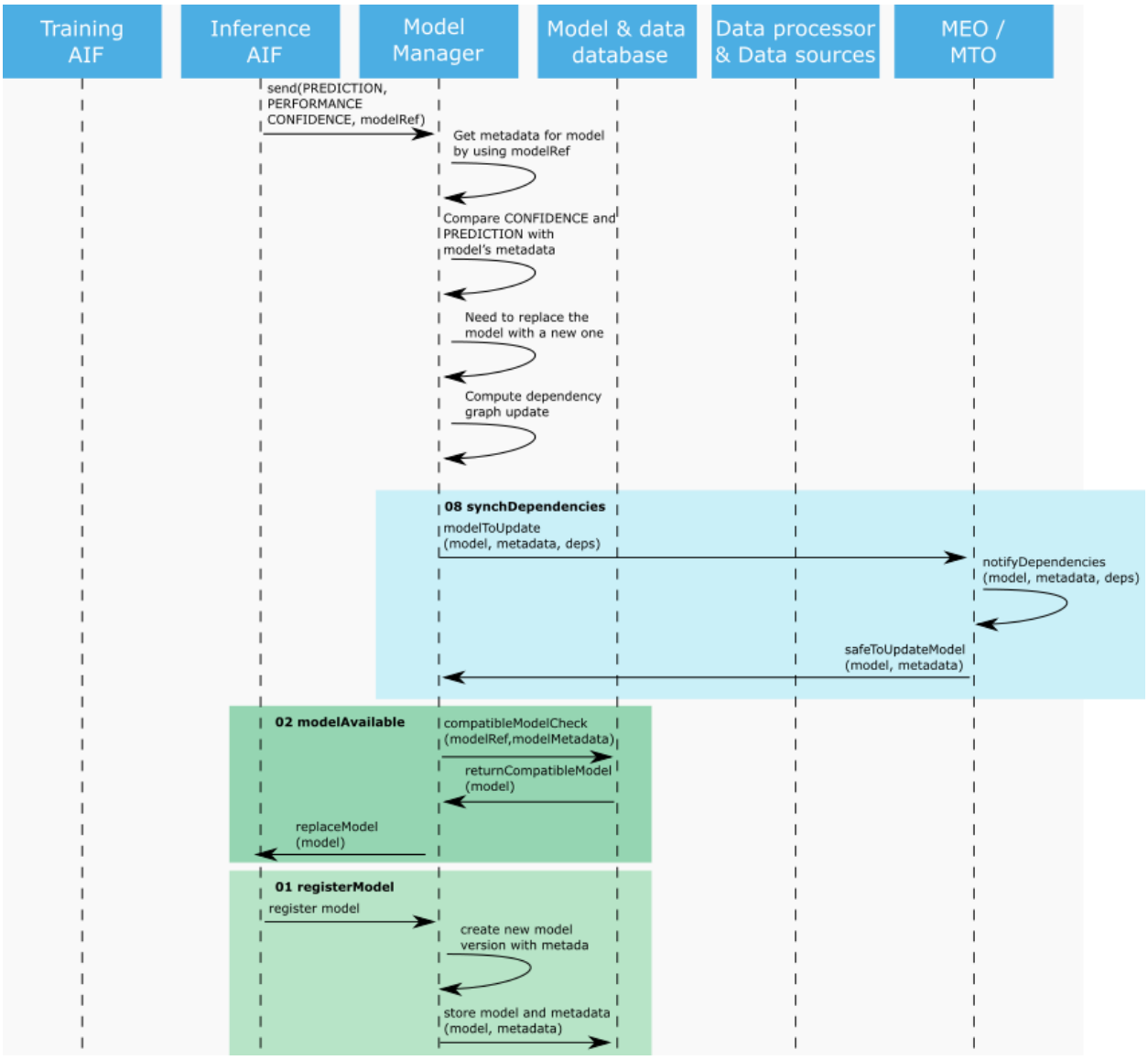


Figure 36 Replace old model with a new model and their dependencies workflow (compatible model is available).

8.B Replace old model with a new model and their dependencies (compatible model is unavailable)

This scenario is like the previous one, but the compatible model is not available in the Model & Data database. To overcome this issue, the Model Manager checks for compatible dataset using the input Source as parameter. However, before this step, the Model Manager will send a notification to the MEO/MTO

about a model replacement for the Inference AIF. This enables a safe replacement that will allow continuous procedure with their dependencies.
After the synchronization process, if the compatible dataset is found, the Model Manager ask the MEO/MTO to instantiate a training AIF using the compatible dataset to train the new model.
If data is unavailable in the database, the Training AIF will obtain new data from data sources, as shown in Figure 37.
If data are found in the database, the Training AIF receives the data without needing to request new data from data sources, as shown in Figure 37.
After this training, the new model is registered in the Model & data database for the Model Manager to keep track of this new model. However, if there is no compatible dataset, the Model Manager asks the MEO/MTO to instantiate a Training AIF to train the new model. The difference here is that the Training AIF will ask for collecting the new data until a train dataset is created. After this training, the model is registered in the model & data database to be tracked by the Model Manager.

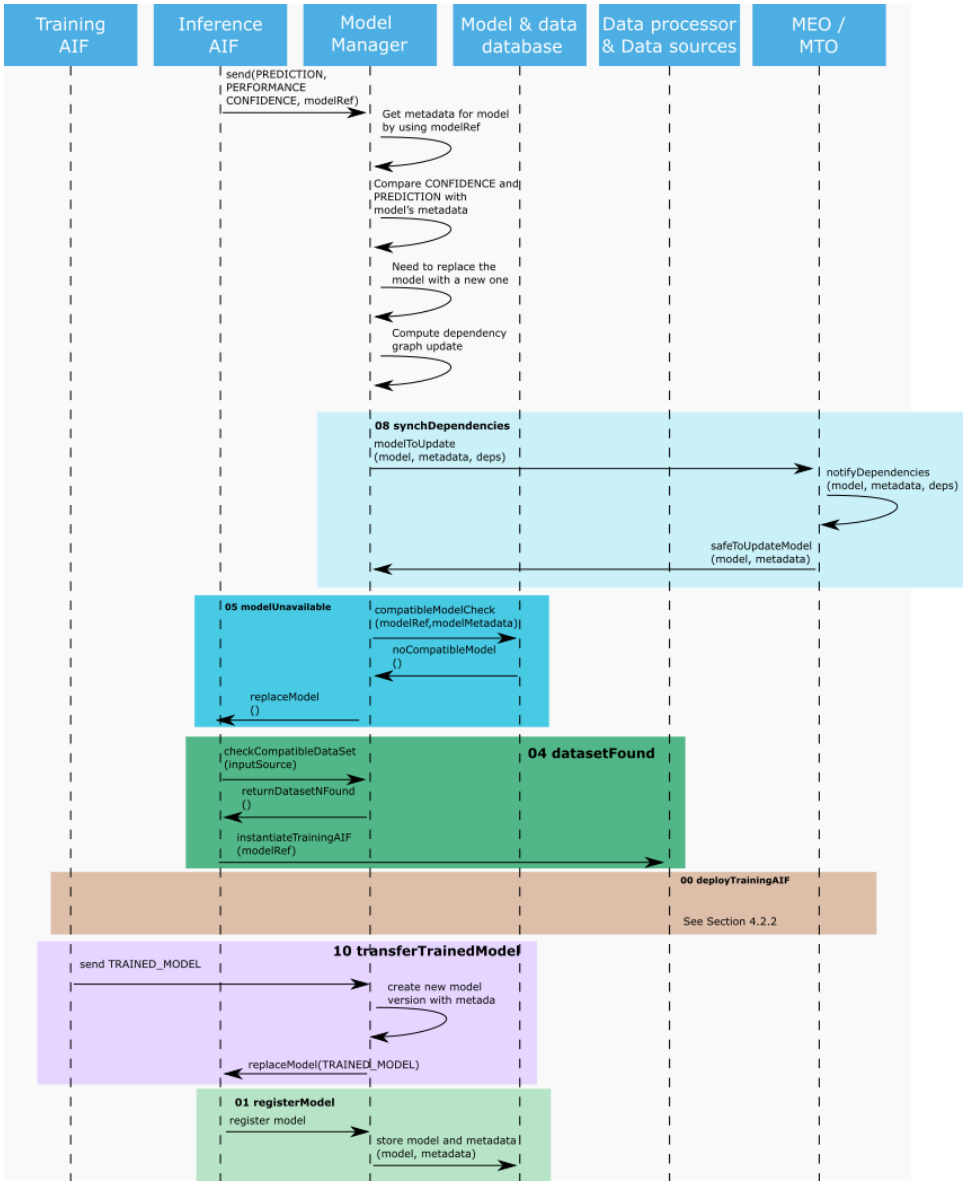


Figure 37 Replace old model with a new model and their dependencies workflow (compatible model is unavailable) and data is available

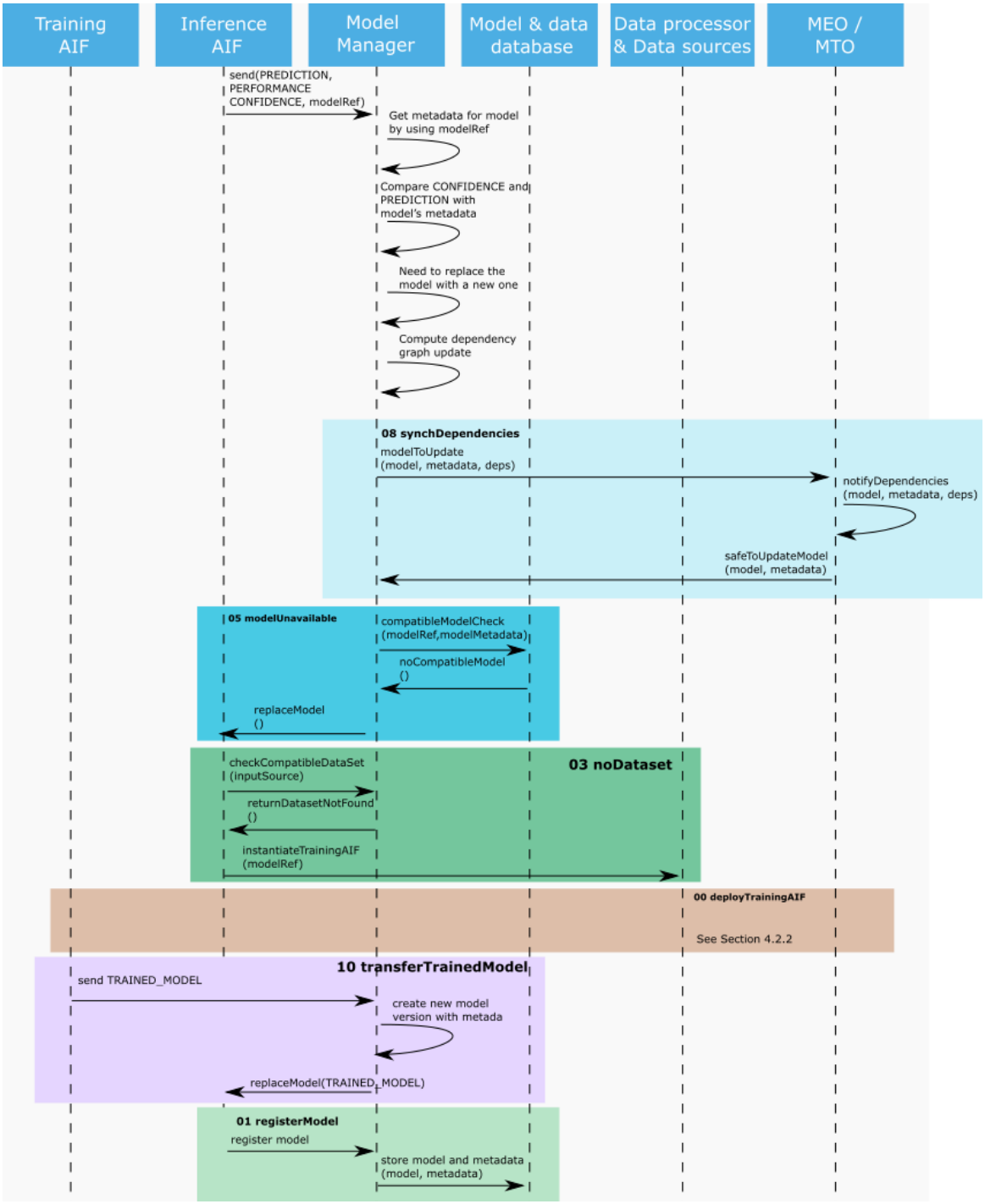


Figure 38 Replace old model with a new model and their dependencies workflow (compatible model is unavailable) and data is unavailable

4.2.2 *MEC workflows*

4.2.2.1. AIF On-Boarding

The process envisioned within the AI@EDGE architecture for application and AIF onboarding is detailed in Figure 39. The onboarding request is performed through the entry point of the NSAP, whose role is played by the Multi-Tier Orchestrator (MTO). Two different approaches can be differentiated in the onboarding process of the AIF Descriptor (AIFD) file. In the first one, the AIFD is just onboarded independently in the platform; by contrast, in the second one, the onboarding process takes place when the instantiation request is received if the corresponding AIFD has not been previously instantiated. More in particular, they can be described as follows:

- Option 1. The Operations Support System (OSS) can store the AIFD directly in the NSAP database, which will return a unique identifier (ID) of the file. This would allow the deployer to have the AIFD pre-onboarded in the system. In deployment time the OSS would have to provide the MTO with the ID in the deployment request to launch the instantiation.
- Option 2. The OSS sends an instantiation request including the AIFD in deployment time. The MTO then, will have to store that AIFD into the database to have control over the System deployments.

In Figure 40 option two is utilized. Notice that in either case, the onboarding process loads the AIFD in the repository maintained at the NSAP by the MTO. However, when this application is instantiated, and a MEC system is selected for this purpose, the corresponding MEC orchestrator loads again locally the AIFD received from the MTO. This allows the MEO to keep a copy of the AIFD to be used, for instance, for AIF migration over the MEO-to-MEO interface.

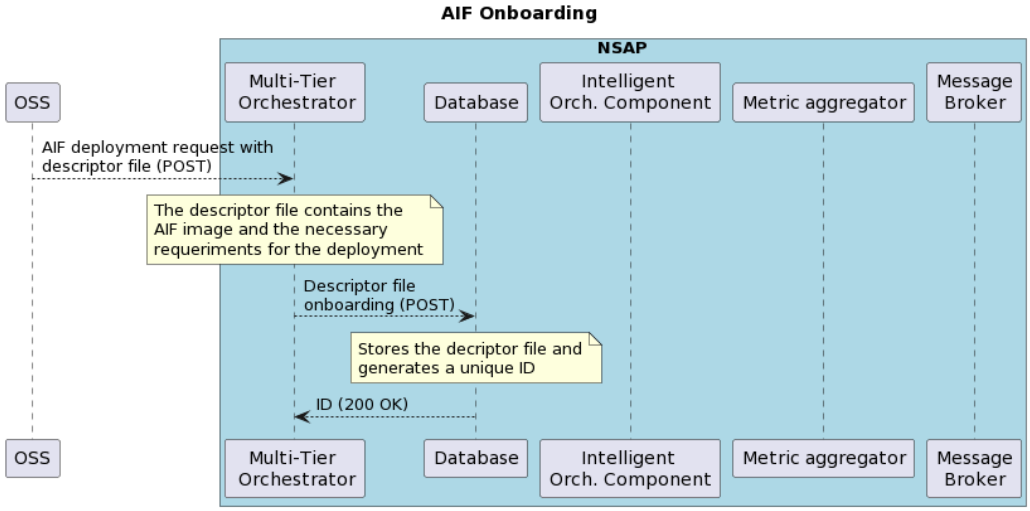


Figure 39 Workflow showing the application onboarding process

4.2.2.2. AIF Instantiation

When a new instantiation request is received at the NSAP, the MTO module then provides the requirements specified by the new application to the Intelligent Orchestration Component (IOC), 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 indicate to which node of a specific MEC system the application must be forwarded through the MTO-to-MEO interface or if it should be allocated into the cloud the first case, the IOC selects the Near Edge node of the MEC System 1 to instantiate the AIF/app. The request is forwarded to the MEO of the selected MEC System. Notice that, as mentioned in the previous subsection, each MEC system stores a copy of the AIFD that is onboarded in the local MEC system's repository once it is selected for instantiation. Figure 40 depicts this process, which 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. The request is forwarded 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 as the one described above but on this occasion the IOC selects the MEC System 2 to instantiate the AIF/app.

In the third 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.

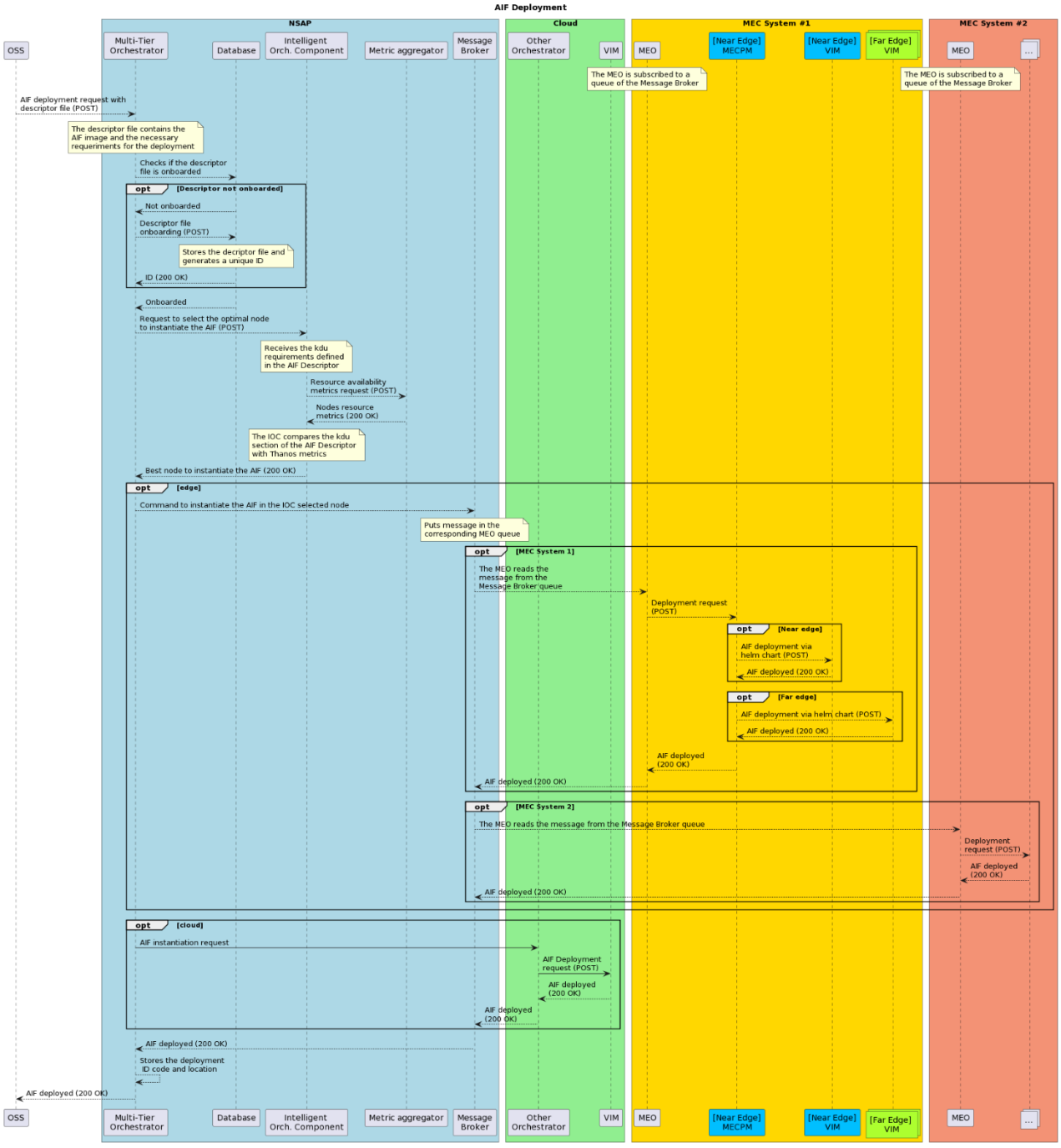


Figure 40 Workflow showing the application instantiation process

When the request received at the NSAP demands hardware acceleration in the descriptor provided, the workflow is slightly different. The hardware acceleration capabilities are handled at MEC system level, where the local IOC selects the most suitable hardware accelerator. The specific process is shown in Figure 41.

The main difference with the previous workflow starts with the output provided by the IOC located at the NSAP. In this case, when detecting hardware acceleration requirements, the IOC limits its decision to a concrete MEC system that can fulfil all the necessities of the application, instead of indicating a specific node. Once the request is received by the MEO of the MEC system selected, it communicates with its local IOC, which will determine the most suitable node depending on the hardware accelerators available and the specific requirements of the application. After this procedure, the MEO has the required information to request the MECPM the instantiation of the application. A similar process would be carried out in the cloud if the site contains the appropriate hardware accelerators, and it is chosen as place for instantiation.

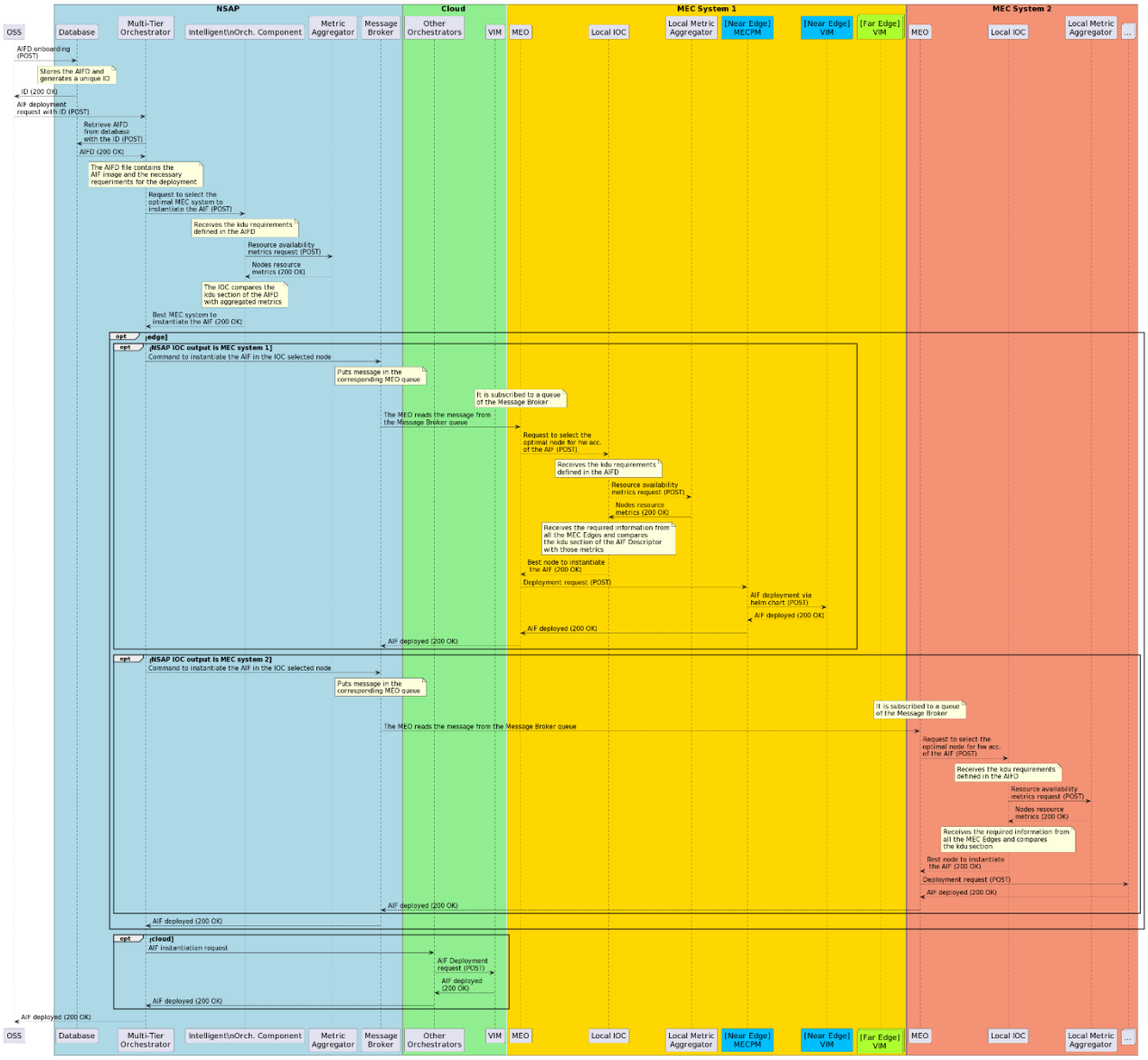


Figure 41 Workflow showing the application instantiation process with Hardware Accelerator

The AI@EDGE architecture also provides network support in the scenario of having an isolated MEC System with no connectivity with the NSAP. The MEC Orchestrator has Standalone (MEO SA)

functionality. This scenario is shown in Figure 42. In this scenario, 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 SA MEO. This module stores the AIFD into a database and forwards the request to the local Intelligent Orchestration Component (Local IOC). The local IOC requests the status information of each underlying system (e.g., available resources and services) from the Metric aggregator. This information is retrieved to make an intelligent decision for the application placement. This decision will indicate to which node of the MEC system the application must be forwarded. Once the decision has been reached, the request is forwarded to the target MEC Platform Manager to initiate the process. 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.

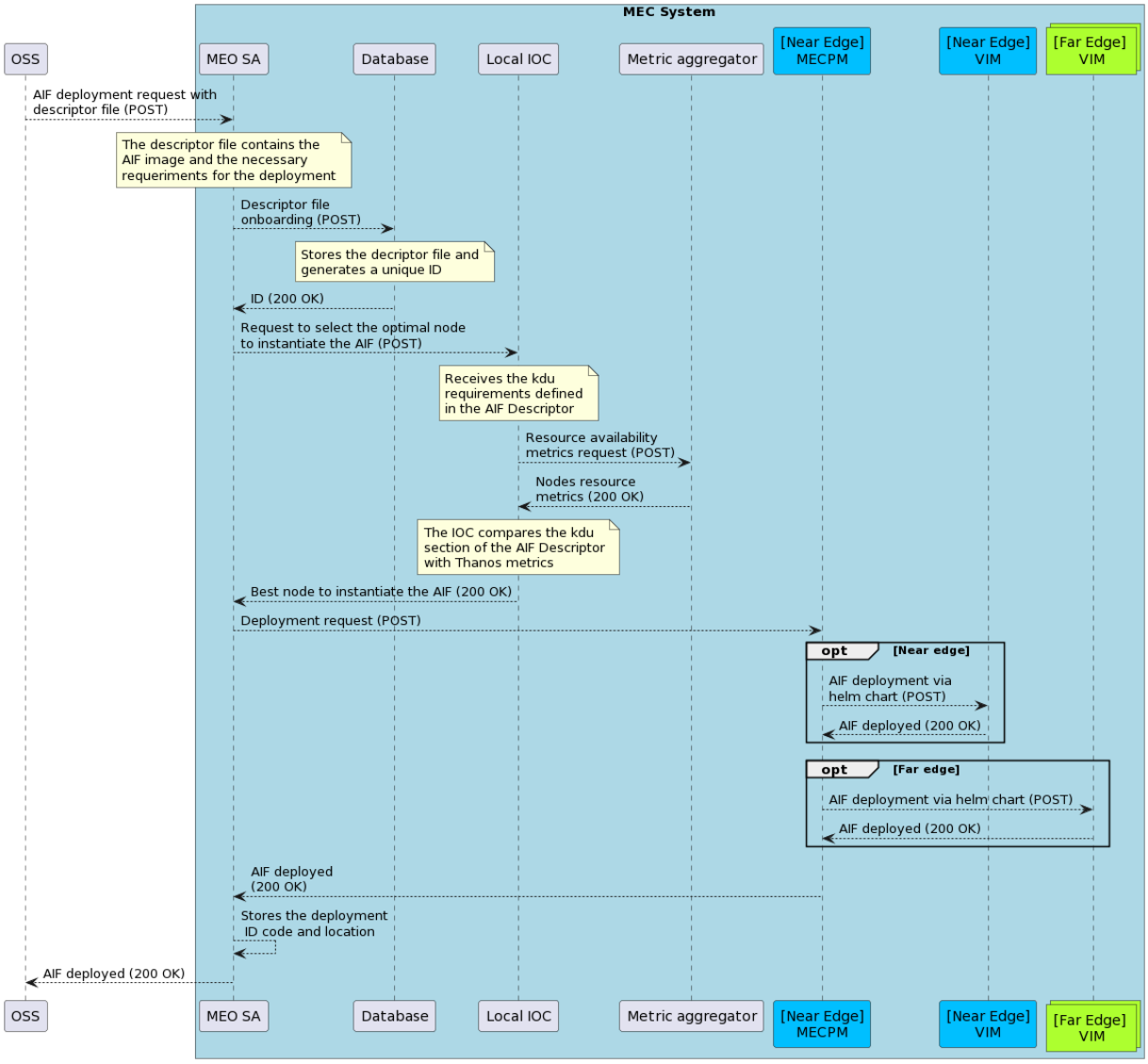


Figure 42 Workflow showing the application instantiation process with only one MEC System

4.2.2.3. Application migration

This section details the workflows considered for application migration processes. AIF migration is defined as the change of the MEC host (and possibly, MEC system) in which an AIF or a MEC application is running. When the migration is performed within the same MEC system, the process is considered stateful. However, when the process takes place across two MEC systems, the migration is stateless, as no prior data form the application is preserved. In particular, three main cases are distinguished: (i) situations in which the application can be migrated to another MEC host within the same MEC system, for instance, due to resource outage, or to the availability of a more suitable hardware accelerator, if required; (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 and (iii) scenarios where no suitable candidate is found within the MEC system and the MTO is not available.

Figure 43 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, for instance, 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. 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 the intelligent module the availability of a MEC platform that satisfies the current needs. Upon a positive response, the MEO 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.

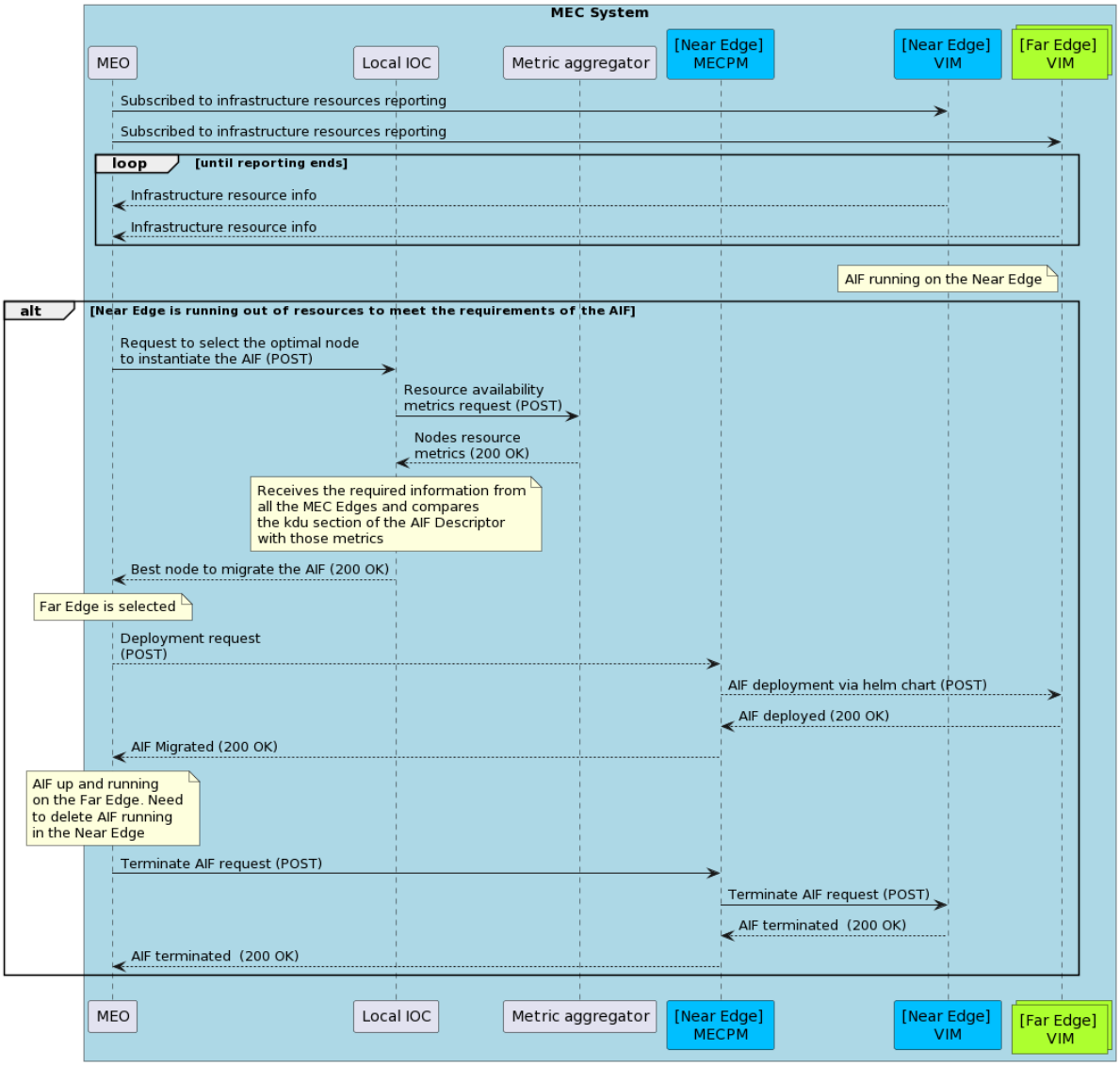


Figure 43 Workflow showing the application migration process within the same MEC systems

Figure 44 showcases the second migration case envisioned, in which the MEC system where the application is currently deployed is not able to satisfy its requirements. 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 inform the MTO of the need of migration through the MTO-to-MEO interface. Then, the MTO will forward the request to the IOC module as in the instantiation scenario. The IOC will determine the most suitable node depending on the requirements of the application. After this decision is taken, the MTO will communicate the decision through the MTO-MEO interface. The receiver MEO passes the request to the MECPM to instantiate the application. Once the AIF has been deployed and running a deletion request is sent to the original MEC System to delete the prior AIF and will free the reserved resources at the VIM.

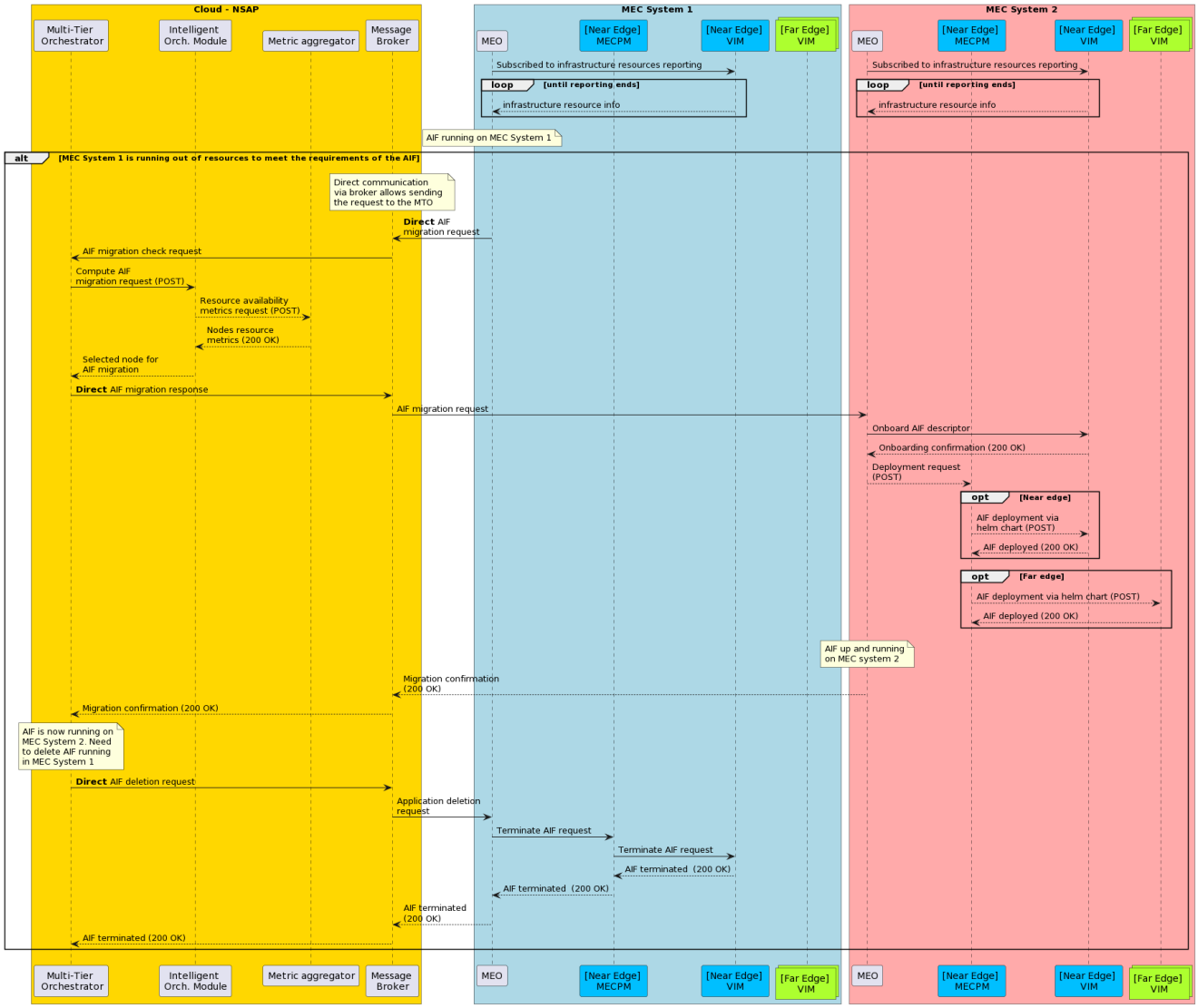


Figure 44 Workflow showing the application migration process between different MEC systems

Figure 45 showcases the first and second migration scenarios taking into consideration hardware acceleration requirements. The hardware acceleration capabilities are handled at MEC system level, where the local IOC selects the most suitable hardware accelerator. In this case, when the local IOC detects a low efficiency in the hardware acceleration a migration request is sent. If any of the hardware available at that MEC System is suitable, the migration is done internally. If not, an external migration is necessary. The MEO would inform the MTO of the need for migration through the MTO-to-MEO interface. Then, the MTO will forward the request to the IOC module, which will determine the most suitable MEC System depending on the hardware requirements. Once the decision is reached the MTO forwards the decision to the destination MEO of the MEC system selected. It communicates with its local IOC, which will determine the most suitable node depending on the hardware accelerators available and the specific requirements of the application. After this procedure, the MEO has the required information to request the MECPM the

instantiation of the application. Once the AIF has been deployed and running the source MEO will proceed with the termination process and will free the reserved resources at the VIM.

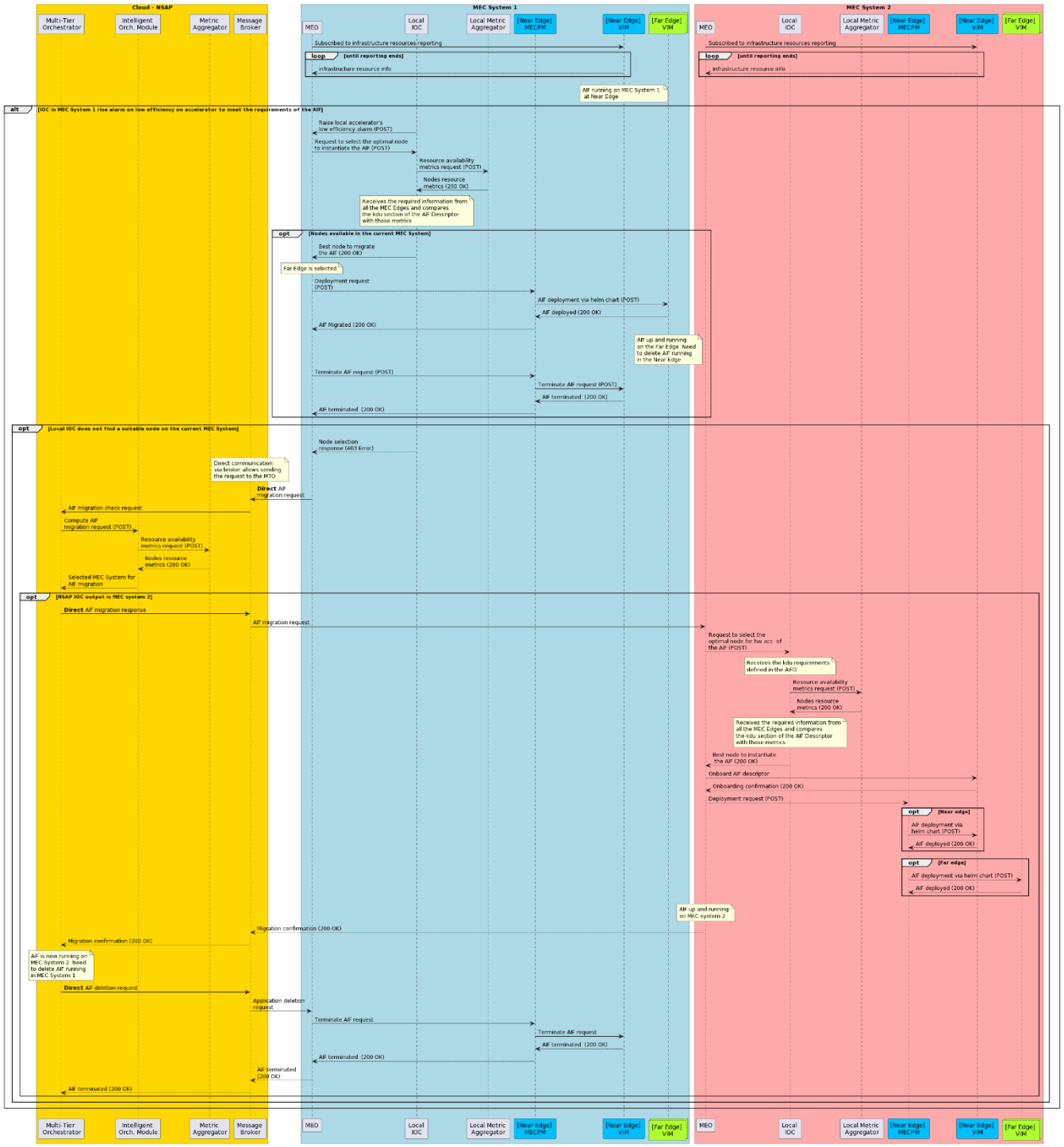


Figure 45 Workflow showing the application migration process between different MEC systems with Hw Acceleration

Figure 46 depicts the third scenario, where no suitable candidate is found within the MEC system and the MTO module is not reachable. It should be noted that the orchestration tasks are designed in such a manner that the MEC systems can continue working independently even if the communication with the MTO or the MTO itself fails. The application is deployed at the near edge of the MEC system 1 and needs to be migrated. To do so, the MEO at the MEC system 1 tries to inform the MTO of the need for migration through the MTO-to-MEO interface. When that connection fails, the MEO sends a broadcast request to the other MEOs available through the broker using the MEO-to-MEO interface. That request asks for the requirements specified at the AIFD for each node of each MEC System. Once that data is retrieved, the source MEO requests its local IOC to select the optimal destination node utilizing the obtained information. Once this decision is taken, the MEO will communicate the decision through the MEO-MEO interface. The receiver MEO passes the request to the MECPM to instantiate the application. Once the AIF has been deployed and running 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.

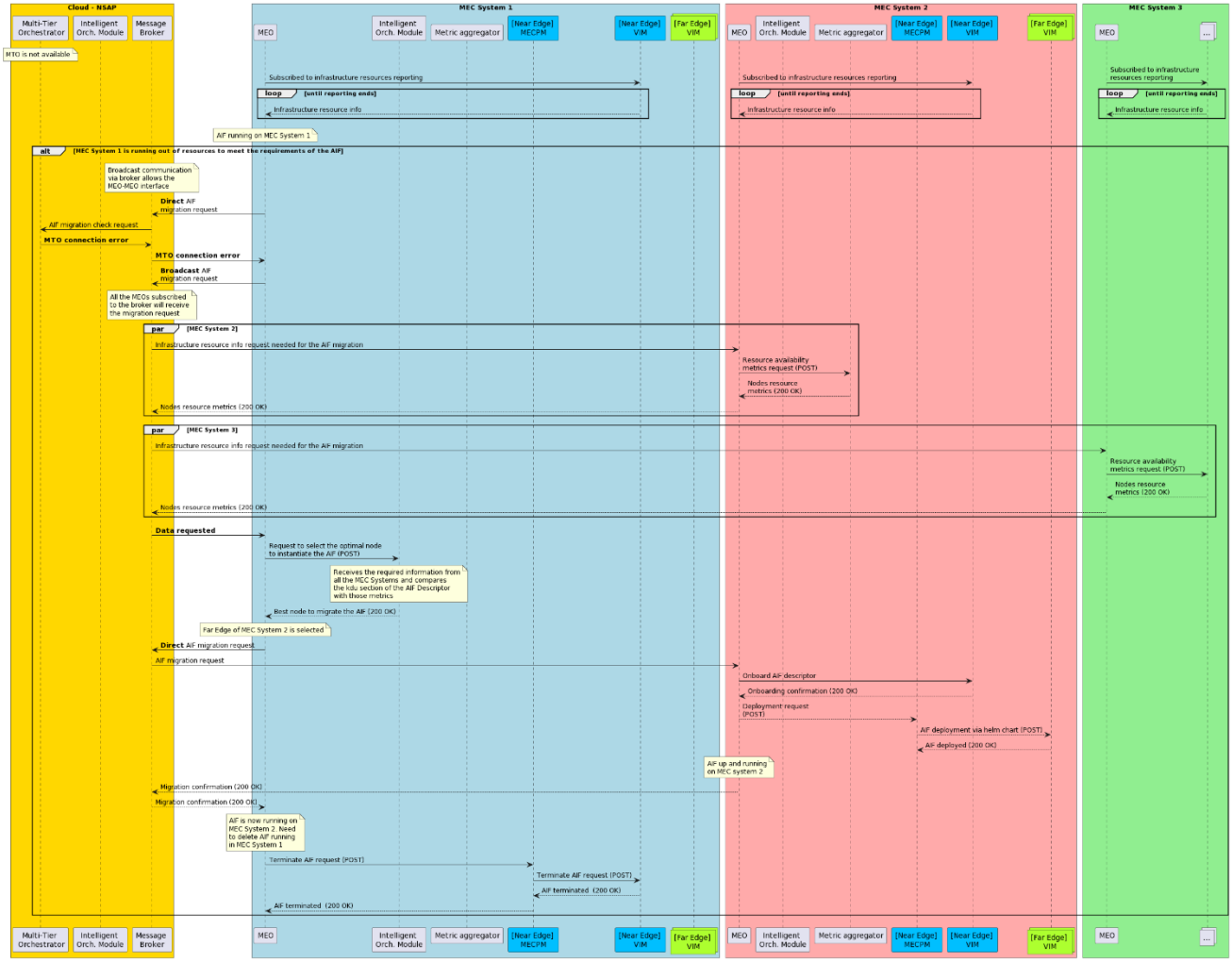


Figure 46 Workflow showing the application migration between different MEC systems when the MTO is not available

4.2.3 *Non-RT RIC workflows*

Figure 47 shows the workflow describing the exchange of data between rAPPs working as consumers and /or producers in order to implement intelligent closed-loop automations by using O-RAN's ICS component as was described in Section 3.1.4.

In the presented workflow, we have considered two rAPPs producing/exposing internal (e.g., RAN data/analytics according to metrics from O1 nodes) and external (e.g., Application data/analytics exposed at the NSAP level) data. Once deployed by the non-RT RIC, the non-RT RIC creates the exposed data types in the ICS and registers the rAPPs as producers of these data types. Then, the non-RT RIC deploys the rAPP 1, which needs this data to perform the optimization of the RAN or to generate a prediction according to an AI/ML algorithm; therefore, the non-RT RIC creates in the ICS specific jobs for serving this data to rAPP 1, defining parameters like the communication method (e.g. REST API, Kafka bus), the reception interval or the processing methods (e.g., Prometheus operations). The ICS notifies the producers about the new jobs and their related parameters, enabling the sending of the required data to the consumer rAPP. Using this data, rAPP 1 can perform an optimisation of the RAN through the O-RAN interfaces exposed by the non-RT RIC (e.g., A1 or O1), as shown in the first option. Alternatively, as depicted in the second option, the rAPP could generate AI-driven predictions, which could also be exposed to other rAPPs through the ICS.

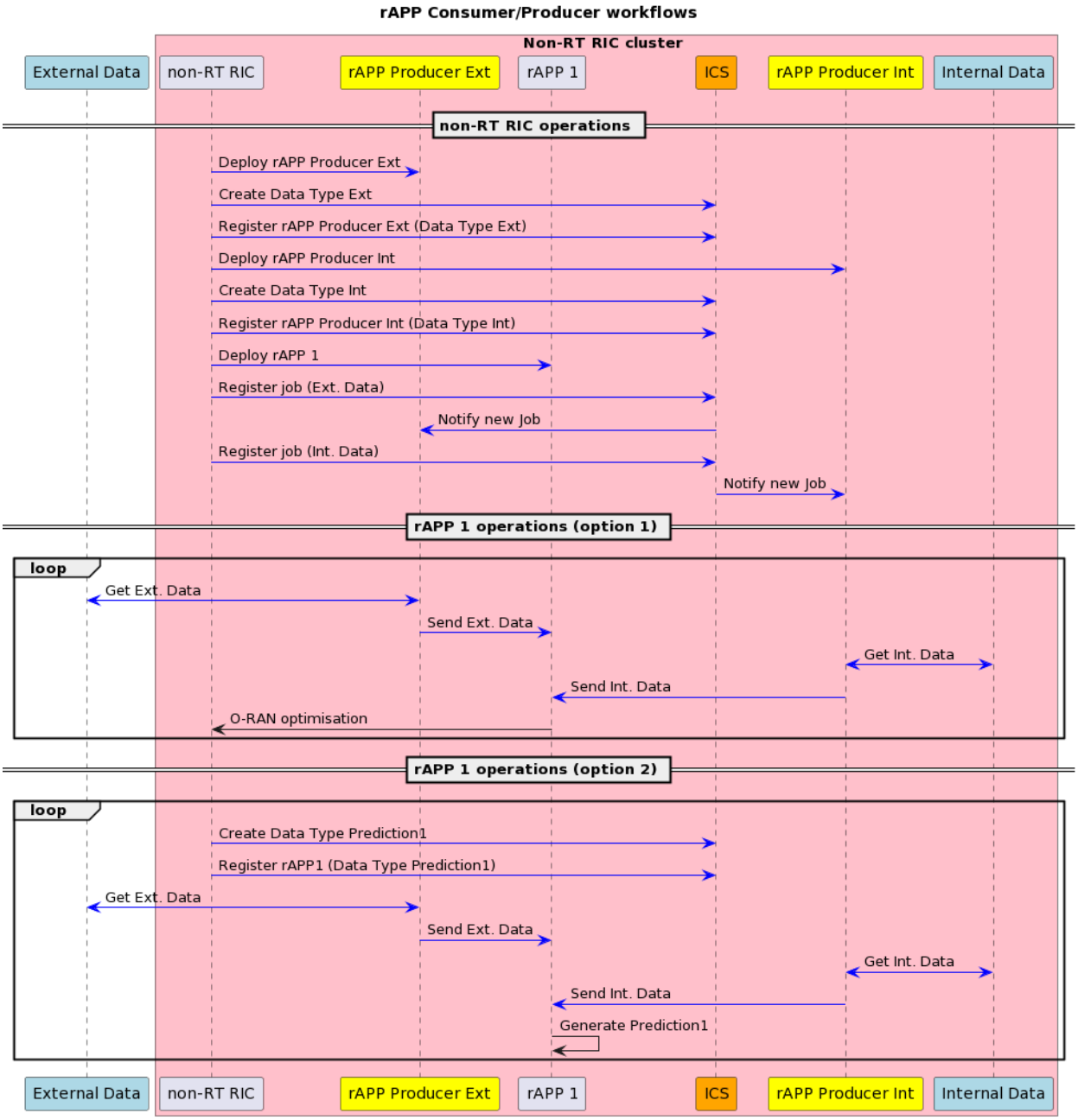


Figure 47 non-RT RIC related workflows: rAPPs as data consumers/producers

5 Dialogues, Drivers and Techno-economic Analysis

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, Transportation, European food supply, Europe's electric power system and European indigenous minorities, in particular the needs of the nomadic Sami people. In section 5.1 we report on updated broader insights drawn from discussing with each category (already introduced in deliverable D.2.2), even without drawing detailed technical conclusions.
In Section 5.2 we introduce Total Cost of Ownership (TCO) as one of main drivers for techno-economic analysis and in Section 5.3 we introduce a general methodology for techno-economic analysis including four consecutive and iterative steps.

5.1 Updates on dialogues

5.1.1 *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. The pandemic took a grave toll among the chronic multi-diseased elderly and the national healthcare systems in Europe were given a lot of extra funds, both in effect postponing any crises resulting from the demographic challenge. Nonetheless, the fundamental problems remain with overwhelming costs and difficulties to recruit enough staff.

The envisioned key component in a solution seems to be self-care, in particular data-driven technology-supported self-care. This self-care would 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 and sensors 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. A number of promising wearable devices seem to be able to serve a multitude of chronic diseases with a single device. 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. Many patients would be able to continue to live as if they were healthy for many years after they got their diagnoses.

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, or even prevent them from falling. 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.

There are also other use cases for edge computing related to e-health, besides the ones associated with the demographic challenge. During our dialogues we encountered a growing use of VR (virtual reality) and AR (augmented reality) glasses for rehabilitation. An interesting example was rehabilitation after neck injuries where traditionally the patient gets a paper with instruction for exercises to do at home. And, equally traditionally, the patient does none or few of these. If instead given a VR headset, the exercises can be done together, but remotely, with a nurse over the Internet, progress can be measured automatically, and feedback can be given to encourage the patient. Any medical analysis of the exercises can be done by edge computing.

Another use case of connected wearables is for triage. It would be useful for ambulance staff when arriving at a disaster scene or accident to put say bracelets on all injured to assist in quickly determining which need the most immediate attention and why, and to monitor the injured to see if their state changes. Many of the parameters needed for this are the same as to track changes in the state of a chronic disease. Edge computing could handle the medical analyses and assist with positioning to know where the wearables are (and the patients wearing them).

5.1.2 *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 and sensors 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. This is one of the main differences between urban and rural living.

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 option 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) in Sweden, one of Sweden's state agencies responsible for the national food supply and crisis management, they 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 that 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, with the option to switch between these two modes, as they will require human assistance to resolve difficult or unforeseen situations. These 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. Edge computing may also be necessary to handle situations like when an ambulance is to pass through roads with autonomous vehicles or a helicopter to fly through airspace with drones. The edge compute fabric will be central to this complex flow of information and data analysis. Like in the e-health use case, edge computing is again seen as a crucial enabling technology.

5.1.3 *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. These perspectives are hopefully exaggerated but edge computing can play an important role in diversifying and robustifying our food supply.

In the dialogues the project has had, two main lines of thinking and action 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 crop 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 like the autonomous transportation system envisioned in Section 5.1.2 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. This could be complemented by high-tech small-scale farming, making use of that technology such as digitalization and AI on one hand, and devices on the other, are getting cheaper. In the same way as self-care in e-health is made possible by cheap data analyses and devices, small scale farming could also become a possibility. Blue-print concepts with names like "the one-acre farmer" could lead to a wave of small-scale farming diversifying food supply.

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. Off-grid and small 5G and 6G systems, like the ones discussed below to be used by the Sami (Europe's probably only indigenous minority), are also suitable for deployments in archipelagos and could serve as a basis for remote aquaculture, in particular if they also have edge computing capabilities.

5.1.4 *Europe's electric power system*

Although deeply troubled by the contribution to climate change and the current energy crisis spawned by Russia´s invasion of Ukraine, strategists in the electric power area seem to have a generally positive outlook on the future. The coming years will see much turmoil in the energy sector with volatile markets and prices, but also a rapid and determined journey away from fossil fuels and thus towards a greener and more sustainable society. This journey will contain a strong build out of infrastructure for generating electricity, but the traditional centralised architecture (based on fossil fuels) will give way to a decentralised architecture based on solar power, wind power and other more local production. An enabler for all this that is largely lacking in the public and political discourse is the price of batteries. The battery experts of Volvo and Polestar noted that the price, size, and weight of batteries had dropped a factor of five in five years and that it is doing so again in the five years to come. This changes a lot in the European energy system as any household or industry will be able to smoothen their load on the grid in a few years from now. Power will no longer be such an issue, only energy.

In this time perspective, the electrical power system will 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 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-fuelled electricity plants, nuclear power plants and hydro-power plants, had a stable output determined by the operators, 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 kilometres 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. In this context, the concept of a virtual power plant is much discussed. A virtual power plants are many small production facilities that are acting so perfectly synchronized that they appear like a single large power plant as seen from the network. This is a task well suited for edge computing. A high level of synchronization can give smaller producers access to markets and prices reserved for large powerplants. In Sweden any plant of 5 MW or more can bid on the so-called frequency control market as ancillary services, which is more lucrative than just selling energy.

5.1.5 *European indigenous minorities – the Sami*

The rural populace of Europe in general, and European indigenous minorities, 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, and we foresee that edge computing will become one. In this context, we worry 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 law 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 sufficiently far out to cover many of the areas where the Sami live and work. In particular, their 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 even further. European society needs to try harder to bring rural areas and minorities into the sphere of digital services that are becoming increasingly important for everyday life.

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 kilometres 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. As far as the project can determine, there are no umbrella-cell-products today that also offer edge computing.

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. ("Fulltäckning" translates to "complete coverage".)

Satellites, especially low-orbit satellites, are also a promising technology when it comes to providing cellular coverage and mobile broadband in rural areas. It could become the final solution to the rural coverage challenge, especially in the north of Europe as the satellite orbits normally converge there. 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, but projects like AI@EDGE and successors could perhaps solve this challenge.

Figure 48 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.

5.2 Main 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 our techno-economic analysis. A techno-economic analysis and work examines primarily costs, benefits, risks, uncertainties, and timeframes to evaluate the attributes of technologies developed and produced in the project. The economic performance of the solutions is calculated considering 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. However, we do not reach that level inside the project, although it is approximated in dialogues with external partners.

In order to evaluate the economic viability of the selected scenarios (use cases) we build a generic TCO. Our 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 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 49.

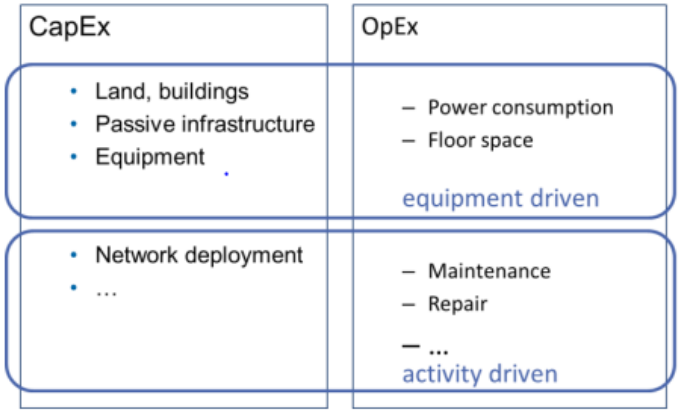


Figure 49 Cost classification

Specific TCO models will be described and developed for each use case in D2.4 in parallel with the development and testing of the use cases (described in D5.1).

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 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 analyse 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.

5.3 General Methodology for Techno-economic Analysis

We have already introduced the parameter of Total Cost of Ownership (TCO) as the main cost element related with every techno economic analysis. With a clear focus on business models (Subsection 5.3.1) and in the context of business analysis, we are proposing a general methodology for techno-economic analysis that consists of four main consecutive and iterative steps as shown in Figure 50. The proposed global methodology includes:

- Step 1. Business models and Value network identification: this can be achieved by firstly defining the different business roles and stakeholders involved, and secondly the interactions between them. Different ways of interactions result in different value network configurations, and accordingly in different individual business models for the stakeholders. This is described in Subsection 5.3.1. and 5.3.2
- Step 2. Business case viability study (TCO): A Total Cost of Ownership (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. Relevant techno economics and some mathematical formulas are introduced in Subsection 5.3.3.
- Step 3. Impact of new technologies on business case: this step consists in investigating the impact of innovative technology on the business case. We are showing the benefits and advantages of the AI@EDGE proposed platform and relevant new technologies in Subsection 5.3.4.
- Step 4. Sensitivity analysis: this step is elaborated to assess the degree of uncertainty (especially this period with high inflation) that links the model outputs to the inputs. This will be reported in the next deliverable D2.4.

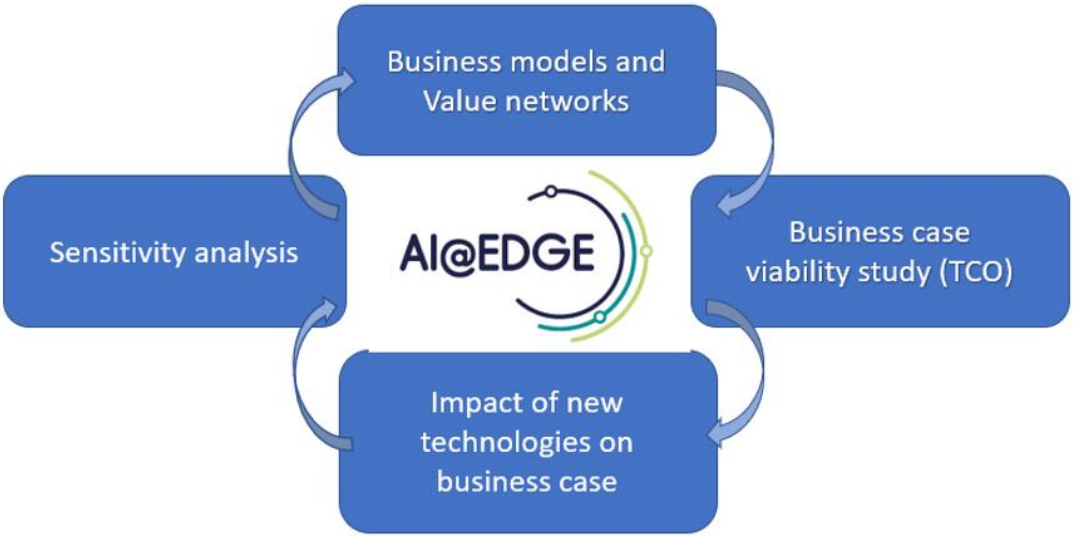


Figure 50 Global Methodology - Techno economics as part of business analysis

In the next sections we proceed with the description and analysis for each of the four steps/pillars of the proposed global methodology for techno economics analysis as part of business analysis.

5.3.1 *Step 1a - Introduction to business models with a focus on 5G / Beyond 5G ecosystems*

Business models have been proposed and developed for both firms and ecosystems. We begin with the definition of the business models for firm (and how it is related with the firm's strategic plan), then we define the 5G ecosystem and propose relevant models for the 5G ecosystem.

The business model definition of each firm ideally is the result of a nested process evolution and development as illustrated in Figure 51. The process starts from a firm's strategic plan that explains which specific long-term goals a firm expects to achieve and how. From its strategic plan the firm derives the business strategy, which sketches the steps needed to achieve the long-term goals. The next step in the process is the definition of a business model which identifies value propositions, customer segments, and concrete steps for the execution of the business strategy and describes how a firm creates, delivers, and captures value in economic, societal and other contexts. Embedded in its business model each firm will define one or more business plans and formulate associated business cases that support the execution of the business strategy.

Figure 51 Business model as part of each firm's strategic plan

Specifically, in fast evolving 5G and beyond 5G market environment (with variable network deployments, products and services and competition of alternative technologies), this process is not so straightforward. The firms need to adapt to competitive pressures through sustainable and reliable business models. In the context of 5G PPP projects, the white paper [33] on *Business Validation in 5G PPP vertical use cases*, proposes a flow of business development activities in 5G and beyond 5G ecosystems to guide the firms participating in these projects on positioning themselves in the ecosystem value network. The business development activities are inspired by lean start-up methodologies and are segregated in four phases (Figure 52):

Customer validation, Solution alignment, *Business Model*, and Growth Trajectory, contributing to a strategic plan for the company (upper layer in the Figure 51 Business model as part of each firm's strategic plan). So, business model in 5G markets is as one the four pillars on business validation in H2020 vertical 5G use-cases.

Figure 52 Business validation for H2020 vertical 5G use cases [33]

Similarly, we assume herein as well, that no firm alone can deliver the full value to the customer. Therefore, before designing a business model, it is necessary to align the understanding with partner firms on how to jointly present a value proposition and deliver a solution which properly meets customer's expectations. When starting with the business model design, the single firm must have realised the dependency on other firms and ruled out the possibility of serving the customer alone.

Before we proceed with the business models for 5G ecosystem, we need first to define such an ecosystem. This white paper [35] provides definitions and early examples of 5G ecosystems, aiming at equipping 5G stakeholders in telecommunication and vertical industry sectors with better understanding of ecosystem dynamics, the processes that take an ecosystem from birth to maturity, and the kind of strategies that are necessary to kick off its evolution. Not the least, they wanted to emphasize that an ecosystem does not evolve and reach volume without potential tensions between stakeholders, which call for the need of balancing strategies and interests, hurdles mitigation and consensus creation. This white paper elaborated on the 5G ecosystem from two perspectives: the provisioning 5G ecosystem, and the 5G vertical ecosystem. In emerging 5G ecosystems, driving firms' strategies must focus on how to mobilize other contributors to take part in value creation. More specifically:

- 5G provisioning ecosystem encompasses those roles and actors who take part in developing, delivering, and providing AI/ML powered 5G services. Traditionally, the telecom industry is seen as a value chain where network operators source the resources necessary to provide fixed and mobile telecommunication services.

- 5G vertical ecosystem black boxes the 5G provisioning ecosystem and focuses on other actors who work closely together as part of vertical industries. While roles and actors from the telecommunication sector are still present in this ecosystem, the emphasis is on yet other roles which apply 5G services in their value creation and can be domain specific.

Since **AI@EDGE is mainly targeting (but not limited to) vertical industries**, we are focusing on 5G vertical ecosystem and the relevant roles. The roles of AI/ML in a 5G vertical ecosystem are part of disaggregation of the 5G Service Customer role in the 5G provisioning ecosystem [35]. A first separation is between the role of the 5G Vertical enterprise customer which purchases 5G services, and the role which support the vertical enterprise customer to create and operate a solution in the vertical domain. While 5G

Service Customer is a customer which pays only for the 5G services (ad hoc solution), the 5G Vertical enterprise is an enterprise which has adopted a 5G Solution, specialized for the needs of the (vertical) enterprise (permanent solution).
The 5G service provided by a 5G Service Provider is one component in such a solution. Thus, seen from the 5G Service Provider side, the supporting role complements a 5G service and the role may be referred to as a complementor. Furthermore, this complementing role consists of many more specific roles, and we therefore refer to the main role in plural – 5G Vertical complementors. It should be noted that the complementors are not only seen as providers of components in AI/ML and 5G empowered solutions; in an ecosystem context complementors are seen as critical holders and developers of knowledge which in turn is the basis for innovation in the vertical domain.
The first two proposed common business models are Security-as-a-Service and AI-as-a-Service.

5.3.1.1. Security-as-a-Service as a potential, common business model

The main goal of AI@EDGE is to build a platform and a set of accompanying tools for enabling secure and automated management, orchestration, and operation of AI-powered services over edge and cloud compute infrastructures, with close to zero-touch of the underlying heterogeneous MEC resources (network, storage, and compute resources).
It is obvious that the overall security of the connect-compute (platform), based on elements with different security levels, should be evaluated at different stages of development. Based on different categories of the CCP (from low to high), different services based on different security levels might target to different potential customers. Such a business model, common to all use cases might be named as a Security-as-a service business model.

5.3.1.2. AI-as-a-Service as a potential, common business model

Additionally, to Security-as-a-Service business model, the AI@EDGE platform will be able to offer AI as serviced for business solutions targeting different type of customers depending on the AIFs which are offered to all use cases. In this case, the business model is called AI-as-a-Service.

5.3.2 *Step 1b - Value network analysis and first models for Use Cases (Vertical Ecosystem)*

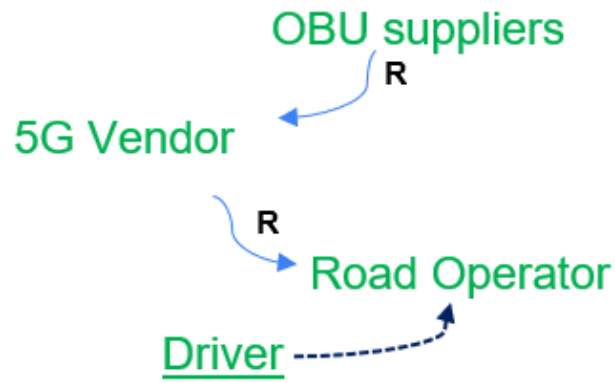
Value network analysis (VNA) is a methodology for understanding, visualizing, using, optimizing internal and external value networks and complex economic ecosystems [36]. "*The methods include visualizing sets of relationships from a dynamic whole systems perspective. Robust network analysis approaches are used for understanding value conversion of financial and non-financial assets, such as intellectual capital, into other forms of value. The value conversion question is critical in both social exchange theory that considers the cost/benefit returns of informal exchanges and more classical views of exchange value where there is concern with conversion of value into financial value or price*". The proposed **e3value methodology** is a stepwise approach to develop business models for networked value constellations. These constellations are networks of enterprises who offer something of economic value to end users. Networks consist of end users (the customers), suppliers, and the suppliers of these suppliers. The e3value approach supposes an ideal network, in which all actors behave honestly and the e3fraud method can be used to analyze sub-ideal behavior, e.g., actors committing a fraud. More specifically:

- Tangible and intangible value streams between all partners: The key point of Verna Allee's Value Network Analysis (VNA) is the shift from a linear value chain with only a few partners to a more complex value network [38] In these value networks intangible value streams are equally important as

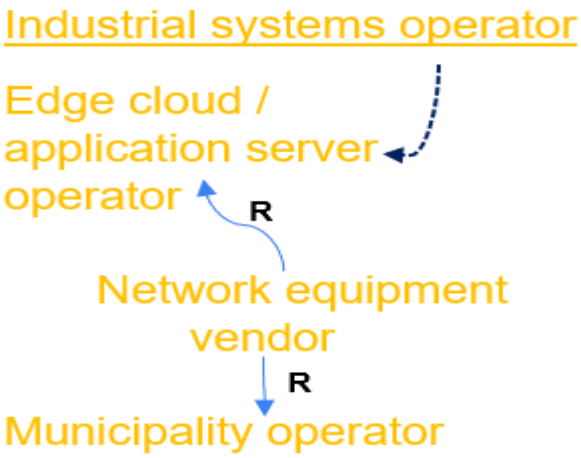
tangible value streams in order to create value. Tangible value streams are contractual like goods, services and money. Intangible value streams, however, are knowledge or intangible benefits which support a product and are not contractual. VNA visualizes both types of value streams between all involved partners. This is in contrast with the e3 value model, which mostly neglects intangible value streams [38]

- Activities or roles without value streams: Each actor can perform roles or activities in order to create value and the mapping of roles on their actors is referred to as a value network configuration. A value network configuration is not unique since different actors could take up different roles. We argue that the value streams in a value network configuration should be visualized between actors, as applied in VNA [37]. In contrast to the e3 value model, value streams between internal activities are not depicted since they do not represent a transfer of ownership [38].
- Economic viability through scenarios: All the compared frameworks recognize the usefulness of scenarios to capture possible future changes of the value network. These scenarios represent different versions of a value network based on giving a value to one or more parameters. The e3 value model, however, solely uses these parameters to determine the economic viability of the network in quantitative numbers. This economic viability is determined by the actors' profitability, which is calculated by identifying their costs and revenues in the value network [39].
- Two types of change: An examination of the e3 value model puts forward that there are two types of change in a value network: a change in the structure of the value network, such as actors (dis)appearing, and a change of its economic viability [39].

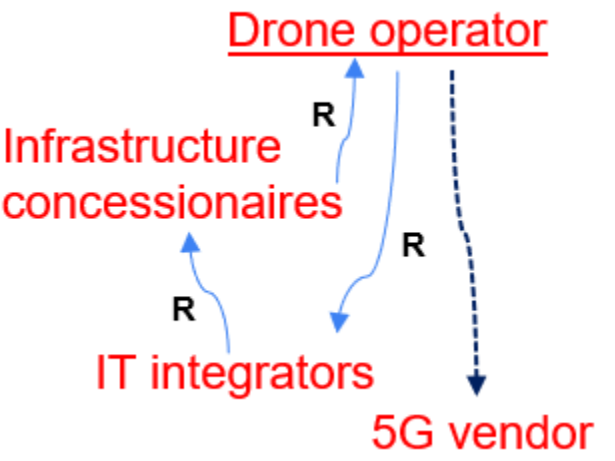
Following the methodology described above, we are proposing, in FigureFigure 53, the initial value network models for each AI@EDGE use case, that could be subject to further analysis. Solid line **R** denotes the relation between partners and dotted line shows the revenue stream from the customer to the supplier of the services. Underlined players correspond to customers, i.e., source of revenue.



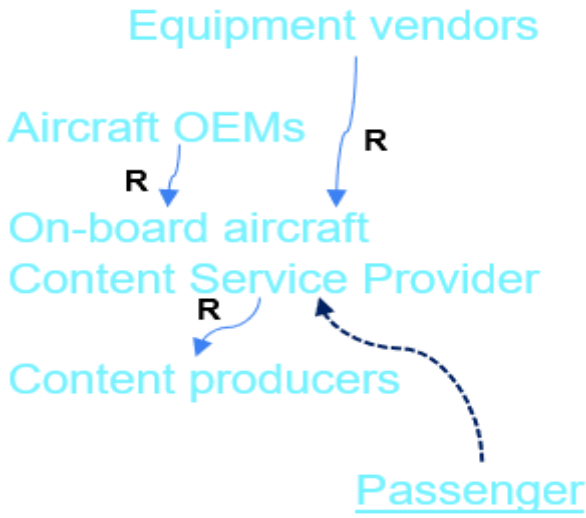
Value network model for Use case 1: Virtual validation of vehicle cooperative perception



Value network model for Use case 2: Secure and resilient orchestration of IoT NWs



Value network model for Use case 3: Edge AI assisted monitoring of linear infrastructures using drones in BVLOS operation



Value network model for Use case 4: Smart content & data curation for in-flight entertainment services

Figure 53 Value network models for AI@EDGE use cases

5.3.3 *Step 2 - Business case viability study*

5.3.3.1. Relevant technoeconomic analyses

In this section we summarize surveys and reports which have as goal related investments, targeting at the technoeconomic aspects of those reports. The common denominator of all the selected reports is the fact

that they aim at finding the most cost- and energy-efficient solution which covers all the requirements of the scenarios that they target at.

Techno-Economic Analysis for Programmable Networks [40] paper compares a programmable with a traditional technoeconomic model. The programmable network has mainly virtual parts and its architecture is based on SDN (Software-Defined Networking) and NFV (Network Function Virtualization). The traditional model refers to the existing networks which are currently used for mobile communication, and they are based on hardware resources without having any virtualization parts. The comparison between these models is occurred in a technoeconomic way and it has to do with base station (BS) cost, power consumption cost, OPEX, CAPEX and TCO cost. More specifically, the BS costs affects both models, but it is noticed that traditional model has higher BS cost, and especially when the amount of BS is increased. This observation is not the same for the SDN networks while the cost of cells is lower and as result the BS cost is not affected much. Power consumption is so important for environmental reasons as for economical ones. In particular, the traditional network needs four times the cost of the SDN network which means that the second one is more cost efficient. The OPEX costs do not affect the overall pricing model that much, making the SDN option a viable solution for operators and providers. The CAPEX cost is high, which is affected by oVS, OFController and other parameters, and it is the main concern of adopters. The analysis shows that the benefits of this technology, the financial profits and the low OPEX cost will contribute to the adoption and integration.

Small cells for Micro-Operators-Deployment Framework [41] paper is about an indoor deployment framework of a uO (micro-operator) in a campus which is incorporating network slicing, network sharing and small cells. Femtocells are low-cost, low-power and plug-and-play devices and they can be deployed with Wi-Fi. They can also replace Wi-Fi access points and use the existing network cabling, local servers, switches and existing internet backbone. The goal of this paper is to examine different scenarios for a campus network (in-building wireless access), having as criterial not only technical such as performance, coverage and spectrum utilization but economical ones, including the Total Cost of Ownership (TCO), capital expenses (CAPEX) – site acquisition, equipment, planning, commissioning and deployment costs- and operational expenses (OPEX) – annual site lease, power usage, operational costs. The result of this techno economic analysis is the cost of femtocells and Wi-Fi access points are almost comparable. Even though, femtocells have higher OPEX (the CAPEX is close to Wi-Fi cost) they present advantages in latency, reliability, performance, and they cover totally the needs of the initial scenario. On the other hand, for the same amount of capacity, it requires less femtocells than Wi-Fi access points which leads to lower TCO. According to the above results, it is proven that small cell network deployment can be cheaper and more beneficial than an exclusive Wi-Fi deployment, even though small cells are more expensive than Wi-Fi access points on a unit basis.

Techno-economic Analysis and Prediction for the Deployment of 5G Mobile Network [42] This techno economic analysis for the deployment of 5G technology over the existing 4G mobile network, took place in Shanghai/China with duration of 6 years. The main goal of this analysis is the benefits and the cost-effectiveness of adoption 5G technology. In order to be performed this model, several parameters have to be defined, such as the predicted number of users, the churn rate that focuses on its impact on revenue of 5G network, the pricing strategy analysis and the evolution of CAPEX and OPEX for a variety of Base Stations classes and different indoor scenarios. Apart from the price and cost, the coverage and the capacity are under consideration. The outcomes show that a good analysis of Price Elasticity of Volume (PED), which is a measure of sensitivity of realized volume to changes in unit price, provides an important margin of benefit. Also, for the current mobile broadband demand of the different scenarios, microcells are the most cost-efficient solution. The network investment model reveals that the deployment of a large amount of new sites is expensive, but this cost can be decreased by the reuse of existing sites or using the carrier aggregation functionality of LTE-A RAT. Finally, it is ascertained a lack of the capacity limited by the

macro sites and in general a limited coverage with small cell solutions such as femtocells, picocells deployed with 5G mmW system and Wi-Fi. To overcome the lack of these, operators should investigate the cooperative layouts of macro sites with femtocells, 5G mmW PBS or Wi-Fi and achieve the trade-offs among capacity, cost, and coverage.

Techno-economic Analysis of 5G Immersive Media Services in Cloud Enabled Small Cell Networks: The Neutral Host Business Model [43] Edge computing and 5G can be a beneficial and interesting combination for vertical industries to develop unconventional services. The study in [43] aims to examine a cloud-enabled Small Cell network, that belongs to a NH, which is going to support Immersive Media Services in Crowded Events. A potential investment in a 5G infrastructure must be reviewed and a planning model to be created to predict the required compute, storage, and radio resources. Also, some economic indices must be taken into consideration such as net present value, internal rate of return and in general the Price evolution. This techno-economic analysis concentrates on the IT and radio infrastructure, that are deployed by NH, for sport stadiums or concert halls where there are a lot of attendees. Particularly, this study, is based on out-turns from three funded research projects which are the SESAME project for the planning model and the IST-TONIC and CELTIC-ECOSYS for the economic part. A Cloud Enabled Small Cell network which can provide Immersive Video Services can be a profitable and viable solution for the above scenario. More specifically, it needs servers, small cells and GPU that mostly contribute to CAPEX by 45%, 25% and 14% respectively. The 58% of OPEX is covered by installation and employee costs. This investment can reach a break point at 6.5 years, and it can be considered fruitful having in the mind that the effective functional period of telecom infrastructures is 15 to 20 years.

5.3.3.2. State of Model inputs and mathematical formulas of TCO for all AI@EDGE use cases:

Following the introduction of the cost model and TCO in D2.3. the cost model is based on several assumptions that should be made to have as realistic results as possible:

- We assume a lifetime of 10 years
- Project horizon is 10 years (hence no hardware renewal is anticipated)
- Hardware installation cost is 15% of the hardware costs
- Maintenance cost for infrastructure is assumed to be 10 % of the original hardware and labour costs (CAPEX)
- The overhead cost is around 22% on top of the sum of the CAPEX and OPEX costs
- Yearly costs are also corrected for inflation (time value of money) with 5% as growth rate
- The TCO is the total cost of ownership for the project lifetime (hence, costs are not discounted)

The Total Cost of Ownership (TCO) of the proposed solution is counted as the sum of the CAPEX, the OPEX of T years and the overhead costs of T years (T is the project horizon).

$$TCO = CAPEX + \sum_{t=1}^{T}(OPEX(t) + Ovhd_c(t))$$

Maintenance costs are counted in the OPEX costs.

$$M = 10\% \times CAPEX$$

Specific calculations (plots) will be presented for each use case in D6.3 in parallel with the development and testing of the use cases. We expect that in some use cases the CAPEX costs will be dominant, on the

other hand in some others the operational and maintenance costs will be higher. Sensitivity analysis (Step 4) will also be reported in deliverable D2.4.

5.3.3.3. Link with exploitation plans

TCO for the deployment of the AI@EDGE platform and the offering of the services, after the end of the project, should be shared among partners (stakeholders) involved in the value network models proposed in Subsection 5.3.2. On the other hand, this is directly linked with the exploitation plan of each partner (Task 6.4) and the strategic plan of each company.

5.3.4 *Step 3 - Impact of new technologies on business case*

The main goal of AI@EDGE is to build a platform and a set of accompanying tools for enabling secure and automated management, orchestration, and operation of AI-powered services over edge and cloud compute infrastructures, with close to zero-touch of the underlying heterogeneous MEC resources (network, storage, and compute resources). One of the key aspects to achieve this vision, is to develop a set of solutions broadly divided into two distinct areas: (i) solutions for the creation, utilization, and adaptation of secure and privacy-aware AI/ML models; and (ii) solutions managing distributed resources inside the operators' infrastructure.
The following are the main AI@EDGE platform objectives: i) design and validate a connect-compute platform enabling the creation of network slicing, ii) extend ETSI MEC/NFV architectures with applications and models capable of providing the AI@EDGE platform with the context and metadata necessary to take automatically actionable decisions and to realize intelligent data and computation offload control and management of applications and services deployed over the decentralized and distributed AI@EDGE platform, iii) investigate different hardware acceleration solutions (FPGA, GPU, CPU) spanning from the terminals to the cloud for highly decentralized and distributed workload management, iv) analyse and compare dual-connectivity monolithic RANs with cross-layer multi-connectivity disaggregated RANs to see if dynamically adapts the network topology to the network conditions.
The new technologies being developed in AI@EDGE project (a connect-compute platform providing solutions based on privacy-aware AI/ML models and managing distributed resources inside the operator's infrastructure) are expected:
(1) to accelerate the business model as part of the strategic plan of each firm and
(2) to provide revenues to the supplier of the services in the coming 5 to 10 years.

5.3.5 *Step 4 - Sensitivity analysis*

Sensitivity analysis (Step 4) will be reported in deliverable D2.4.

6 Key Performance Indicators

This Chapter summarizes the process of exploring the different steps and 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 tries to define a common format for KPIs description, which encapsulates all aspects that are being tackled from a holistic point of view.
In order to ease this dynamic process, a KPI matrix was introduced, which encapsulates all the KPIs characteristics. Since providing deliverable D2.2, some changes were introduced to the KPI Matrix in order to better adjust the format to the core characteristics of the AI@EDGE Platform and to align with the 5GPPP template, as discussed in Section 6.1. These changes are reflected not only in the KPI Matrix itself in Section 6.2, but also in the new tool, KPI Console, introduced and described in Section 6.3.

6.1 Integration with the 5GPPP TMV WG Template

As stated in the previous deliverable (D2.2), in order to capture KPIs in a holistic and comprehensible manner, the AI@EDGE project proposed a template based on fields that contain essential information and data that define a KPI. Based on this approach, the template takes the form of a KPI Matrix, which contains the core characteristics of a specific KPI. The main domains introduced in the first draft of the KPI Matrix are as follows:

- 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.

The new revision of the KPI Matrix (version 2), contains the following two fields:

- 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.

Together with collecting and defining AI@EDGE KPIs, there was an initiative for collaboration with the 5GPPP Test, Measurement and KPIs Validation Working Group (TMV WG), as its objectives bear close relation with the work already implemented for the AI@EDGE Project. Upon collaboration with the TMV WG, the KPI Matrix was enhanced with fields that were adopted directly from the template proposed by the 5G PPP TMV formula of collecting and defining KPIs.
The added value of this modification is encapsulated in the following columns that were adopted to the KPI Matrix (Technical), producing the version 2:

- Where to measure the KPI
- Relevant components (Parts of the platform that directly affect KPIs)
- Project enhancement (How meeting the target value for said KPI drives to innovation)
- Comments (General remarks regarding the specific KPI)

6.2 KPIs Matrix

With the modifications described in the previous subsection, the structure of the new KPI Matrix (Technical) is as follows and shown in Table 1. We should note at this point that the fields marked as TBD (to be defined) are as of now not yet decided or defined, but they are going to be in future editions of the Matrix.

Domain [ID]	Group [ID]	KPI Description [ID]	Use Case Nr / All / Generic	Threshold (Number / Qualitative Description)	Achieved Value	Where to Measure	Components Involved	Project enhancement	Comments
Technical [T]	Networking [N]	Vehicle Density [TN1]	1	1200 vehicles/km2	1200 vehicle/km2	- 5G network Emulators - Driving Simulators	AV/ADAS Vehicle simulators Network Simulators,MEC Platform	Development of cooperative perception AI Application (on-board Vehicle and on Edge)	N/A
		Drone Range [TN3]	3	> 20km	TBD	Drone flight plans	Drone Fleet	Deployment in larger coverage surface	N/A
		Data Rate/Client for Streaming [TN4]	4	> 15 Mbps	TBD	UEs connected to in-flight network	Edge Platform, 5G Networks	Providing high-end in-flight entertainment (IFE)	N/A
		Aggregate In-Cabin Throughput Density [TN4]	4	≥ 20 Mbit/s/sqm	TBD	Cabin WLAN	Cabin experimental setup	User Experience	N/A
	Computing [C]	Latency V2V [TC1]	1	< 160 ms	160 ms	TBD	Vehicles	TBD	N/A
		Latency V2N [TC1]	1	≤ 2000 ms	2000 ms	5G network Emulators Driving Simulators Network Simulators	AV/ADAS Vehicle simulators, Network Simulators, MEC Platform	Evaluation of the C-V2X simulation environment	N/A
		Control Signal Latency [TC3]	3	≤ 50 ms	TBD	TBD	TBD	TBD	N/A
		Video Processing Latency [TC3]	3	≤ 50 ms	TBD	Applications that deploy video streaming content	Multi-Media streaming services, 5G/6G Networks, Edge Computing	Ultra-fast services in regards to video streaming services	N/A
	AIF [A]	Robust AIs [TA2]	2	< 5% detection rate decrease against adv. samples	TBD	TBD	AI algorithms deployed in the AI@EDGE platform	Security and Cybersecurity	N/A

		Fast Detection [TA2]	2	Local within 1s, global within 1m	TBD	TBD	AI algorithms deployed in the AI@EDGE platform	TBD	N/A
		False Alarm Rate ~~à possibly on AIF group~~ [TA2]	2	Rate: < 0.1 %	0.1%	Anomaly Detection Algorithms	ML Anomaly Detection algorithms, Datasets, Data output from UC2	Reliability, Robustness	N/A
		Known-Attack Detection ~~à possibly on AIF group~~ [TA2]	2	Detection Accuracy ≥ 97 %	97 %	ML Algorithms running on edge devices	ML Anomaly Detection algorithms, Datasets, Data output from UC2	Security, Robustness	N/A
	Reliability [R]	Service Deployment Time [TR4]	4	TBD	TBD	TBD	TBD	User Experience	N/A
		Service Recovery Time [TR4]	4	≤ 180 s	TBD	UEs connected to in-flight network	Edge Platform, 5G Networks, IFE	Providing high-end in-flight entertainment (IFE)	N/A
		Curated Content Delivery Time [TR4]	4	≤ 180 s	TBD	TBD	TBD	TBD	N/A
		Content Curation Precision of Recommendation [TR4]	4	≥ 80 %	TBD	TBD	TBD	TBD	N/A
		Number Of Served Passengers [TR4]	4	12 for demonstration	TBD	Plane Cabin (experimental setup)	TBD	TBD	N/A
		Communication Reliability [TR1]	1	99.9%	99.9999 %	TBD	TBD	User Experience	N/A
		Control Signal Packet Loss [TR3]	3	≤ 1 %	1 %	Direct E2E links in Edge Computing Systems, eNBs, ground stations etc	Edge Computing Systems, 5/6G stations, 5G Core Network etc	Enhance communications reliability, improve edge platforms, build robust wireless links	N/A

Table 1 KPIs Matrix (Technical) – Version 2

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	2	

Table 2 KPIs Matrix (Societal, Economic and Environmental)

6.3 *Introduction to the KPI Console*

In this section we introduce the KPI Console, an ad-on tool for the management and monitoring of the KPIs defined for the AI@EDGE platform. Moreover, this tool could be also used as starting point for the further developments with addition of new functionalities relevant for the AI@EDGE platform, both from the users operating AI@EDGE platform point of view and from the business users' point of view.

6.3.1 *KPI Console Platform*

The KPI Console is an open-source web application, which can be deployed both in development and in production servers. The home page welcomes the user, and it provides a side menu which contains all the console's main functionalities, as it is shown in the Figure 54. It is important to state at this point that this tool is highly and easily customizable, so if during the project the need arises for other KPI attributes to be dynamically queried, the current implementation can easily adopt this modification.
The KPI Console platform was designed to be a main standpoint, which will serve as a centralized hub for all kind of KPI data to be saved, shared, and disseminated throughout the span of the AI@EDGE project components, thus making the collection, measurement, and definition of KPIs in a more agile process.
In the left sidebar, the 4 main functionalities (as of the earliest version) can be accessed:

- KPI List: displays the section of the KPI Matrix containing every attribute regarding the KPIs
- Measurements: displays the section of the KPI Matrix containing the values of the attributes that provide measurements for the KPIs
- About specific KPI: implements a page containing query forms that take a KPI as an input and print out the relevant information
- Feedback: provides a contact form, through which a user can upload comments/observations/other useful point of information

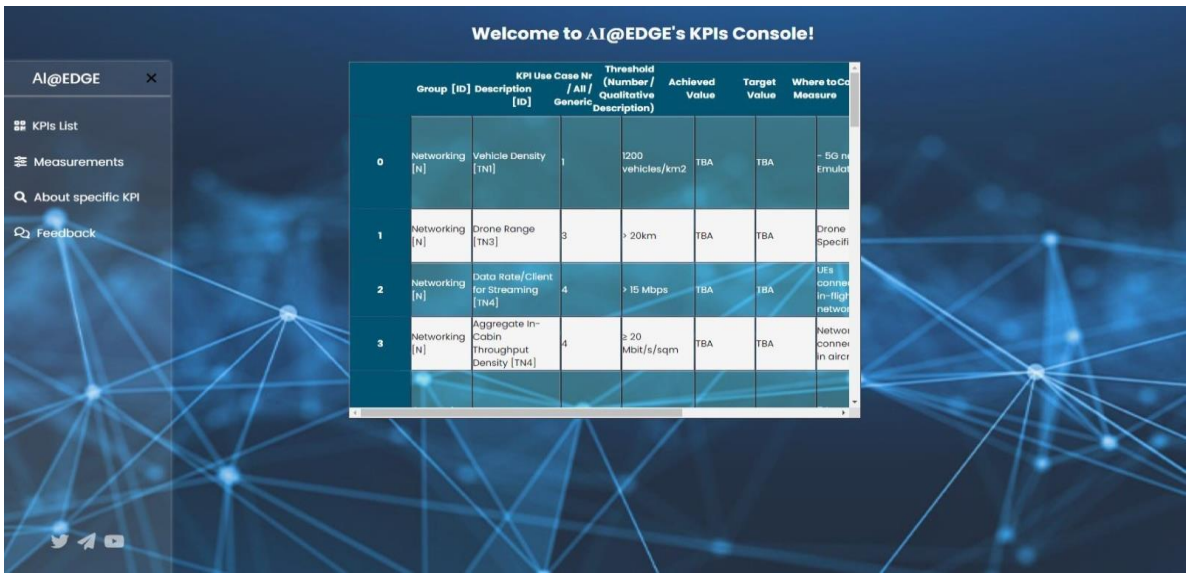


Figure 54 Home page of the KPI Console

6.3.2 *Dynamic KPI console and real-time monitoring*

There are several aspects related to the monitoring of the KPIs, depending on the specific indicators being tracked and the needs of the AI@EDGE platform and use cases. Some relevant aspects include:

1. **Establishing a system for collecting data on the KPIs**: this may involve deploying and using software tools to track and record data on the KPIs on a regular basis.
2. **Setting benchmarks and targets**: using historical data or industry standards to establish benchmarks and targets for the KPIs, to measure the performance against targets and goals.
3. **Regularly reviewing the KPIs**: regular review of the data on the KPIs to identify trends and patterns, and to assess progress towards achieving targets and goals, that could be done through regular meetings, reports, or other communication channels.
4. **Taking action based on the KPI data**: once the data on the KPIs has been reviewed, it's important to take action to address any issues or opportunities identified. This may involve implementing new strategies, adjusting processes, or making other changes to improve performance.

One way that users can implement real-time KPI monitoring is by using dashboard software or other data visualization tools. These tools allow users to create visualizations of their key metrics and display them on a single screen or dashboard, allowing decision makers and other stakeholders to monitor the performance of the KPI quickly and easily. Another approach to real-time KPI monitoring is to set up automatic alerts that notify decision makers and other stakeholders when key metrics reach a certain threshold or deviate from expected values. This allows users to be notified immediately when there is a problem or potential issue, allowing them to take timely action to address the issue and prevent it from escalating.
Real-time KPI monitoring is an essential tool for businesses and organizations that want to improve their performance and stay on track towards their targets and goals. By continuously monitoring and tracking key metrics, users can identify potential issues early and take timely action to prevent them from becoming major problems. This allows businesses and organizations to maintain a high level of performance and ensure that they are achieving their desired results.
Overall, effective KPI monitoring involves establishing a system for collecting and analysing data on the indicators, setting benchmarks and targets, and using the insights gained to drive improvement.
Given the fact that the KPI Console provides a software solution that contains crucial data regarding the AI@EDGE platform KPIs, it could be considered and enhanced to perform also monitoring over the KPIs. The core idea is to identify the parts of the AI@EDGE platform that can feed measurement data regarding specific KPIs, and then connect via API endpoints with the KPI Console Web Application which will subsequently calculated and displayed a near real-time. This possible future enhancement of the KPI Console as a monitoring tool could be provided as an individual component that can be dynamically embedded in the overall AI@EDGE system architecture, thus aiding in aspects such as orchestration, reliability, security, and robustness.
On future deliverable, all components of the AI@EDGE architectural design will be examined, performing the selection of the individual components that can be actors to this procedure of KPIs monitoring.

7 Conclusions

This deliverable presents the contributions of tasks T2.2 and T2.3: the consolidated system architecture and interfaces (Milestone MS2.6), the techno-economic analysis (Milestone MS2.7) and the KPIs (Milestone MS2.5).

The AI@EDGE consolidated system architecture corresponds to a network and service automation platform (NSAP) 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 had an impact on the definition of the components, interfaces, and workflows.

The AI@EDGE Architecture wanted position is to provide capabilities that go beyond the specific requirements of AI@EDGE use cases, aiming to support the following topics, which have been addressed with **respective AI@EDGE architecture contributions**.

- artificial intelligence "in-platform" inclusion (Network and Service automation intelligence) enabling better usage of infrastructure resources, i.e., high performance converged computing and communication platform; **AIFs as a concepts and interfaces, AIF descriptor.**
- artificial intelligence "on-platform" inclusion (End-user Application intelligence) enabling better end-user services quality of experience in various application domains (verticals such as automotive), easy and flexible deployment of the third-party AIFs; **AIFs as a concepts and interfaces, AIF descriptor.**
- data pipeline and data governance framework, being the data-driven platform, while preserving privacy and security of data in multi-stakeholder environment; **AI/ML data pipeline including model lifecycle management and data collection components, aiming at an underlying data-driven architecture.**
- E2E overall system orchestration and management, with orchestration and management for AIFs workflows, seamlessly integrated with network and resource orchestration and management, enabling open, collaborative, distributed ecosystems**. A NSAP architecture with new orchestration components, such as MTO, IOC and IARM; a CCP architecture with MEO, IOC and IARM; and new interfaces such as MEO to MEO and MTO to MEO.**

The presented contributions are based on the inputs from WP3, WP4 and WP5, which will be further described in deliverables D3.2, D4.2 and D5.2, respectively, and on leveraging state-of-the-art standards, open-source solutions, and outputs of other 5G-PPP projects. Thus, the consolidated system architecture describes the interactions between MEC and 5G System through NSAP and CCP platforms to enhance AIFs and Network automation operations, defining in particular the concrete functionalities of the Intelligent Orchestration Component (IOC).

Chapter 5 on dialogues, drivers and techno-economic analysis presented an updated version of the different dialogues on e-health, transportation, etc., followed by an introduction of TCO as one of main drivers for techno-economic analysis and general methodology the techno-economic analysis in the project.

Chapter 6 on key performance indicators summarized the process of exploring the different steps and components developed under the cloak of the AI@EDGE platform, to define a set of Key Performance Indicators (KPIs) within the project. The methodology tries to define a common format for KPIs description, which encapsulates all aspects that are being tackled from a holistic point of view, resulting in a so called KPI matrix and KPI console, both described in the chapter.

The techno-economic analysis work presented in D2.2 and now in D2.3 will continue to evolve until the end of the project, providing views on 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. Regarding the KPIs, this deliverable groups and extends the UC-related KPIs introduced in deliverable D2.2, presenting the updated KPI matrix organized in four basic pillars-domains: Technical, Societal, Economic and Environmental. This matrix will evolve during the project according to the input received from the different partners, experimentations and use cases.

As an example of the next step, WP5 intends to deep dive into how AI@EDGE use cases could exploit the consolidated system architecture during the next months, and these learnings will be described in conjunction with the techno-economic analysis in D2.4.

References

[1] University of Oulu – 6G Flagship – *6G White Paper on Edge Intelligence* - isbn9789526226774.pdf
[2] Kliegr, T., Bahník, Š., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*, 103458.
[3] Kshirsagar, D., & Shaikh, J. M. (2019, September). Intrusion detection using rule-based machine learning algorithms. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)* (pp. 1-4). IEEE.
[4] Jha, K. K., Jha, R., Jha, A. K., Hassan, M. A. M., Yadav, S. K., & Mahesh, T. (2021, December). A brief comparison on machine learning algorithms based on various applications: a comprehensive survey. In *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1-5). IEEE.
[5] Jiang, W. (2022). Cellular traffic prediction with machine learning: A survey. *Expert Systems with Applications*, 117163.
[6] El Mrabet, M. A., El Makkaoui, K., & Faize, A. (2021, December). Supervised machine learning: a survey. In *2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)* (pp. 1-10). IEEE.
[7] Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, *7*, 65579-65615.
[8] Çalışır, S., & Pehlivanoğlu, M. K. (2019, April). Model-free reinforcement learning algorithms: A survey. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.
[9] O-RAN xApp Descriptor - O-RAN-xApp Descriptor.
[10] ETSI GR MEC 017, “Mobile Edge Computing (MEC); Deployment of Mobile Edge Computing in an NFV environment v1.1.1”, February 2018.
[11] ETSI GS MEC 00, “Mobile Edge Computing (MEC) Terminology v1.1.1”, March 2016.
[12] O-RAN specifications Downloads (orandownloadsweb.azurewebsites.net)
[13] Michele Polese, Leonardo Bonati, Salvatore D’Oro, Stefano Basagni, Tommaso Melodia. Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges. DOI 10.1109/COMST.2023.3239220
[14] “3GPP TS 23.502. Technical specification group service and system aspects; system architecture for the 5G system (5GS),” v16.7.0.
[15] “ITU Focus group on machine learning for future networks including 5G,” https://www.itu.int/en/ITU-T/focusgroups/ml5g/.
[16] ITU-T Y.3172 - Architectural framework for machine learning in future networks including IMT-2020
[17] Generic Network Slice Template - https://www.gsma.com/newsroom/wpcontent/uploads//NG.116-v7.0.pdf
[18] 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.
[19] S. Kekki et al., "MEC in 5G networks," ETSI white paper n. 28, June 2018.
[20] O-RAN Alliance, “O-RAN R1 interface: General Aspects and Principles 3.0”, October 2022
[21] O-RAN Alliance, “O-RAN Architecture Description 7.0”, October 2021.
[22] O-RAN Alliance, “O-RAN AI/ML Workflow Description and Requirements 1.03”, October 2021
[23] ETSI GS MEC 003, "Multi-access Edge Computing (MEC); Framework and Reference Architecture v2.2.1", Dec. 2020.
[24] 3GPP, "System architecture for the 5G System (5GS)," TS 23.501, Rel. 17, V17.4.0, May 2022.
[25] 3GPP, "Procedures for the 5G system (5GS)," TS 23.502, Rel. 17, V17.7.0, Dec 2022.

[26] 3GPP, "Network Exposure Function Northbound APIs," TS 29.522, Rel. 17, V17.5.0, May 2022.
[27] 3GPP, "T8 reference point for Northbound APIs," TS 29.122, Rel. 17, V17.5.0, May 2022.
[28] 3GPP, "5G System Enhancements for Edge Computing," TS 23.548, V17.2.0, May 2022.
[29] 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.
[30] O-RAN Alliance, "O-RAN Near-Real-time RAN Intelligent Controller Architecture & E2 General Aspects and Principles v1.01", July 2020.
[31] O-RAN Alliance WG3, "Near-Real-time RAN Intelligent Controller, E2 Application Protocol (E2AP) O-RAN.WG3.E2AP-v02.03", Technical Specification
[32] O-RAN Alliance WG3, "Near-Real-time RAN Intelligent Controller E2 Service Model (E2SM), O-RAN.WG3.E2SM-v02.01", Technical Specification
[33] Gavras, Anastasius, Durkin, Patrick, Fletcher, Simon, Hallingby, Hanne Kristine, & Mesogiti, Ioanna. "Business Validation in 5G PPP vertical use cases" (2020). https://doi.org/10.5281/zenodo.3775405
[34] D5.1, Ecosystem analysis and specification of Business and Economic KPIs, project 5GVINNI, July 2019 [Online]. Available: https://zenodo.org/record/3345665
[35] 6G – IA Whitepaper on 5G Ecosystem Business Modelling (discussions/calls November 2022)
[36] V. Allee, "Value Network Analysis and value conversion of tangible and intangible assets," Journal of Intellectual Capital, vol. 9, no. 1, pp. 5-24, 2008.
[37] V. Allee, "Value Network Analysis and value conversion of tangible and intangible assets," Journal of Intellectual Capital, vol. 9, no. 1, pp. 5-24, 2008.
[38] https://www.thevalueengineers.nl/tutorials/model-analyze-value-model/
[39] V. Allee, "Reconfiguring the value network," Journal of Business strategy, vol. 21, no. 4, p. 36-39, 2000.
[40] C. Bouras, A. Kollia, "Techno-Economic Analysis for Programmable Networks," pp. 2-19.
[41] J. S. Walia, H. Hämmäinen, M. Matinmikko, "5G Micro-operators for the Future Campus: A Techno-economic Study".
[42] G. Smail, J. Weijia, "Techno-economic Analysis and Prediction for the Deployment of 5G Mobile Network".
[43] P. Paglierani, I. Neokosmidis, T. Rokkas, C. Meani, K.M. Nasr, P.S. Khodashenas, "Techno-economic Analysis of 5G Immersive Media Services in Cloud Enabled Small Cell Networks: The Neutral Host Business Model".
[44] N. -E. -H. Yellas, B. Addis, R. Riggio and S. Secci, "Function Placement and Acceleration for In-Network Federated Learning Services," - Oct 2022