HEXA-X-II

A holistic flagship towards the 6G network platform and system, to inspire digital transformation, for the world to act together in meeting needs in society and ecosystems with novel 6G services

Deliverable D6.2
Foundations on 6G Smart Network Management Enablers

Co-funded by the European Union

Hexa-X-II project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101095759.

Date of delivery:	30/10/2023	Version:	1.0
Project reference:	101095759	Call:	HORIZON-JU-SNS-2022
Start date of project:	01/01/2023	Duration:	30 months

Document properties:

Document Number:	D6.2
Document Title:	Foundations on 6G Smart Network Management Enablers
Editor(s):	Ricard Vilalta (CTT)
Authors:	A. Gallego, A. Labrador, E. Lluesma, A. Ramos (ATO), R. Muñoz, P. Alemany, Ll. Gifre, C. Manso (CTT), W. Tavernier, J. Miserez (IME), G. Landi, M. De Angelis, P. G. Giardina (NXW), J. Ordonez-Lucena, R. Nicolicchia (TID), S. Rodríguez (OPT), T. Dimitrovski, N. Toumi (TNO), A. Zafeiropoulos, I. Tzanettis (ICC), C. Ayimba (UC3), V. Lamprousi, S. Barmpounakis, P. Demestichas (WIN), M. Karaca, L. Karaçay (EBY), J. Castaneda, F. Brito (EAB).
Contractual Date of Delivery:	30/10/2023
Dissemination level:	PU
Status:	Final
Version:	1.0
File Name:	Hexa-X-II D6.2_v1.0

Revision History

Revision	Date	Issued by	Description
0.1	03.2023	Hexa-X-II WP6	Table of Contents
0.2	04.2023	Hexa-X-II WP6	First draft of SoTA
0.3	05.2023	Hexa-X-II WP6	First draft of Enablers
0.4	25.07.2023	Hexa-X-II WP6	First complete draft
0.5	05.09.2023	Hexa-X-II WP6	After internal review
0.6	30.09.2023	Hexa-X-II WP6	After external review
1.0	30.10.2023	Hexa-X-II WP6	Submitted version

Abstract

This report presents the design foundations for the envisioned enablers that will be implemented as part of the S*mart Network Management* framework that is targeted in Hexa-X-II project. This initial design work has been performed based on the analysis of 6G drivers and KPIs/KVIs, elaboration on the functional objectives regarding Management and Orchestration (M&O) established in Hexa-X-II Description of Action (DoA), study of background information elements that have been identified in the State of the Art (SoTA) and the 1st end-to-end Hexa-X-II system blueprint.

The Hexa-X-II *Smart Network Management* enablers aim to build a programmable cloud-native M&O framework for the future 6G networks, that will be including mechanisms that define a 6G enabled trustworthy environment, orchestration mechanisms for managing the deployment of 6G services over heterogeneous resources across the IoT-to-edge-to-cloud continuum, sustainable AI/ML-based network control solutions with optimal energy efficiency as well as zero-touch M&O mechanisms for closed loop automation, guaranteeing compliance with relevant 6GKVI/ KPIs while reducing OPEX.

The current version of the *Smart Network Management* enablers presented in this document will be considered into the next (2nd) design iteration of the end-to-end Hexa-X-II system design in a bottom-

up approach. The current enablers will be later analysed in order to be compliant with the 6G end-to-end architecture objectives. Further end-to-end requirements will be taken into account to provide an updated and complete design of the enablers expected in the project by June 2024.

Keywords

Smart Management, Orchestration, Control, Monitoring, Telemetry, Zero-Touch, Automation, Trust, Multi-cloud, Compute Continuum, Sustainable AI, Network Digital Twins, Closed Loop.

Disclaimer

Funded by the European Union. The views and opinions expressed are however those of the author(s) only and do not necessarily reflect the views of Hexa-X-II Consortium nor those of the European Union or Horizon Europe SNS JU. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

The Smart Network and Services Joint Undertaking (SNS JU) 6G Flagship project Hexa-X-II leads the way to next generation 6G end-to-end (E2E) system design, based on integrated and interacting technology enablers. The project will continue the track of the previous 5G-PPP Hexa-X project, which has laid the foundation for the global communication network of the 2030s by developing the 6G vision and basic concepts. As a continuation of Hexa-X the interaction between the cyber-physical world and human world will be further evolved with the advancement of information, communication, and computation technologies towards a pervasive human centred cyber-physical world in 2030. To reach this 6G vision, Hexa-X-II has set up the goals to design the system blueprint of a sustainable, inclusive, and trustworthy 6G platform, which will require novel enablers regarding Smart Management & Orchestration (M&O) functionalities.

This report is the first public deliverable released by the "*Smart Network Management*" *Work Package* in Hexa-X-II project and its ultimate goal is to provide the initial design foundations for the envisioned enablers that will be implemented as part of the smart network management framework that is targeted in Hexa-X-II. The identification of the smart network management enablers and their initial design foundations have been performed based on the analysis of 6G drivers and KPIs/KVIs, elaboration on the functional objectives regarding Management and Orchestration (M&O) established in Hexa-X-II Description of Action (DoA), study of background information elements that have been identified in the State of the Art (SoTA), including previous Hexa-X Management and Orchestration principles, and the initial end-to-end Hexa-X-II system blueprint provided in [HEX223-D21].

The main drivers behind the development of 6G technology have been analysed providing an objective analysis of their impact. Environmental, social, and economic factors have identified influencing the adoption and deployment of 6G technology. Then the potential benefits and challenges associated with these drivers have been studied. Besides Hexa-X KPIs have been examined, which will be used later on in the project to measure the success and efficacy of the M&O strategies proposed. The overarching objectives and goals for M&O in the Hexa-X-II project have been elaborated, highlighting the crucial role played by effective management strategies in realizing the full potential of 6G technology.

The main part of this deliverable covers the description of the design foundations for the following 11 proposed Smart Management and Orchestration enablers:

- Enabler 1: Programmable flexible network configuration
- Enabler 2: Programmable network monitoring and telemetry
- Enabler 3: Integration fabric
- Enabler 4: Trustworthy 3rd party management
- Enabler 5: Multi-cloud management mechanisms
- Enabler 6: Orchestration mechanisms for the computing continuum
- Enabler 7: Sustainable AI/ML-based control
- Enabler 8: Trustworthy AI/ML-based control
- Enabler 9: Network Digital Twins
- Enabler 10: Zero-touch closed loop governance
- Enabler 11: Zero-touch multiple closed loop coordination

For each enabler, the deliverable presents: i) the main motivation or which problem each enabler aims to solve, ii) its objectives, iii) a more detailed description, iv) the main State of The Art (SoTA) items in which each enabler relies on, and how the enabler goes Beyond SoTA items, v) identified internal components and interfaces, and vi) its relationship with other enablers.

The identified list of enablers are expected to have significant positive impact towards 6G especially in relation to the following aspects: i) Increase of network automation and network autonomy and consequently also reducing operational expenses (OPEX) (Enablers 1, 3, 7, 8, 9, 10, 11), ii) Environmental sustainability, network efficiency and decarbonization (Enabler 7) iii) Trustworthiness (Enablers 3, 8) iv) Improved performance in terms of zero-perceived latency and higher speed (Enablers 2, 5, 6).

The M&O enablers are planned to be validated by means of two Proof of Concept (PoC) being integrated to build concrete use cases: PoC#A.1 about Sustainability and trustworthy-oriented orchestration in 6G, and

PoC#B.1 about AI-assisted end-to-end lifecycle management of a 6G latency-sensitive service across the compute continuum.

The output of this document will be considered into the next design iteration of the Hexa-X-II end-to-end system design in a bottom-up approach so the different enablers presented in this document will be analysed in order achieve the 6G E2E architecture objectives. Besides, the work presented in this deliverable sets up a framework for further research and development within the project in the next phases:

- Further requirements will be incorporated in the following design iteration, once that work done in the project regarding use cases analysis is released by Dec 2023.
- The overall updated Hexa-X-II system blueprint regarding M&O functionalities will be considered for the update of the current enablers in following iterations in the project in a top-down approach.
- Next deliverable by June 2024 will cover the initial Design of 6G Smart Network Management Framework. This report will provide early results on the 6G smart network management enablers, contributing to the 3rd iteration of the overall Hexa-X-II system blueprint.

Table of Contents

List of Tables

List of Figures

Acronyms and abbreviations

Term	Description
ADT	Application Data Transfer
AI	Artificial Intelligence
API	Application Programming Interface
BSS	Business Support System
CAPIF	Common API Framework for 3GPP northbound APIs
CCL	Closed-Control loop
CI/CD	Continuous Integration/Continuous Deployment
CL	Closed loop
CLC	Closed loop Coordination
CNCF	Cloud Native Computing Foundation
CNF	Containerized Network Functions
COINRG	Compute In the Network Research Group
CPU	Central processing unit
DetNet	Deterministic Networking
DMM	Distributed Mobility Management
DMS	Deployment Management Services
DN	Data Network
DNN	Data Network Name
DNS	Domain Name System
DSP	Digital Service Provider
ENI	Experiential Networked Intelligence
ESG	Environmental, social and governance
ETSI	European Telecommunications Standards Institute
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GMPLS	General Multiprotocol Label Switching
GNN	Graph Neural Network
HNF	Hybrid Network Functions
IBN	Intent-based networking

IETF	Internet Engineering Task Force
ILP	Integer Linear Programming
IMS	IP Multimedia Subsystem
IRTF	Internet Research Task Force
ISG	Industry Specification Groups
ITU	International Telecommunication Union
K8s	Kubernetes
KNF	Kubernetes Network Functions
KPIs	Key Performance Indicators
LCM	Life Cycle Management
LTS	Long Term Support
M&O	Management and Orchestration
MADINAS	MAC Address Device Identification for Network and Application Services
MDAF	Management Data Analytics Function
MEC	Multi-access Edge Computing
ML	Machine Learning
MPLS	Multiprotocol Label Switching
MQTT	Message Queuing Telemetry Transport
MVP	Minimum valuable product
NM	Network Modelling
NFV	Network Functions Virtualization
NFVI	Network Function Virtualization Infrastructure
NFVO	Network Function Virtualization Orchestrator
NIC	Network Interface Card
NSSAI	Network Slice Selection Assistance Information
NWDAF	Network Data Analytics Function
ONAP	Open Network Automation Platform
ONF	Open Networking Foundation
OPS23	Operations and Management Area Working Group
OS	Operating System
OSS	Operations Support System

PDU	Protocol Data Unit
PNF	Physical Network Functions
QoE	Quality of Experience
QoS	Quality of Service
RAW	Reliable and Available Wireless
REST	Representational state transfer
RSD	Route Selection Descriptor
RSVP	Resource Reservation Protocol
SAI	Securing Artificial Intelligence
SDO	Standard Development Organizations
SMO	Service Management and Orchestration
SoTA	State of the Art
SSC	Session and Service Continuity
SFC	Service Function Chaining
TAPI	Transport API
TEAS	Traffic Engineering Architecture and Signaling
TIP	Telecom Infra Project
TLA	Trust Level Agreement
TSN	Time Sensitive Networking
UE	User Equipment
URSP	UE Route Selection Policy
VMAF	Virtualized MEC application function
VNF	Virtual Network Function
ZSM	Zero touch network and Service Management

1 Introduction

1.1 Objective of the document and Hexa-X-II project context

The overall objective of this document is to provide the **initial design foundations for the envisioned enablers** that will be implemented as part of the **smart network management** framework that is targeted in Hexa-X-II project.

An enabler is understood in Hexa-X-II as a set of essential components, technologies, and processes that facilitate the efficient and effective control, coordination, and optimization of complex systems and resources, particularly in the realms of 6G networks. Hexa-X-II will design an end-to-end 6G system blueprint aiming to build a sustainable, inclusive, and trustworthy 6G platform based on integrated and interacting technology enablers.

The identification of the smart network management enablers and their initial design foundations presented in this document has been performed based on:

- 6G drivers and KPIs/KVIs analysis elaborated by Hexa-X-II project in [HEX223-D11].
- Analysis of the functional objectives regarding Management and Orchestration (M&O) established in Hexa-X-II Description of Action (DoA).
- Background information elements that have been identified in the State of the Art (SoTA) exploration, including Hexa-X Management and Orchestration principles.
- Initial end-to-end Hexa-X-II system blueprint provided in [HEX223-D21].

The output of this document will be considered into the next design iteration of the Hexa-X-II end-to-end system design in a bottom-up approach. First, the top-down approach involves an iterative design process wherein the end to end system blueprint for the overall 6G platform will be elaborated based on the 6G use-cases requirements and the system architecture design principles. Second, the bottom-up approach designs and provides the different components of the system as enablers. The enablers selection process will consider pros and cons of each potential enabler and component developed or considered for achieving the 6G E2E architecture objectives. A checklist of what can be considered in technical components/enablers for the alignment with the E2E performance and operation targets can then be used as feedback towards enabler design as well as E2E system design (see [HEX223-D21] for further details of this process.

Figure 1-1 represents the initial overall 6G end-to-end system blueprint that is explained in [HEX223-D21]. The enablers presented in the current deliverable correspond to be part of the “management & orchestration” box that is part of the “pervasive functionalities”.

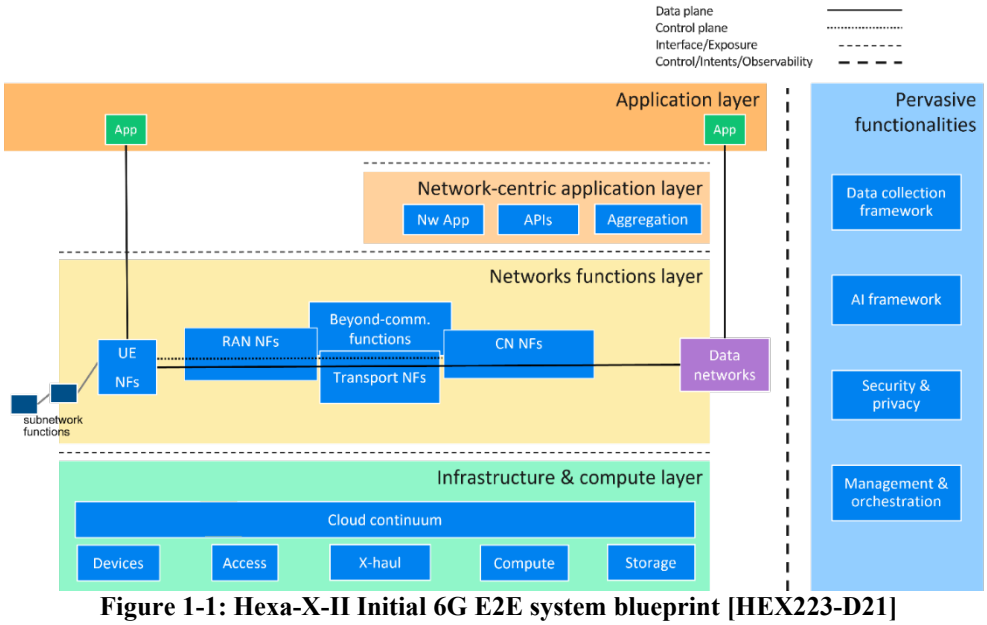


Figure 1-1: Hexa-X-II Initial 6G E2E system blueprint [HEX223-D21]

Besides, at end-to-end application management level there is also an ongoing related work regarding automation that is being described in [HEX223-D21]. Those enablers at application level will be interfacing with the enablers presented in this document (especially related with the work on monitoring, integration fabric, trustworthy management, service orchestration and control loops coordination in sections 3.2, 3.3, 3.4, 3.6, 3.11).

Also general end-to-end security and privacy considerations (that can be found in [HEX223-D21]) are considered in the work on the smart network management enablers.

The other pervasive functionalities, e.g., AI and Data frameworks that are also relevant to M&O are being tackled in [HEX223-D32] in a general end to end approach.

The current set of M&O enablers in this document will be further refined to validate and to make sure they appropriately contribute to achievement of 6G KPIs and KVIs in the next design iteration in the project through the end-to-end system validation [HEX223-D21]. The final list of Hexa-X-II use cases and requirements will be expected by December'23. Those requirements will be considered in the next iteration of the current set of smart network management enablers design that is planned to be released by June'24.

1.2 Hexa-X Management and Orchestration Structural View

Hexa-X-II is the continuation of previous 5G-PPP Hexa-X project, whose main results will be considered to design the overall 6G platform. Regarding Management and Orchestration specifically, though an extensive SoTA analysis has been collected in section 7 as an annex, here below Figure 1-2 represents the Management and Orchestration structural view that was elaborated in Hexa-X project, which is being considered a key reference for the current work in Hexa-X-II.

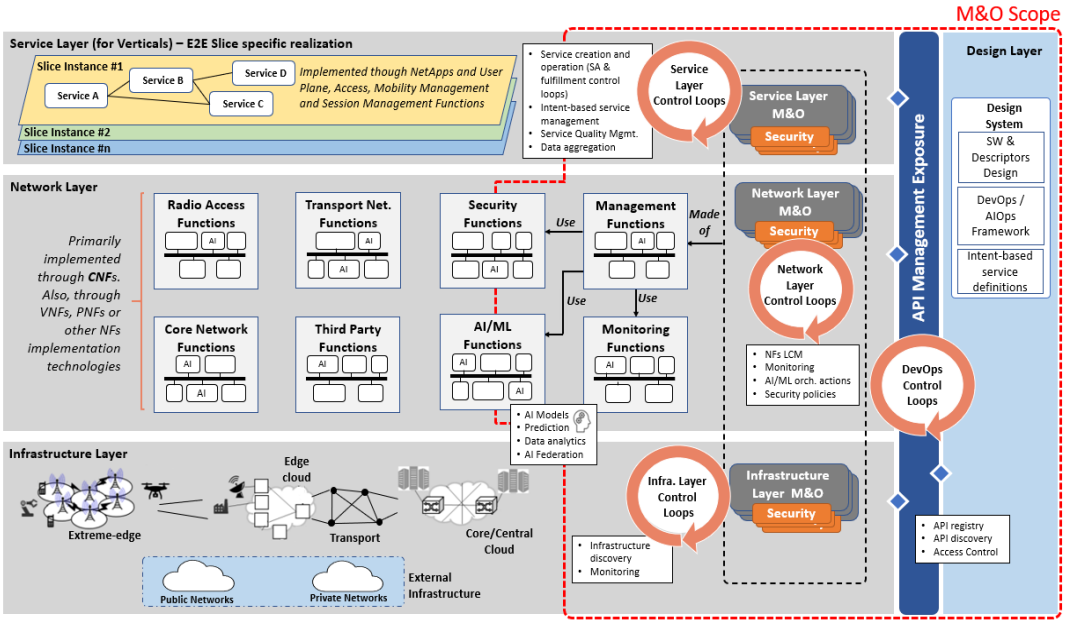


Figure 1-2 Hexa-X Management and Orchestration Structural View [HEX22-D62]

The main design principles of the Hexa-X M&O structural view are the following:

- Clear split between Managing Objects and Managed Objects, aligned with the OSI management protocols.
- Design Layer added to represent 3rd party software providers.
- Extreme-edge and public/private networks included as part of the Infrastructure Layer.
- Four controls loops included, for Infrastructure, DevOps, Network Layer, and Service Layer.
- Functions are associated in different groups at the Network Layer, and they are generic. The M&O Scope includes AI/ML Functions and Security Functions.
- AI collaborative components are distributed across the network.
- Communication among layers (and with external resources) is carried out through the API Management Exposure block, following a cloud-native approach.

The smart network management enablers identified in Hexa-X-II presented in this document are compliant with the aforementioned principles and aim to go a step further providing a system design approach of the main innovative aspects that were conceptually described in Hexa-X, e.g. orchestration across the compute continuum, the inclusion of control closed loops, or the concept of AI-based orchestration, with objective of further increasing level of automation and performance. Besides, new aspects are being included in Hexa-X-II smart management enablers such as network digital twins, programmability of transport networks through SDN, and more stringent requirements regarding KVI/KPI are being considered, with inclusion of specific enablers on trustworthiness for 3rd party management and AI/ML, as well as mechanisms to contribute to environmental sustainability.

1.3 Structure of the document

The current document is structured as follows:

- Section 1 sets the context for the document and provides an overview of its contents. It outlines the objectives and goals of the document, explaining its importance in the context of the Hexa-X-II project and how it relates with other work and documents in the project, as well as with previous work from Hexa-X project. Additionally, this section introduces the structure of the document and how each section contributes to achieving the overall objective.
- Section 2 focuses on the drivers behind the development of 6G technology. The section firstly identifies and analyzes the environmental, social, and economic factors influencing the adoption and deployment of 6G technology. It explores the potential benefits and challenges associated with these drivers as well as the implications into the enablers later presented in this document. The main considered KPI, related to Hexa-X KPI, are presented which will be used later on in the project to measure the success and efficacy of the proposed management and orchestration strategies. Finally, the overarching objectives and goals for the management and orchestration of the Hexa-X-II project are elaborated.
- Section 3, which is the main part of the document, provides a comprehensive list of the identified enablers proposed for the effective management and orchestration of 6G technology. For each enabler, the document presents motivation, objectives, description, State of The Art and beyond SoTA analysis, analysis of components and interfaces, and relationship with other enablers.
- In Section 4, the document outlines the planned Proof-of-Concept scenarios, where selected enablers are planned to be demonstrated. It describes the methodology, testbeds, and expected outcomes of these proofs of concept, aiming to validate the feasibility and effectiveness of the proposed strategies.
- Section 5 of the document provides a summary of the key findings and insights, and presents potential future directions for further work.
- In the Annex, the current State of the Art (SoTA) in 6G technology relevant for M&O is provided. It is divided into three subsections: SDO and Open-Source Software, Industry Fora and Research Projects. SDO and Open-Source Software focuses on the standards and advancements proposed by standard development organizations (SDOs) and contributions from the open-source software community. Industry Fora covers the contributions and developments from various industry fora, explores the ongoing efforts and collaborations within the industry to advance 6G technology, and identifies key trends and innovations. Research Projects subsection discusses noteworthy research projects in the EC SNS domain. It examines their methodologies, and potential implications for the Hexa-X-II project.

2 Management & Orchestration Drivers and Objectives Analysis

This section provides an overview of the main KVIs and KPIs as defined in Hexa-X [HEX23-D14] and Hexa-X-II [HEX223-D11] that can be affecting the enablers design related to management and orchestration.

Then, the settled objectives for Hexa-X-II on Smart Network Management are elaborated, indicating the related KPIs as well as the initial identification of enablers for each of them. When analysing the proposed objectives in the context of the M&O framework, it is important to decompose the objectives into smaller subobjectives to ensure that each objective is achieved successfully. In addition, related KPIs should be provided for each subobjective so as to later evaluate its performance and effectiveness.

This objectives' analysis has allowed to identify the set of enablers that we aim to design in order to accomplish those objectives. The identification of those enablers is introduced in this section for each of the objectives (and will be the main scope of the following section 3).

NOTE: Aligned with the rest of the Hexa-X-II project terminology, the term "*enabler*" in Hexa-X-II is defined as something that facilitates, empowers, or makes possible a particular action, process, or outcome in network management procedures. Depending on the specific enabler, this is either a tool, a technology or a specific resource that enables or contributes to the achievement of a desired goal or objective. An enabler can comprise one or more components.

The enablers should be designed to support the development and deployment of the M&O framework and ensure that the objectives are achieved successfully.

2.1 Environmental, Social and Economic 6G Drivers and KVI analysis

In Hexa-X-II D1.1 [HEX223-D11], a set of foundations are provided for the design of a sustainable, inclusive and trustworthy 6G. These foundations consider environmental sustainability, social sustainability and economic sustainability aspects, taking into account that these aspects are inseparable and have direct and indirect impacts on one another. Environmental sustainability refers to the development of an 6G ecosystem that include mechanisms for lifecycle assessment of the environmental footprint of 6G technologies. Social sustainability focuses on the well-being of individuals and communities and the production of positive social impact through the 6G technologies. Economic sustainability aims to achieve long-term economic growth, considering activities that are aligned with the environmental and social sustainability. Following, we shortly refer to each of these aspects as drivers for the development of 6G technologies, the motivation for designing the different enablers in line with these drivers, and the design of different enablers.

2.1.1 Driver #1: Environmental Sustainability

Driver #1 for 6G networks is Environmental Sustainability. It aims to reduce the impact of the network on the environment by decarbonizing electricity and processes, promoting circularity, and reducing raw material use. One approach to achieving this is to reuse 5G and legacy equipment as much as possible, instead of building entirely new infrastructure.

In addition, it is important to build modular and durable equipment, including devices, to reduce the amount of e-waste and increase the lifespan of the equipment. The goal is to target net-zero carbon emissions, digital inclusion, circular economy, and biodiversity.

To measure the success of achieving these goals, KPIs such as carbon footprint reduction, raw material use reduction, and e-waste reduction can be used. In addition, monitoring the circularity of the network and the lifespan of the equipment can also serve as KPIs to measure the success of achieving the objectives related to environmental sustainability.

Enablers such as renewable energy sources and sustainable design principles can support the development and deployment of the M&O framework for 6G networks in a way that is aligned with the objectives related to environmental sustainability (**see more details in enabler described in section 3.7**). By incorporating these enablers into the development process, the M&O framework can be optimized to reduce its impact on the environment and contribute to a sustainable future.

2.1.2 Driver #2: Social Sustainability

It aims to ensure that the network is trustworthy and promotes digital inclusion. Trustworthiness is essential for the successful deployment of the M&O framework. It includes factors such as security, privacy, and reliability. KPIs such as the number of security breaches, privacy violations, and system downtime can be used to measure the success of achieving the objectives related to trustworthiness.

Digital inclusion is another critical factor in social sustainability. It is essential to ensure that everyone has access to the benefits of the 6G network. This includes people living in rural or remote areas, those with disabilities, and those with low incomes. KPIs such as the percentage of the population with access to the 6G network and the percentage of underserved communities that have access to the network can be used to measure the success of achieving the objectives related to digital inclusion.

Enablers such as user-centered design and data governance frameworks can support the development and deployment of the M&O framework for 6G networks in a way that is aligned with the objectives related to social sustainability. By incorporating these enablers into the development process, the M&O framework can be optimized to promote trustworthiness and digital inclusion, leading to a more sustainable and equitable future. **Enablers in sections 3.4 and 3.8 especially focus on trustworthiness and security aspects**.

2.1.3 Driver #3: Economic Sustainability

It aims to ensure that the network is designed and operated in a way that is economically sustainable, while also providing value to its users. One approach to achieve this is to adopt a value-based design for the 6G network. This means that the design should be based on the value it provides to its users, rather than just the technology itself. KPIs such as customer satisfaction, revenue growth, and cost efficiency can be used to measure the success of achieving the objectives related to value-based 6G design.

Another aspect of economical sustainability is to develop a sustainable business model for the 6G network. This includes finding ways to generate revenue that are financially sustainable in the long term, while also being socially and environmentally responsible. KPIs such as profitability, return on investment, and social and environmental impact can be used to measure the success of achieving the objectives related to sustainable 6G business model innovation.

Moreover, building the 6G network to meet the needs of urbanization is an essential part of economical sustainability. This includes factors such as scalability, flexibility, and interoperability that are key design drivers considered in multiple enablers presented in this document. KPIs such as network coverage, capacity utilization, and network reliability can be used to measure the success of achieving the objectives related to building the 6G network to meet urbanization needs.

Open innovation and co-creation can support the development and deployment of M&O solutions for 6G networks in a way that is aligned with the objectives related to economical sustainability. By incorporating open innovation into the development process, the M&O framework can be optimized to provide value to its users while also being economically sustainable, leading to a more viable and prosperous future. **Enablers in sections 3.1, 3.2, 3.3, 3.5, 3.6, 3.9, 3.10, and 3.11 contributes to increase network automation, which in turn contributes to reduce Operational Costs (OPEX) and therefore a key aspect for economic sustainability.**

2.2 Hexa-X-II Smart Network Management Objectives Analysis

The enablers presented in this document (section 3) are in the scope of the Smart network management and orchestration (M&O) pervasive functionalities that Hexa-X-II aims to implement. Those M&O functionalities need to consider both network and cloud resources as part of the 6G overall system, while contributing to achieving 6G KPIs and KVIs: ultra-low zero-perceived latency, trustworthiness and sustainability. Furthermore, Hexa-X-II M&O mechanisms aim to widely exploit programmability as well as AI/ML techniques. Network intelligence will be applied not only to increase automation, but also, to abstract network complexity given the high heterogeneity of resources in 6G. Besides, concrete AI-based network management algorithms will be designed to reduce energy consumption and therefore contributing to achieve the sustainability goal. Hexa-X-II will also provide specific inter-domain orchestration mechanisms jointly for

cloud and network resources, taking care of guaranteeing isolation among potentially shared resources by different tenants, as well as control separation in order to ensure a trustworthy ecosystem. The overall goal of the M&O functionalities is divided in Hexa-X-II in the following five objectives, which are depicted in Figure 2-1 and subsequently explained:

- Design and develop a programmable **cloud-native micro-service-based Management and Orchestration (M&O) framework** for the future 6G networks, which is represented in the centre of the figure as all the rest of objectives/enablers will be relying on it.
- Design and develop mechanisms that collectively define a 6G enabled **trustworthy** environment, with a **user-centric integration fabric** that ensures **multi-tenancy** support and SLA verifiability.
- Develop **synergetic orchestration** mechanisms for managing the deployment of 6G services over heterogeneous resources **across the IoT-to-edge-to-cloud continuum**.
- Design and implement **robust and trustworthy AI/ML-based network control** solutions with optimal **energy efficiency and sustainability** targets.
- Design and develop **zero-touch M&O mechanisms for closed loop automation** and continuous service assurance, guaranteeing compliance with relevant 6G KPIs while reducing OPEX.

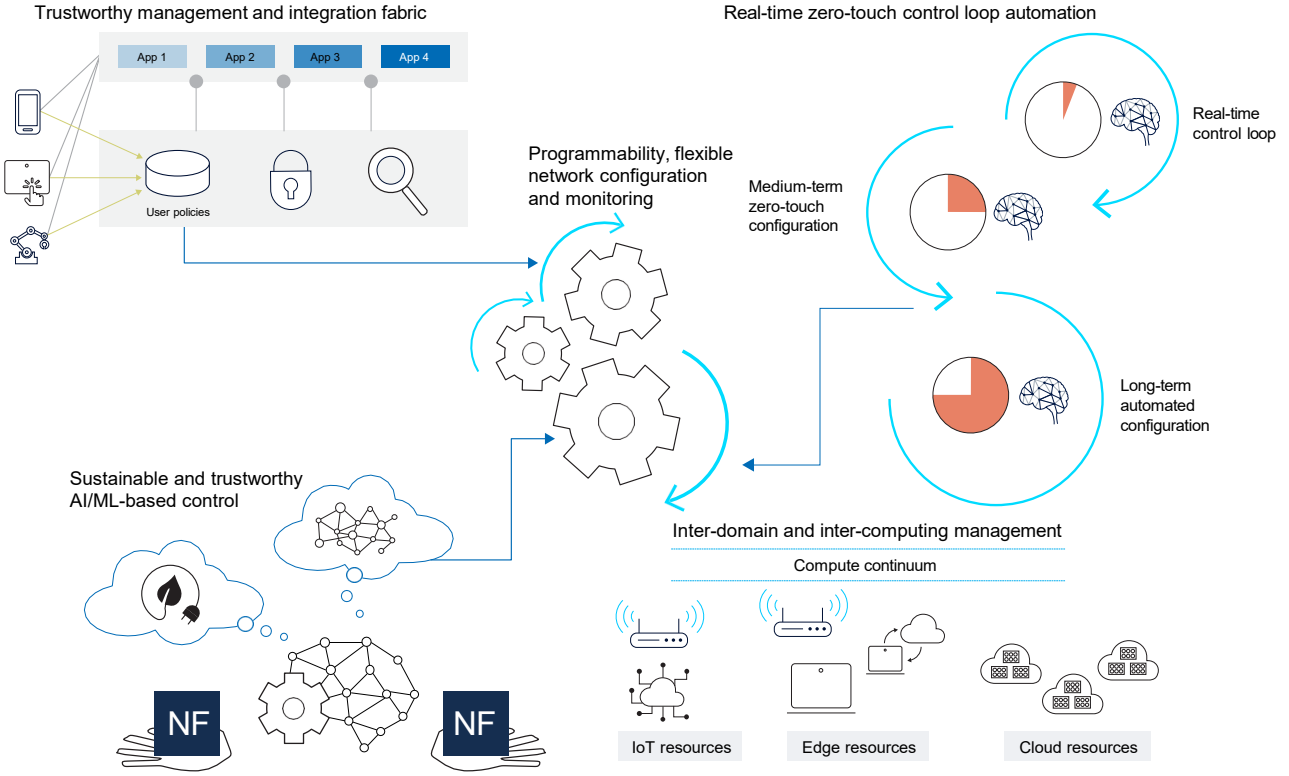


Figure 2-1: Hexa-X-II technical concepts on Smart Network Management

In [HEX22-D62], Hexa-X project defined Key Performance Indicators (KPIs) and Core Capabilities regarding Management and Orchestration (M&O) for 6G. These KPIs are important to ensure the successful development and deployment of the M&O framework with the expected performance in 6G networks. Hexa-X-II KPIs will be provided by Dec'23. In order to identify possible KPI/KVIs at this preliminary stage, we have based them on the previous KPI introduced in Hexa-X.

2.2.1 Cloud-native micro-service-based M&O framework

Design and develop a programmable cloud-native micro-service-based Management and Orchestration (M&O) framework for the future 6G networks.

2.2.1.1 Subobjectives

- Requirement definition of programmable cloud-native micro-service-based Management and Orchestration (M&O) framework. These requirements should include features such as scalability, reliability, flexibility, programmability, security, and automation. The proposed requirements shall be aligned with proposed in [HEX21-D12].
- M&O architecture definition with focus on scalability, flexibility, and resiliency. The proposed architecture shall be aligned with End-to-End architecture as described in [HEX223-D21].
- The M&O framework components may include:
  - Service discovery: A service discovery component helps to discover and locate the microservices that are part of the M&O framework.
  - API gateway: An API gateway provides a unified interface for accessing microservices.
  - Orchestration engine: An orchestration engine helps to coordinate the activities of microservices to achieve a specific goal.
  - Monitoring and analytics: A monitoring and analytics component helps to collect data and analyse it to optimize the performance and reliability of the M&O framework.
- Analyse proposed M&O frameworks: Once the architecture, components, and technology stack are defined, it is necessary to analyse the specific solutions that might fit the proposed architecture of the M&O framework.

2.2.1.2 Cloud-native micro-service-based M&O framework KPI/KVI

Some of the identified relevant KPI/KVIs for the proposed M&O objective are:

- Latency [s]: This KPI measures the time it takes for a request to be processed by the M&O framework. In a 6G network, where latency is expected to be extremely low, this KPI becomes particularly important. A low latency indicates that the M&O framework can process requests quickly, leading to a better user experience.
- Creation time [s]: This KPI measures the time it takes to create and deploy a new microservice within the M&O framework. A low creation time is important as it allows developers to quickly deploy new microservices and update existing ones, leading to faster innovation and improvements in the M&O framework.
- Availability [%]: This KPI measures the percentage of time the M&O framework is available and can handle requests. In a 6G network, where high availability is critical, this KPI becomes particularly important. A high availability percentage indicates that the M&O framework can handle requests even during high traffic periods or in the event of failures.
- Reliability [%]: This KPI measures the ability of the M&O framework to perform consistently and accurately over time. In a 6G network, where reliability is crucial, this KPI becomes particularly important. A high reliability percentage indicates that the M&O framework can consistently perform as expected, leading to a better user experience.

2.2.1.3 Identification of Enablers

The current objective is planned to be achieved by the design and implementation of the following two enablers:

Enabler 1: Programmable flexible network configuration

This enabler allows the M&O framework to configure and adapt the network based on changing requirements, traffic patterns, and use cases. A programmable network configuration enables the M&O framework to be flexible and agile, allowing it to handle dynamic workloads and adapt to changes in the network topology.

Enabler 2: Programmable network monitoring and telemetry

This enabler allows the M&O framework to collect and analyse data about network performance and user experience. By providing rich telemetry data, the M&O framework can identify performance bottlenecks, diagnose issues, and optimize network performance.

2.2.2 6G enabled trustworthy environment

Design and develop mechanisms that collectively define a 6G enabled trustworthy environment, with a user-centric integration fabric that ensures multi-tenancy support and SLA verifiability.

2.2.2.1 Subobjectives

- Identification of tenets for Hexa-X-II system to become a trustworthy 6G service platform that can be programmatically accessed by 3rd parties, including customers from B2B and B2B2C market segments (e.g., enterprise customers, application developers and aggregators). The purpose is to allow a secure yet frictionless network-application integration, to enhance user experience or unleash new use cases in the digital service ecosystem.
- Designing and developing operator internal mechanisms to provide 3rd parties with segregated yet customized management spaces to control their applications and manage their interaction with Hexa-X-II M&O framework resources.
- User-centric network management. Offer end users with optimal and personal experience, according to i) user preferences, ii) service requirements, and iii) network context.
- Assisting end users to gain access to hosted applications, while keeping their data securely stored, preventing any unauthorized entity to read/modify them.
- SLA translation and Trust Level Agreements (TLA), with the design of mechanisms that allow translating SLA components (<KPI list>, <KVI list>, <TLA list>) into network level configuration/control and monitoring actions, within and across domains.
- SLA enforcement, with the design of closed loop mechanisms that ensure that the SLA components are met during the service lifetime, and that allow 3rd party (and associated users) to continuously verify them.
- Designing M&O framework as a service-based management architecture, with capabilities conveyed into fine-grained loosely coupled microservices deployed (and quickly updated) in highly distributed environments. The scope includes network, cloud, management, and AI capabilities. This architectural approach allows 3rd parties to programmatically customize (and aggregate) capabilities according to their specific needs and scoped spaces.
- Analyzing plausible service bus options allowing for secure, reliable and observable communication between the microservices forming the M&O framework, including service mesh and message broker solutions, and compare their merits/drawbacks.
- Identifying a subset of mechanisms for Hexa-X-II PoC#B.1, prototype a service bus for M&O framework governance, and integrating it.

2.2.2.2 6G enabled trustworthy environment KPIs/KVIs

Some of the identified relevant KPI/KVIs for the proposed M&O objective are:

- Latency [s]: This KPI measures the time it takes to exchange a request-response message in the M&O framework. This exchange includes i) capturing the request issued by the API consumer, ii) routing it to the target API consumer, for its processing, iii) request processing time, and iv) send response back to the API consumer. A low latency indicates that the M&O framework can process requests quickly, leading to a better experience for framework consumers (including 3rd parties).
- Programmability [%]: This KPI measures the ease for M&O framework consumers to configure and compose (aggregate) existing capabilities, following cloud-native practices in the context of microservice-based architecture. A high programmability percentage means that the M&O framework can be i) quickly customized, to suit different requirements and use cases, and ii) easily extended/upgraded, with the rollout/roll-back of new features.
- Availability [%]: This KPI measures the percentage of time that the service bus is up and running, exchanging messages and process data across microservices. It reflects the traffic overload protection and protection to denial-of-service attacks.
- Reliability [%]: This KPI measures the ability of the service bus to exchange messages within a configurable upper latency bound, which is always lower than the timeout associated to HTTP protocol.

- Security by design [Boolean]: The M&O system natively incorporates cryptographically secure techniques to mutually authenticate individual microservices and encrypt the traffic between, and access control solutions to authenticate, authorize and later audit interactions with external consumers.
- Scalability [%]: This KPI measures the number of the 3rd parties (tenants) that are able to gain access (and duly authorize and audit with proper access control mechanism) to Hexa-X M&O framework resources, for the purpose of monitoring, control and/or configuration.

2.2.2.3 Identification of Enablers

The current objective is planned to be achieved by the design and implementation of the following two enablers:

Enabler 3: Integration fabric

This enabler allows the M&O framework to become a service-based management architecture (SBMA), with management functions split into loosely coupled microservices that can be dynamically deployed/removed in highly distributed infrastructures, using serverless solutions.

Enabler 4: Trustworthy 3rd party management

This enabler allows the M&O framework to provide three tenets: i) resource controllability, with the provision of segregated yet customized management spaces to 3rd parties; ii) user-centric network management, with personalized service experiences while respecting EU regulation with regards to data privacy and consent management; and iii) SLA enforcement and their verifiability from 3rd parties (and associated users).

2.2.3 Synergetic orchestration mechanisms across the compute-continuum

Develop synergetic orchestration mechanisms for managing the deployment of 6G services over heterogeneous resources across the IoT-to-edge-to-cloud continuum.

2.2.3.1 Subobjectives

- Manage deployment of network services and applications across multi-cluster resources in the computing continuum from IoT to edge to cloud.
- Develop intent-driven orchestration mechanisms for deployment of cloud-native applications with strict QoS requirements (e.g., latency-sensitive application components) across resources in the continuum.
- Develop synergetic/collaborative orchestration mechanisms based on hierarchical decision-making and distribution of control points in multiple orchestration entities (e.g., by taking advantage of multi-agent systems and collaborative AI mechanisms).
- Examine solutions where interaction between network providers and OTT players (edge/cloud providers) may take place for optimal deployment of applications. Check network northbound APIs that can be consumed by edge/cloud computing orchestration platforms.
- Develop AI-assisted orchestration mechanisms, taking advantage of modern cloud-native observability stacks.
- Develop federated data management approaches across the continuum, considering security and privacy aspects.

2.2.3.2 Synergetic orchestration mechanisms KPI/KVI

Some of the identified relevant KPI/KVIs for the proposed M&O objective are:

- Deployment time [s]: time required from moving from a deployment request to an operational instance of a network service graph or an application graph.
- Availability [%]: percentage of time that the orchestration platform is up and running in a fully-functional mode.
- Elasticity/Scaling time [s]: average time for creating new instances of horizontally scalable components.

- Reliability [%]: the ability of the orchestration platform to perform consistently and accurately over time.
- Programmability [%]: the percentage of functions offered by the orchestration platform that can be managed (e.g., through open Application Programming Interfaces (APIs))
- Maintainability [Degree]: zero-touch mechanisms can be applied to self-healing actions, detecting failures and restoring the communication services in an autonomous manner.
- Automation [Degree]: the orchestration mechanisms that operates in a fully autonomous manner, with reactive, proactive or predictive approaches, without requiring any intervention from the system administrator.
- Scalability [Degree]: the orchestration mechanisms working in parallel on various network domains, slices and services can help to efficiently manage more scalable networks reducing the complexity of handling manually the interdependencies across distributed networks and co-existing services.

2.2.3.3 Identification of Enablers

The current objective is planned to be achieved by the design and implementation of the following two enablers:

Enabler 5: Multi-cloud management mechanisms

This enabler focuses on the development of solutions for the deployment of network services and applications over a multi-cluster/cloud infrastructure. Multi-cluster management tools are considered (e.g., Liqo, Karmada, Open Cluster Management) where resources across the continuum can be registered and be manageable. Intent-driven hierarchical decision-making mechanisms are going to be considered.

Enabler 6: Orchestration mechanisms for the computing continuum

This enabler focuses on the development of orchestration mechanisms for the computing continuum, considering proper resource abstraction for the representation of devices from IoT to edge to cloud infrastructure. Collaborative AI mechanisms are going to assist orchestration platforms to improve the efficiency of the provided actions and to facilitate collaboration among different orchestration entities.

2.2.4 Robust and trustworthy AI/ML-based network control solutions

Design and implement robust and trustworthy AI/ML-based network control solutions with optimal energy efficiency and sustainability target.

2.2.4.1 Subobjectives

- Design of natively secure and sustainable AI/ML-based network control solutions.
- Develop effective and performant AI/ML-based network control solutions for reaching a trade-off between energy efficiency and performance for the network in model training and inference.
- Analyse the applicability and efficiency of different defence mechanisms against adversarial and privacy attacks targeting AI/ML models.
- Incorporate explainable AI methods to improve the trustworthiness of AI/ML based control solutions.
- Design and develop accurate network digital twins to enable a safe and effective AI/ML model training.

2.2.4.2 Robust and trustworthy AI/ML-based network control solutions KPI/KVI

Some of the identified relevant KPI/KVIs for the proposed M&O objective are:

- Energy savings [W]: This KPI measures the ability of the network control solutions to minimize energy consumption and is a crucial metric for assessing energy efficiency and sustainability according to targets.
- Availability [%]: This KPI measures the percentage of time the network and AI/ML control solutions are available, which reflects their robustness to faults and attacks.

- Performance Consistency [%]: This KPI assesses the ability of AI/ML control solutions to achieve their performance objectives consistently and accurately over time. It's important to note that the interpretation of performance consistency can vary; while system reliability and AI solution reliability may differ as metrics, this measure focuses specifically on the consistency of AI/ML control solutions in meeting their goals. A high-performance consistency score indicates the effectiveness and robustness of these control solutions.
- AI/ML models training (and inference) time [s]: This KPI measures the time and indirectly the number of resources and energy consumed from training the AI/ML models and performing inference, which has an impact on the energy efficiency and sustainability targets.
- Security by design [Boolean]: A network control system that is secure by design natively incorporates the security and defence mechanisms which improves robustness and trustworthiness.
- Resiliency [%]: The ability of the network control system to mitigate adversarial or security attacks and continue to operate correctly improves resiliency and trustworthiness in its capabilities to deliver the required performance under all circumstances.
- Carbon footprint [kg CO_2e]: This KPI measures the total greenhouse gas (GHG) emissions caused directly or indirectly by the operation of AI/ML-based network control solution, and the application of decisions made by the network control solution.

2.2.4.3 Identification of Enablers

The current objective is planned to be achieved by the design and implementation of the following three enablers:

Enabler 7: Sustainable AI/M-based control

This enabler allows the AI/ML-based control solutions to minimize energy consumption in the network while maintaining performance. It also minimizes the energy consumption of the AI/ML control loop itself by reaching a trade-off between performance and energy usage. Reducing energy consumption of the network and the AI/ML control loop contributes to reaching the energy efficiency and sustainability targets.

Enabler 8: Trustworthy AI/M-based control

This enabler focuses on providing security solutions for the AI/ML-based control loop by incorporating defence measures against adversarial and privacy attacks. It also leverages solutions from Explainable AI (XAI) to provide human readable explanations for the reasoning behind the output of ML models. Those methods help to make the AI/ML-based control solutions more robust, secure and trustworthy.

Enabler 9: Network Digital Twins

This enabler allows the usage of digital replicas of the networks that share the same high-level properties and replicate their behaviour to train the AI/ML models with no impact on the network from the exploration decisions. The network digital twins also allow the prediction of service KPIs to perform proactive measures and are constantly updated to accurately represent the state of the network, which improves the robustness and trustworthiness of the AI/ML-based control solutions.

2.2.5 Zero-touch M&O mechanisms

Design and develop zero-touch M&O mechanisms for closed loop automation and continuous service assurance, guaranteeing compliance with relevant 6G KPIs while reducing OPEX.

2.2.5.1 Subobjectives

- Designing an architecture for zero-touch automation operating at the service, network and infrastructure layers.
- Identifying functions, open interfaces and workflows for zero-touch automation based on CL mechanisms applicable to a selected set of multi-dimensional scenarios:
    - Time scale: CL for real-time and short-term control actions vs. medium or long term (re-) planning and optimization actions.

  - Domain scope: CL at access, transport, and core network domain; CL at extreme edge, edge and cloud domains.
  - Layer scope: CL at infrastructure, network, and service layer.
  - Stakeholder scope: CL with per-slice, per-tenant, per-MNO, per-service provider scope.
- Defining functions for the various CL stages (monitor, analyse, decide, execute), as well as technologies and interfaces for each of them, including their interaction with the functions of the overall 6G M&O system.
- Analysing deployment models for virtualized CL functions, including workflows for their instantiation and lifecycle management and efficient strategies for their placement in the extreme edge/edge/cloud continuum.
- Analysing the applicability and efficiency of reactive, proactive and predictive strategies in a relevant set of zero-touch automation scenarios.
- Identifying a subset of CL functions and zero-touch automation workflows for Hexa-X-II PoC#A.1 and PoC#B.1 (see Sections 4.1 and 4.2), for their implementation and integration.
- Defining strategies to combine zero-touch automation mechanisms working with different scopes and time scales, coordinating multiple concurrent CLs with the objective of guaranteeing the overall stability of the system and the inter-domain/inter-layer consistency.
- Identifying the implications of coordinating multiple CLs operating at different administrative domains, in terms of interfaces, data ownership, capability exposure, access control.
- Analysing alternative models for cooperative CLs, including peer, hierarchical and nested CLs.
- Analysing alternative CL collaboration models, including CL delegation and escalation, and identifying strategies for their mix in cross-layer and cross-domain automation.
- Design, implementation and evaluation of a representative subset of mechanisms for (i) conflict detection, mitigation, and resolution; (ii) global assessment of decisions from multiple concurrent CLs; (iii) coordination and arbitration of multi-objective CLs.

2.2.5.2 Zero-touch M&O mechanisms KPI/KVI

Some of the identified relevant KPI/KVIs for the proposed M&O objective are:

- Latency [s]: zero-touch automation mechanisms can help to reduce the latency experienced by communication or application services, e.g., triggering automatically the scaling of network or service functions to avoid overloading or replicating/moving UPFs and edge application functions closer to the users.
- Storage Capacity and Processing Capacity optimization [%]: zero-touch M&O can automatically optimize the usage of storage and processing resources on the basis of their dynamic availability in the extreme-edge/edge/cloud continuum and the requirements of the running, planned or predicted services.
- Programmability [%]: programmability is the key enabler for CLs' implementation, both in terms of monitoring and execution actions. Moreover, a fundamental sub-objective for this research item is to make the CL functions themselves fully programmable, through the support of open and potentially standard APIs. This would simplify their integration in the 6G management system and allow to coordinate multiple CLs, including their internal stages, in a programmable manner.
- Energy savings [W]: zero-touch automation can adopt the energy efficiency as one of the optimization criteria to take decisions and drive re-configuration actions, e.g., moving functions in aggregated computing points or preferring green-powered computing resources in the allocation strategies.
- Reliability [%]: proactive or even predictive approaches can be applied in the various CLs, with early recognition of underperforming resources or service degradation that leads to anticipated re-configuration actions, thus improving the global reliability of the system.
- AI/ML models training time [s]: since AI/ML techniques can be applied to the analysis and decision stages of a CL, this KPI can influence the performance of the zero-touch automation itself – especially in case of re-training. Therefore, the CL governance mechanisms for CL functions' provisioning and lifecycle management should allocate resources for ML models (re-)training (when applicable) taking into account this constraint.
- Maintainability [Degree]: zero-touch mechanisms can be applied to self-healing actions, detecting failures and restoring the communication services in an autonomous manner. This contributes to the

level of maintainability of the entire system. Moreover, the programmability of the CL functions gives the network administrator the possibility to operate and configure the various CLs for maintenance purposes.
- Elasticity [Degree]: zero-touch automation decisions and actions can be applied to the dynamic optimization of the resource usage on the basis of the variable availability of storage and computing capacity in the various domains of the infrastructure, in compliance with the service requirements.
- Automation [Degree]: the zero-touch automation is based on CLs that operate in a fully autonomous manner, with reactive, proactive or predictive approaches, without requiring any intervention from the system administrator.
- Scalability [Degree]: the coordination of multiple CLs working in parallel on various network domains, slices and services can help to efficiently manage more scalable networks reducing the complexity of handling manually the interdependencies across distributed networks and co-existing services.

2.2.5.3 Identification of Enablers

The current objective is planned to be achieved by the design and implementation of the following two enablers:

Enabler 10: Zero-touch closed loop governance

This enabler allows the M&O framework to manage the lifecycle of single CLs, handling the provisioning and runtime management of virtual functions for the CL stages (monitoring, analysis, decision, execution), their interconnectivity with the other M&O functions, and providing open interfaces for the governance of CL stages and the sharing of related information from/with external elements.

Enabler 11: Zero-touch multiple closed loop coordination

This enabler allows the M&O framework to coordinate multiple concurrent CLs operating with different scopes or time scales, to guarantee the overall consistency of their decisions (reducing potential conflicts, combining multiple objectives and/or arbitrating among contrasting decisions) and the correct order in the execution of the related commands.

3 Initial design of Smart Network Management Enablers

This section goes into the details of the enablers that have been identified in the previous section based on the Smart Network Management objectives analysis for the effective management and orchestration of 6G Technology. Figure 3-1 represents how the different enablers map into the current end-to-end Hexa-X-II system blueprint [HEX223-D21]. All enablers tackled in this document are part of the Management & Orchestration functional block, while there are some of them that would be also part of other pervasive functionalities: this is Enabler#2 being part of the data collection framework, Enablers #7 and #8 as part of the AI framework and Enablers #4 and #8 part of the Security & Privacy one. Besides, it is worth to highlight that though Enablers #1, #5 and #6 are also part of the M&O block they are tightly related with the resource infrastructure layer.

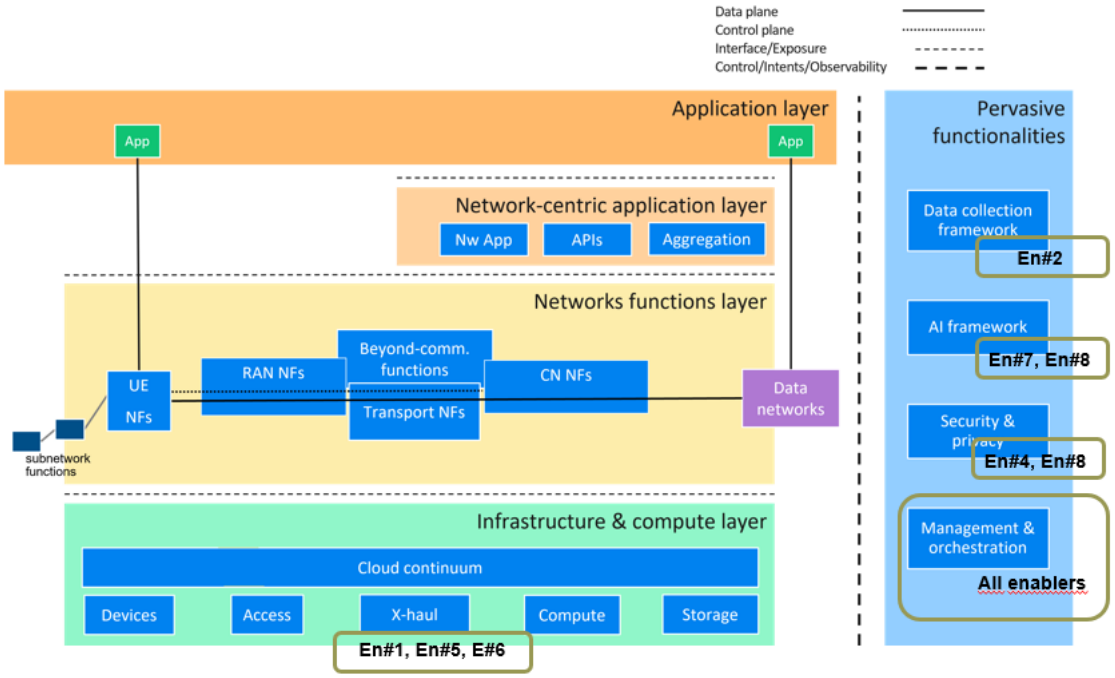


Figure 3-1: Hexa-X-II initial system blueprint and its relationship with M&O enablers

For each enabler, this section presents its motivation and problem it solves, its objectives, high-level description, relevant State of the Art (SoTA) solutions considered as starting point and how each enabler goes beyond SoTA, its initial envision components and interfaces, and its relationship with other enablers.

3.1 Enabler 1: Programmable flexible network configuration of transport networks

3.1.1 Motivation

The need for programmable and flexible network configurations has become essential in upcoming 6G networks. Transport networks have traditionally been often static and inflexible and were not able to keep up with the changing requirements and traffic patterns of digital technologies such as cloud/edge computing and IoT. By automating routine network operations and using programmable network configurations aligned with industry standards, organizations can improve network efficiency, reduce downtime, and increase network agility.

Legacy network infrastructure may be incompatible with new technologies, requiring significant updates or replacement. A programmable network that can adapt to new technologies and integrate them seamlessly is necessary to keep pace with changing business requirements. Furthermore, network security is a critical

concern for businesses, and a programmable network that can detect and respond to potential security threats in real-time is necessary to ensure business continuity and data protection.

In summary, programmable, and flexible network configurations are necessary to meet the changing demands of modern networks. By using programmable network configurations, organizations can improve network efficiency, agility, and security while keeping up with changing business requirements.

3.1.2 Objectives

The main objectives of Enabler 1 are:

- Enable Dynamic Network Configuration: The primary objective of this enabler is to facilitate dynamic network configuration through programmability. It aims to enable network administrators to easily configure and modify network elements, such as switches, routers, and access points, using software-defined networking (SDN) principles. This objective involves developing tools, protocols, and interfaces that allow network administrators to define and modify network configurations in real-time, enabling agile network management.
- Enhance Network Flexibility: The objective of enhancing network flexibility aims to enable rapid adaptation and reconfiguration of network elements to address changing network requirements. It involves developing flexible programming models, frameworks, and interfaces that enable network administrators to define and deploy customized network functions, protocols, and services. This objective empowers network operators to quickly adapt their network infrastructure to support new applications, accommodate traffic fluctuations, and optimize resource utilization.
- Enable Network Element Programmability: This objective focuses on enabling network element programmability, allowing administrators to define and modify the behavior of individual network elements. It involves programmable interfaces and control frameworks that facilitate the development and deployment of custom software on network devices. Network element programmability allows for the implementation of innovative forwarding strategies, customized protocol stacks, and specialized services to address specific network requirements and optimize network performance.
- Facilitate Network Management and Orchestration: This objective aims to provide robust management and orchestration capabilities for programmable networks. It involves developing tools, APIs, and frameworks that simplify the monitoring, configuration, and orchestration of programmable network elements. Network management and orchestration functionalities enable centralized control, automation, and policy enforcement across the network infrastructure, improving operational efficiency, reducing configuration errors, and enhancing network security.
- Ensure Interoperability and Standards Compliance: This objective focuses on promoting interoperability and adherence to industry standards for programmable network configurations. It involves supporting open interfaces and standardized protocols that enable interoperability between different vendors' programmable network elements. By promoting interoperability and standards compliance, this objective fosters vendor-neutral solutions, encourages innovation, and avoids vendor lock-in, ensuring a diverse ecosystem of programmable network technologies.

3.1.3 Description of the solution

The solution/enabler for "Programmable flexible network configuration" is designed to provide service providers with a flexible, programmable, and scalable approach to network management. This solution is based on several key technologies and approaches, including software-defined networking (SDN), application programming interfaces (APIs), and cloud-based network management platforms.

SDN provides a programmable and flexible approach to network configuration by separating the network control plane from the data plane. This separation enables service providers to centrally manage and orchestrate network policies and configurations, providing a high level of control and flexibility in network management. This allows for faster and more efficient network configuration, deployment, and maintenance.

APIs provide a standard way for applications and services to interact with the network infrastructure, enabling programmable automation and orchestration of network functions and policies. This allows for easy integration

of new applications and services into the network, as well as the ability to automate and orchestrate network functions and policies for improved operational efficiency.

Cloud-based network management platforms provide a scalable and flexible approach to network management and orchestration. These platforms allow service providers to manage and monitor network configurations and policies from a central location, regardless of the physical location of network devices. This provides a high level of flexibility and scalability in network management, enabling administrators to easily manage and orchestrate network configurations across large and geographically dispersed networks.

3.1.4 SoTA and Beyond SoTA

The related SoTA in programmable flexible network configuration includes Telecom Infra Project (TIP) Open Optical & Packet Transport (OOPT) and the Mandatory Use Case Requirements for SDN for Transport (MUST), which define the architecture and interfaces for this domain [TIP21]. Additionally, the open-source cloud-native ETSI TeraFlowSDN controller (TFS) plays a significant role in enabling programmability and flexibility in network configurations, as detailed in Section 7.1.6.2. These advancements contribute to the development of efficient and adaptable networks, empowering operators to optimize their infrastructure and deliver innovative services.

In addition to the SoTA, there are several key areas that are being explored to further enhance programmable flexible network configuration:

- Firstly, there is a focus on designing general interfaces that simplify the control of new device types through Device plugins and Service plugins. These interfaces aim to provide a standardized approach to incorporating diverse network devices into the software-defined networking (SDN) framework, enabling easier integration and management of a wide range of devices.
- Another important aspect is aligning the SDN architecture with ETSI MEC 015 [mec-015], specifically related to Enabler 6 (Orchestration mechanisms for the computing continuum), to offer Bandwidth Management (BWM) Services to MEC applications. By integrating BWM services into the SDN framework, network operators can efficiently allocate and manage bandwidth resources for MEC applications, ensuring optimal performance and quality of service for edge computing deployments. Furthermore, there is a focus on analysing and extending the Zero-touch Service Management (ZSM) architecture for SDN networks control and management, particularly related to Enablers 10 (Zero-touch closed loop governance) and 11 (Zero-touch multiple closed loop coordination). This involves exploring how the ZSM framework can be adapted and enhanced to effectively handle the control and management of SDN-based networks, enabling automated provisioning, configuration, and monitoring of network resources. To support these advancements, a proposal is made to contribute ETSI TeraFlowSDN (TFS) as a TIP reference implementation. This involves conducting a comprehensive analysis of TFS in a whitepaper, identifying any missing interfaces that may be required to align with TIP requirements. Additionally, the maturity of TFS is evaluated, determining its readiness for deployment and production environments. Based on this analysis, prioritization of features and functionalities is proposed to ensure the alignment of TFS with TIP standards. Lastly, support for ETSI TFS release 3.0 and 4.0 is emphasized to ensure compatibility and continued improvement of the TFS controller.

3.1.5 Identification of possible components and interfaces

Figure 3-2 presents the proposed initial high-level architecture for Enabler 1, which involves a hierarchical approach with an End-to-End (E2E) controller as the parent controller and technological domain controllers as child controllers, as detailed in [TIP21]. Specifically, it mentions an E2E SDN (Software-Defined Networking) orchestrator and technological domain SDN controllers in the IP (Internet Protocol), Optical, Time-Sensitive Networks (TSN) / Deterministic Networks (DETNET) and other domains.

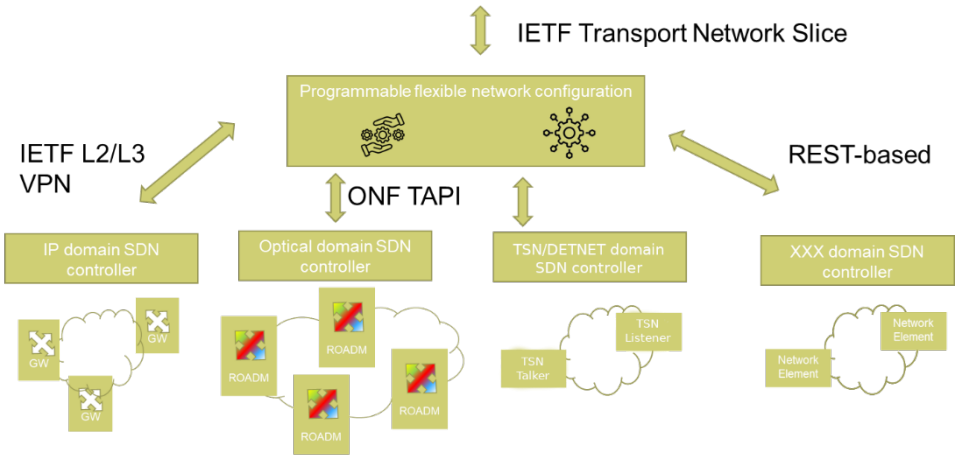


Figure 3-2: Proposed architecture for Enabler 1 - Programmable flexible network configuration

End-to-End (E2E) SDN Orchestrator is the parent controller responsible for managing and orchestrating the entire network infrastructure using Software-Defined Networking. The E2E SDN orchestrator oversees the coordination and control of the network across different technological domains. The E2E SDN orchestrator can rely on existing solutions, such as ETSI TeraFlowSDN (Section 7.1.6.2).

Technological Domain SDN Controllers are child controllers that operate within specific technological domains of the network. In this case, it mentions IP, Optical, TSN/DETNET, and other domains, but there may be additional domains depending on the network architecture. Each technological domain controller is responsible for managing and controlling the SDN functions and resources within its respective domain. Each specific technological domain controller might be based on several SNS project solutions, for example PREDICT-6G work could be used for DETNET.

Overall, the hierarchical approach with the parent E2E SDN orchestrator and child technological domain SDN controllers allows for a modular and distributed control of the network, where each controller handles specific domain-related tasks while the E2E orchestrator ensures end-to-end coordination and management across the entire network infrastructure.

3.1.5.1 E2E SDN Orchestrator

As the parent controller, the E2E SDN Orchestrator assumes the role of overseeing the coordination and control of the network across different technological domains. It acts as a unifying entity that harmonizes the operations of various network elements, ensuring seamless communication and efficient utilization of resources.

One of the primary objectives of the E2E SDN Orchestrator is to coordinate multi-domain network orchestration. In modern network architectures, different domains, such as data center networks, wide area networks (WANs), edge networks, and cloud infrastructure, coexist and interact to deliver end-to-end services. The E2E SDN Orchestrator plays a crucial role in integrating and managing network elements from diverse domains. By facilitating the coordination and orchestration of network operations across these domains, the E2E SDN Orchestrator enables consistent policies, optimized resource utilization, and efficient communication throughout the entire network infrastructure.

Furthermore, the E2E SDN Orchestrator supports service-level orchestration. In addition to managing the network infrastructure, it provides mechanisms for defining and managing services that traverse multiple network domains. By considering the end-to-end service requirements, the E2E SDN Orchestrator ensures that services are provisioned and delivered seamlessly across the network. It enables the dynamic allocation and optimization of network resources, allowing services to adapt to changing demands and conditions. The service-level orchestration capabilities of the E2E SDN Orchestrator enhance the overall efficiency, scalability, and agility of the network, providing a flexible and adaptable infrastructure for delivering a wide range of applications and services.

The exposed interfaces are described in Table 3-1.

Interface name	Service descriptions	Potential consumers	Possible Standards
NBI E2E SDN Orchestrator	Support for Transport Network Slices	OSS/BSS	IETF draft-liu-teas-transport-network-slice-yang
SBI E2E SDN Orchestrator	Layer 2 Virtual Private Network (L2VPN) Service Delivery	Technological Domain SDN controller	IETF – RFC 8466
SBI E2E SDN Orchestrator	Connectivity Service, Topology Discovery, Telemetry	Technological Domain SDN controller	ONF Transport API

Table 3-1: E2E SDN Orchestrator exposed interfaces

3.1.5.2 Technological-Domain SDN controller

The technological-domain SDN controllers are responsible for managing and controlling the SDN functions and resources within their respective domains. The domains can vary depending on the network architecture, but commonly mentioned domains include IP, Optical, TSN/DETNET, and others. Each Technological Domain SDN Controller focuses on a particular domain and is tailored to address the specific requirements and characteristics of that domain.

For instance, an **IP SDN Controller** would be designed to manage and control the IP-based network elements, such as routers and switches, within the network infrastructure. It would handle tasks like routing, traffic engineering, QoS enforcement, and network optimization within the IP domain.

Similarly, an **Optical SDN Controller** would specialize in managing and controlling the optical network elements, such as wavelength routers and optical switches. It would facilitate functions like lightpath provisioning, dynamic spectrum allocation, and optical resource optimization. The Optical SDN Controller enables programmability and control of the optical domain, allowing for efficient utilization of optical resources and seamless integration with other domains.

Another is **Time-Sensitive Networking (TSN) or Deterministic Networking (DETNET)**. These controllers are responsible for managing and controlling the real-time and time-sensitive communication requirements within the network. They ensure precise and deterministic delivery of critical data, including time synchronization, traffic scheduling, flow registration, resource reservation, and computing end-to-end routes. It remains to be seen whether these controllers can be implemented by building upon the work of research projects such as PREDICT-6G (Section 7.3.2.2).

In addition to the mentioned domains, there can be **other technological domains SDN controllers** depending on the specific network architecture and requirements. For example, domains like wireless networks, cloud infrastructure, IoT networks, or VNF could have dedicated SDN Controllers to manage and control their respective elements and functionalities.

By having technological domain SDN Controllers operating within specific domains, the network architecture becomes more modular and scalable. Each controller focuses on its domain's unique characteristics, allowing for optimized management and control of the associated SDN functions and resources. These controllers work in conjunction with the E2E SDN Orchestrator to achieve a comprehensive and efficient network management framework, providing end-to-end coordination and control across different technological domains.

The exposed interfaces are described in Table 3-2.

Interface name	Service descriptions	Potential consumers	Possible Standards
NBI (IP)	Layer 2 Virtual Private Network (L2VPN) Service Delivery	E2E SDN Orchestrator OSS/BSS	IETF – RFC 8466

SBI (IP)	Router configuration and monitoring	Technological Domain SDN controller	OpenConfig gNMI
NBI (Optical)	Connectivity Service, Topology Discovery, Telemetry	E2E SDN Orchestrator OSS/BSS	ONF Transport API
SBI (Optical)	ROADM configuration	Technological Domain SDN controller	OpenROADM

Table 3-2: Technological-Domain SDN controller exposed interfaces

3.1.6 Relationship with other Enablers

Enabler 2, Programmable network monitoring and telemetry, is closely related to the programmable flexible network configuration enabler. Telemetry and monitoring information provides service providers with real-time visibility into network performance and usage, enabling them to detect and respond to potential issues in real-time. By leveraging telemetry and monitoring information, administrators can automate the process of detecting and responding to potential issues, improving operational efficiency, and reducing the risk of downtime. In terms of the relationship between these two enablers, Enabler 1 focuses on day 1 operations, which involves the initial deployment and configuration of the network. Enabler 2, on the other hand, focuses on day 2 operations, which involves ongoing monitoring and management of the network. By using Enabler 2, service providers can make use of telemetry and monitoring information to react to potential issues in real-time, ensuring optimal network performance and reducing the risk of downtime.

Enabler 3, Integration Fabric, will expose Enabler 1 capabilities and northbound interfaces towards Operations Support System (OSS) and Business Support System (BSS). To this end, integration of both enablers will be considered.

Another important aspect is aligning the SDN architecture with ETSI MEC 015, specifically related to **Enabler 6** (Orchestration mechanisms for the computing continuum), to offer Bandwidth Management (BWM) Services to MEC applications.

The proposed advancements in programmable flexible network configuration have relationships with other enablers and tasks in the networking domain. They align with **Enablers 9 and 10**, which focus on close loop operation and involve automated decision-making processes.

3.2 Enabler 2: Programmable network monitoring and telemetry

3.2.1 Motivation

The increasing complexity of modern networks, including hybrid and multi-cloud environments, as well as distributed architectures and microservices, has made network monitoring a challenging task. With so many interconnected devices and systems, it can be difficult to gain a comprehensive view of the network as a whole, leading to blind spots and potential security vulnerabilities. One of the key problems addressed by programmable network monitoring and telemetry is the limited visibility provided by traditional network monitoring tools. Many legacy monitoring tools focus on individual devices or metrics, providing a narrow view of network performance that can make it difficult to identify problems and optimize network operations. By contrast, programmable network monitoring and telemetry solutions offer a more comprehensive view of the network, providing real-time data on end-to-end network behaviour. This enables network administrators to quickly identify and troubleshoot issues, reducing network downtime and improving operational efficiency. Another challenge faced by service providers is the time-consuming and error-prone nature of manual network monitoring and troubleshooting. With the expected complexity of upcoming 6G networks, it can be difficult to manually monitor every device and system, leading to missed issues and longer resolution times. By automating network monitoring and troubleshooting using programmable network monitoring and telemetry solutions, service providers can save time and reduce the risk of human error, leading to more reliable and efficient network operations.

One of the key benefits of programmable network monitoring and telemetry is improved network security. By monitoring network traffic in real time, administrators can quickly detect and respond to security threats, such as unauthorized access attempts or data exfiltration. This allows for a proactive approach to network security that can help prevent breaches and minimize their impact.

Another benefit of programmable network monitoring and telemetry is faster troubleshooting and remediation of issues. By collecting and analysing real-time data about network performance, administrators can quickly identify the source of issues and take corrective action. This can help minimize downtime and reduce the impact of network issues on users and applications.

Finally, current monitoring tools can provide focus on a wealth of data, but limited insights into network behaviour. Programmable network monitoring and telemetry solutions go beyond traditional monitoring tools by providing a more granular view of network performance and other network related metrics, such as energy consumption and carbon footprint, not present so much in current telco monitoring solutions, that however will be key in 6G. This will enable the detection of potential security threats and optimize network operations in real time. By leveraging technologies such as NETCONF, REST, YANG, gRPC, gNMI and SNMP, programmable network monitoring and telemetry solutions offer a powerful and flexible approach to network monitoring that can meet the needs of even the most complex and distributed networks.

3.2.2 Objectives

The objective of programmable network monitoring and telemetry is to collect and enable the analysis of service and network performance data, potentially in real-time, across multiple domains. This enabler focuses on developing mechanisms that allow network administrators to gather granular and accurate information about the network's behavior, performance, and utilization across multiple network domains (such as extreme edge, RAN, edge, transport, core, and service layer). The proposed solution must be agnostic to the processed data, as example, it could be used in the context of service-level monitoring and end-to-end validation.

By employing programmable network monitoring and telemetry, network administrators can dynamically define and collect relevant data points, such as traffic statistics, latency measurements, packet loss rates, and resource utilization metrics, from different network elements. The data can be collected in real-time or near real-time, providing up-to-date insights into the network's health and performance. Multi-vendor solutions will be considered, as well as authentication and privacy mechanisms will be studied in the scope of security, privacy and system level resilience.

Moreover, programmable network monitoring and telemetry shall enable the aggregation and correlation of performance data across multiple network domains. This objective involves developing frameworks and protocols that facilitate the seamless integration and analysis of data from various technological domains. By combining and correlating data from different domains, network administrators gain a holistic view of the network, enabling them to identify potential bottlenecks, diagnose performance issues, and optimize network resources more effectively.

Additionally, programmable network monitoring and telemetry framework shall support advanced analytics and ML techniques. By providing programmable interfaces and standardized protocols for data collection, this objective enables the integration of analytics tools and algorithms to process and analyze network performance data. This will allow for the identification of patterns, anomalies, and trends in the network's behavior, empowering network administrators to make data-driven decisions, proactively address issues, and optimize network performance and reliability.

3.2.3 Description of the solution

Programmable network monitoring and telemetry involves the use of SDN and multiple automation technologies to collect and analyse data about network traffic and performance in real time, from virtual and physical network elements and applications.

The proposed programmable network monitoring and telemetry enabler shall:

- collect, process and export data from multiple sources (e.g., from multiple layers, including services and applications) in a multi-vendor, scalable and flexible approach.

- enable better decision-making based on real-time data. By collecting and analysing data about network traffic and performance, administrators and network automation frameworks (such as provided by Enabler 10) can gain insights into network usage patterns, identify areas where network resources are underutilized, and optimize network configurations to improve performance and reduce costs.
- handle data on energy consumption of the different network elements (including compute resources) is also needed. These data will allow the analysis and development of algorithms that provide network deployments that consider leveraging energy consumption.
- leverage a range of technologies and protocols, including NETCONF, REST, YANG, gRPC, gNMI, and SNMP. These technologies enable network administrators to collect and analyse data about network traffic and performance in real time, automate network management tasks, and orchestrate network functions and policies.
- Introduce authentication and privacy mechanisms in relationship with security, privacy and system level resilience.
- consider cloud-scale scalability, and to avoid monitoring bottlenecks. To this purpose, metrics and alerts could be processed in a sequence of independent steps. Several necessary steps are envisioned, such as metric/alert acquisition, metric/alert normalization, metric/alert visualization, alert evaluation, and alert publication.

3.2.4 SoTA and Beyond SoTA

The SoTA in network monitoring and telemetry encompasses a great range of solutions catering to different scopes, including networks, services, applications, and end-to-end (E2E) measurements. Numerous monitoring solutions have emerged to address these diverse monitoring needs. These solutions provide visibility into network performance, service availability, application health, and the overall end-user experience. They enable operators to monitor and analyze various metrics and parameters to identify issues, optimize performance, and ensure the desired quality of service.

Two notable projects in the current landscape of network monitoring and telemetry are OpenTelemetry (Section 7.1.6.4) and Prometheus (Section 7.1.6.5). Beyond the SoTA in network monitoring and telemetry, there are several advancements and objectives that can be considered. From an architectural perspective, there should be a focus on introducing a scalable framework that integrates various components such as integration with GNMI protocol and OpenTelemetry, allowing for efficient and flexible monitoring capabilities. Additionally, some initial efforts are being made to contribute towards the analysis of monitoring extreme-edge devices, taking into account the unique challenges and requirements of these devices from the operational perspective. Moreover, the TSN/DETNET domain also demands special monitoring capabilities for delay-sensitive traffic, which are currently under investigation in wireless networks. Further research efforts will focus on extending this work to wired/wireless networks, investigating the trade-off between in-band and out-of-band telemetry, and algorithms to find the optimal placement of telemetry functions in such networks.

Data fusion of different signals is another area of interest beyond the SoTA, aiming to extend network metrics by combining multiple data sources to provide a more comprehensive view of network performance. Moreover, analysis mechanisms are being developed to correlate signals and identify events that require remediation actions, enabling proactive response to network issues.

In terms of APIs and components, there is an emphasis on adapting and integrating the OpenTelemetry specification, facilitating seamless integration and interoperability with existing monitoring solutions. This ensures consistency and standardization in telemetry data collection and analysis.

Regarding energy monitoring, there are several solutions in the SoTA that could be considered, e.g. Kepler [KEP23], Scaphandre [SCA23], which are currently not being used in telecom networks. Beyond the SoTA, Scaphandre is planned to be used as part of the current Enabler 2 to improve energy observability metrics throughout the whole 6G continuum for both network and computing resources.

ETSI TeraFlowSDN (TFS) monitoring involves persisting key performance indicators (KPIs) and setting alarms to effectively monitor the performance of the TFS controller. The development of a scalable framework for TFS further enhances its monitoring capabilities, enabling efficient management of network resources. Additionally, the open-source 5G monitoring platform (Section 7.1.6.3) is being enhanced to facilitate its integration into closed loop systems, where monitoring data can inform automated remediation actions.

3.2.5 Identification of possible components and interfaces

A programmable network monitoring and telemetry architecture typically consists of three components: collectors, processors, and exporters (Figure 3-3).

- Collectors are responsible for getting data into the system. They can be push-based or pull-based. Push-based collectors send data to the system without being asked, while pull-based collectors only send data when they are requested.
- Processors are responsible for what to do with the data that is received from the collectors. They can perform a variety of tasks, such as filtering the data to remove irrelevant or duplicate data, transforming the data into a format that is easier to analyse, correlating the data with other data sources, and alerting on abnormal or unexpected events.
- Exporters are responsible for sending the processed data to other systems. They can be push-based or pull-based. Push-based exporters send data to other systems without being asked, while pull-based exporters only send data when they are requested.

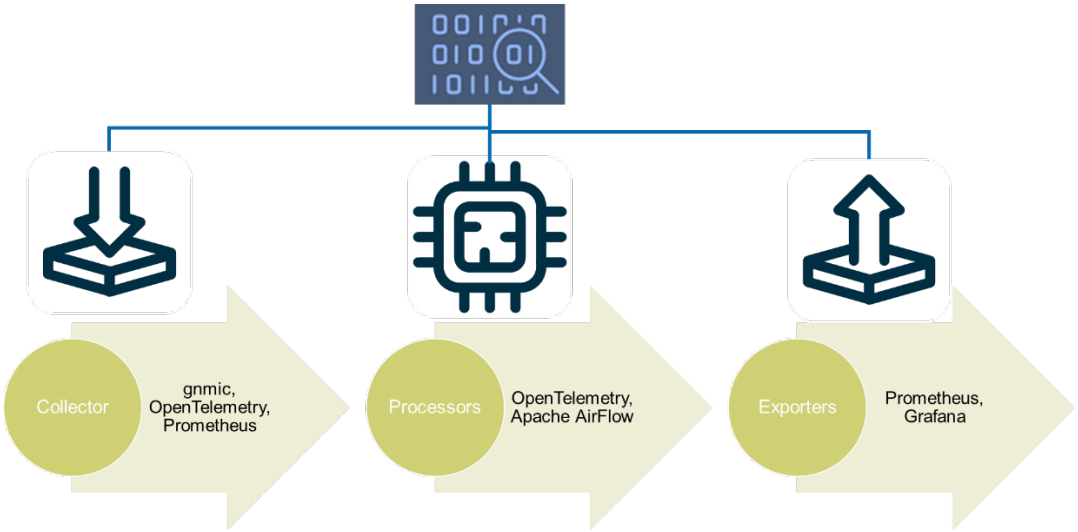


Figure 3-3: Proposed architecture for Enabler 2 - Programmable network monitoring and telemetry

The envisioned architecture for this enabler should:

- be able to scale both horizontally and vertically, thus it can be adapted to meet the specific needs of any organization.
- be based on a scalable framework that can support the collection, processing, and analysis of large amounts of data. This framework should be able to handle the high volume and velocity of data.
- consider extreme-edge devices monitoring. These devices are typically located at the edge of the network, and they generate a large amount of data. The proposed architecture should be able to collect this data and analyze it to identify potential problems.
- be able to fuse data from different sources. This data can be fused to create derived metrics that provide a more complete view of the network. For example, the proposed architecture could fuse data from network devices, applications, and cloud providers to create a metric that measures the overall health of the network.
- use ML to correlate signals and identify events that call for remediation actions. This ML-based analysis can help organizations to identify potential problems before they cause outages or security breaches.

Possible technical solutions to be integrated as part of this enabler are the following:

- A distributed architecture based on OpenTelemetry (Section 7.1.6.4) concepts for data collection can be used to collect telemetry data from a variety of sources, including applications, containers, and

cloud providers. This data can then be used to generate distributed traces, which can be used to troubleshoot problems and improve the performance of a distributed system. Distributed tracing is a technique for tracking the flow of requests through a distributed system. It can be used to identify performance bottlenecks, troubleshoot errors, and understand how different components of a system interact with each other. Analysis mechanisms can be used to correlate signals and identify events that call for remediation actions. For example, an analysis mechanism could be used to correlate logs, metrics, and traces to identify a specific error that is causing performance problems.

- Scaphandre [SCA23] will be used as part the current enabler to be applied to improve energy observability metrics throughout the whole continuum for both network and computing resources (Figure 3-4). A dedicated data base focused on energy telemetry will be implemented which will be storing the real time measurements on cloud and network sources as well as available power sources (for later use of energy sources selection mechanism in enabler#7)

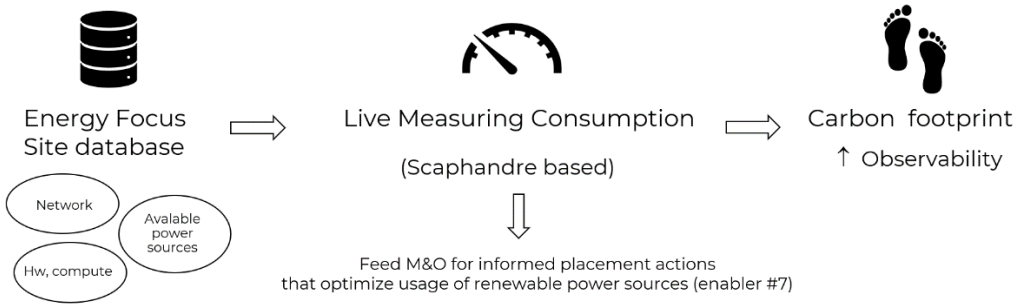


Figure 3-4: Energy Monitoring focus based on Scaphandre for environmentally sustainable target

- TeraFlowSDN (TFS) controller (Section 7.1.6.2) monitoring will be used to provide a scalable framework for data collection, processing and export, and will be extended, e.g. to support persistent KPI that can be predefined and configured (not only defined during run-time), or alarm settings to provide different type of event notifications
- 5G Monitoring Platform (as described in Section 7.1.6.3) can be enhanced to improve its integration in closed loops. This can be done by adding additional data sources related to the transport network, improving the interface for network monitoring programmability, and providing mechanisms for data distribution in hierarchical deployments and access control for data sharing. The monitoring platform can be enhanced to include additional data sources related to the transport network and to provide an improved interface for network monitoring programmability. This would allow users to create custom scripts and queries to monitor the network and identify potential problems, giving users more flexibility and control over how the network is monitored. It would also allow users to automate the monitoring process. Finally, the monitoring platform can be enhanced to provide mechanisms for data distribution in hierarchical deployments and access control for data sharing.

3.2.6 Relationship with other Enablers

The advancements in network monitoring and telemetry have relationships with various enablers and tasks within the networking domain.

In the context of **Enabler 3**, which focuses on the integration fabric, the developments in network monitoring and telemetry provide valuable input and data that can be integrated into the broader integration fabric architecture. This integration allows for seamless data flow between different components and systems, enhancing the overall monitoring and analytics capabilities of the integrated fabric.

In the context of **Enabler 6**, which focuses on the edge-cloud continuum, the developments in network monitoring and telemetry provide valuable insights into the performance and behavior of edge infrastructure. This enables efficient monitoring and management of edge resources, ensuring optimal service delivery and performance in edge-cloud deployments.

Furthermore, in the context of **Enabler 7**, which deals with the applicability and sustainability of AI, the advancements in network monitoring and telemetry enable the utilization of AI techniques for data analysis and decision-making.

Regarding **Enabler 10**, which involves the relationship with Zero-touch Service Management (ZSM), the advancements in monitoring and telemetry contribute to the analytics capabilities of ZSM. By providing rich telemetry data and metrics, the monitoring solutions enhance the analytics processes of ZSM, enabling proactive detection of issues, predictive maintenance, and efficient management of services and resources.

3.3 Enabler 3: Integration fabric

3.3.1 Motivation

Hexa-X-II derived solutions defines a **set of capabilities** that can be **customised** and **aggregated** to deliver digital services to a myriad of 6G stakeholders, including hyperscalers, application service providers, enterprise customers (e.g., verticals) and public institutions, as reported in [HEX223-D21]. These capabilities can be classified into different groups, depending on categorization criteria:

- Semantics. This criterium allows differentiating the following capability types: i) *network capabilities*, provided by network equipment and functions from access, transport, core and local data/edge network domains; ii) *cloud capabilities*, provided by virtualization enablement platforms; iii) network & *service management capabilities*, provided functions inherent to OSS (Operations Support Systems); and v) *zero-touch capabilities*, provided by AI/ML functions and data platforms. These 5 categories can be further divided depending on the technology used. For example, network capabilities can include optical, IP and microwave solutions in the transport network domain, or O-RAN and non-O-RAN capabilities in the access domain. Likewise, as for cloud capabilities, we can differentiate between those provided by IaaS platforms (e.g., Openstack), CaaS (Containers as a Service) platforms (e.g., Kubernetes), or the ones provided by novel delivery solutions, e.g., serverless computing.
- Infrastructure domains. This criterium allows clustering capabilities according to the stakeholders owning them. In the device-edge-cloud continuum, which defines the stratum for e2e service delivery, one can come up with: i) *telco capabilities*, including the ones which are within the perimeter of PLMN (Public Land Mobile Network) domain; ii) *on-prem capabilities*; including those that are instantiated within private facilities, e.g., factories, transportation hubs, etc; iii) *hyperscale capabilities*, executing on regional cloud facilities.
- Operation domains. Similar to the previous criterium, but with focus on who operates the capabilities. It is worth noting that the stakeholder owning a capability is not always responsible for operating it. For example, in a private network infrastructure, the infrastructure owner can outsource the operation to a Managed Service Provider (MSP), typically the business unit of an operator. Based on this criterium, capabilities can be categorized as: i) *operator managed* capabilities and ii) *3rd party managed* capabilities.

In this complex multi-stakeholder environment, with several capabilities of different nature and owned/operated by different stakeholders, the actual aggregation and customization activities needs to be done using programmable compositional patterns, following:

- cloud-native practices, on the context of serverless architectures.
- plug-and-play approaches

This is where the integration fabric turns in.

3.3.2 Objectives

The integration fabric is necessary for two main reasons. On the one hand, a **profound re-factoring of management services**, making them modular and stateless, so that they can be deployed as microservices which can be scaled and containerized independently. On the other hand, a complete **system APIfication**, replacing traditional point-to-point interfaces with HTTP-based RESTful APIs made available for consumption. To policy and manage the interaction between the producers and consumer of the different microservices through these APIs, an integration fabric is needed.

The main objective to overcome the presented limitations is the definition of service bus to allow liquid and frictionless interoperation between Hexa-X-II capabilities. The purpose of an integration fabric is to facilitate the integration of management services from different management domains.

3.3.3 Solution description

The integration fabric is an advanced system component that enables liquid interoperation between API producers and consumers, regardless their administrative domains nor the actual capability in scope. The integration fabric implements a set of features, including:

- Connectivity. This includes i) *service communication*, using a set of communication channels; ii) *service discovery*, enabling the consumer to discover the available microservices, their endpoints and supported capabilities, iii) *service registration*, which handles addition/removal of services to/from the set that can be discovered; and iv) *service access control*, implementing needed AuthN/Z solutions for the invokers to become authorized service consumers. In what relates to service communication, it is worth noting that the content of the communication channels includes event notifications, streams and data objects that are generated by the service producer asynchronously, and that can be reported using either query or subscribe-notification mechanisms. When creating new channels, the producer can specify channel properties such as QoS, content type, restriction, policies (e.g., filtering criteria) and available communication styles (pull vs push). Consumers can subscribe to one or more channels, according to their interests.
- Reliability. This is achieved by means of different resilience patterns, including: i) circuit breaking, designed to remove endpoints that persistently return error messages from a load-balanced group; and ii) timeouts; iii) retry and iv) fallback.
- Security. In the security-by-default, defence in-depth and zero-trust network environments claimed for 6G, this requires implementing: (i) mutual TLS (mTLS) (variation of TLS in which two sides of a communications channel verify each other's identity, instead of only one side verifying the other, usually used in zero trust framework), using cryptographically secure techniques to mutually authenticate individual microservices and encrypt the traffic between them; ii) sidecar and perimeter proxies, working as policy enforcement points, and iii) identity and certification management solutions.
- Observability. This includes i) the ability to monitor telemetry and metrics, including latency, traffic, error and situation; ii) distributed tracing, intended to be used when metrics do not provide enough information to troubleshoot a problem or understand an unexpected behaviour; and iii) topology graph visualization, to have the full picture on traffic flows and microservice status.

The integration fabric can be typically implemented using service bus, either as a service mesh (see Section 3.3.5.1) or a message broker (see Section 3.3.5.2).

3.3.4 SoTA and Beyond SoTA

3.3.4.1 Standardization landscape

As for **standardization bodies**, it is worth noting the work done at ETSI ZSM, with the definition of a modular, scalable and extensible system architecture [zsm-002] aiming to help operators in their automation transformation. One of the key elements of the ZSM architecture framework is precisely the integration fabric. This is a management function that allows composing and communicating capabilities between them, yet through a controllable abstraction layer.

As for open-source reference solutions, **Openslice** [OPE23] deserves attention. **Openslice** is a prototype open-source OSS that provide operator with automation means for the delivery and assurance of slicing services. Figure 3-5 pictures the different components building up the Openslice solution suite (for more details on them, see [OTR20]). This reference solution is aligned with ETSI ZSM architecture, as proved in [zsm-poc2]. As noted, the service bus maps well to ZSM integration fabric.

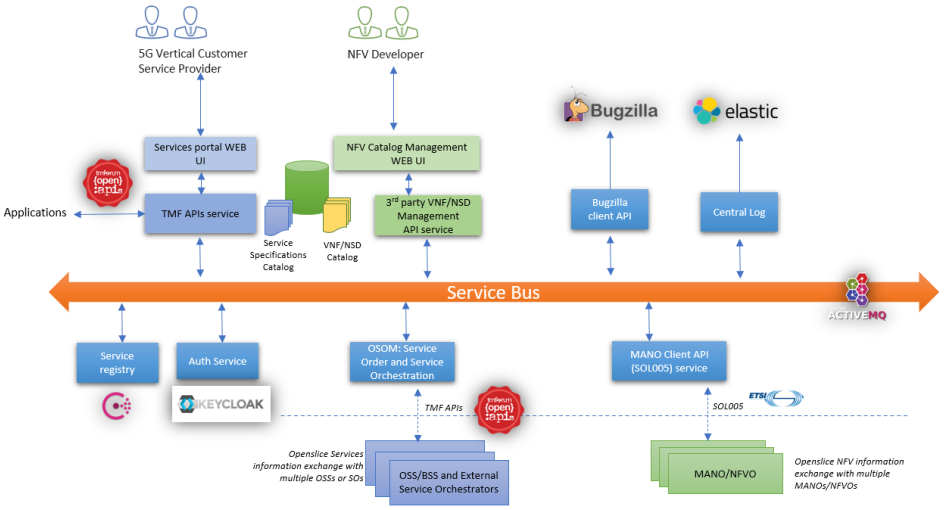


Figure 3-5: Openslice reference architecture. Source: [OPE23]

With regards to the **PoCs**, it is worth noting ZSM PoC#6: "Security SLA assurance in network slices". This PoC exhibits a security-oriented closed loop as well as the implementation of a High-Level Architecture (HLA) based on the ZSM reference architecture, across many domains and sites coupled by an Integration Fabric. The PoC architecture is built on top of the INSPIRE-5Gplus project's [INS20] general specification of the HLA. This HLA complies with the design guidelines, management services, and domains of the ETSI ZSM standard architecture.

3.3.4.2 Service mesh solutions

Service mesh is a software architecture that enables developers to connect, control, monitor, and secure microservices architectures. Main commercial solutions work with any microservice regardless of its platform, source, or vendor, providing a unified layer between application services and the network. In general, service mesh solutions provide observability, robust communication, and control even as the number of microservices in an application increase. It supports workloads running in both containers and virtual machines (most important solutions are mainly built on top of k8s) and can manage the interactions between them. As a general structure, service mesh is divided into a data plane and a control plane:

- The data plane consists of a set of proxies (Envoy), deployed as sidecars alongside the application.
- The control plane manages and configures proxies to route traffic.

Regarding the control plane and sidecar proxy one of the most used is Envoy. **Envoy Proxy** [ENV23] is an open-source, high performance, small footprint edge and service proxy. It is intended to be run as a sidecar adjacent to each service in an application, distancing the network from the essential operational logic. Load balancing, resilience features like timeouts, circuit breakers, and retries, as well as observability and metrics, are all features offered by Envoy. Envoy can be used as an API gateway for networks, critical feature in delivering service discovery utilities. Additionally, Envoy gathers detailed metrics about the traffic it traverses and makes them available for viewing and use in tools like Grafana. The landscape of service mesh solutions is very wide, in this document is addressed a comparative, that can be seen in Table 3-3, between three of them, **Istio** [IST23], **Consul** [HAS23] and Linkerd [LIN23].

	Istio	**Consul**	**Linkerd**
Infrastructure comparison			
Platforms	K8s, VMs	K8s, VMs	K8s
Sidecar proxy	Envoy	Envoy (others available)	Linkerd-proxy
Per-node agent required	No	Yes	No

Traffic management comparison			
Circuit breaking	Yes	Yes (Envoy)	No
Fault injection	Yes	Yes	Yes
Rate Limitation	Yes	Yes (Envoy)	Yes
Observability comparison			
Prometheus compatibility	Yes	Yes	Yes
Integrated Grafana	Yes	No, but compatible	Yes
Distributed tracing	Yes (OpenTelemetry)	Yes	Yes (OpenTelemetry)
GUI	No, but Kiali available	Yes	No, only Grafana dashboard
Deployment and other features comparison			
Multi cluster	Yes	Yes	Yes
Mesh expansion	Yes	Yes	No
Deployment	CLI, Helm, IstioCtl	Helm	CLI, Helm
Complexity	High	Medium	Low

Table 3-3: Service mesh opensource comparison

3.3.4.3 Message broker solutions

RabbitMQ [RAB23] is a distributed message broker. It uses mostly Advanced Messaging Queuing Protocol (AMQP) for secure transfer of messages. The communication is based on producers that publish messages to an exchange which routes the messages to different queues. Based on the message routing key, the exchange creates a binding with the queue and routes the messages to the queue. The message stays in the queue until a consumer connects to the queue and subscribes to the message. There are different exchange types, i.e. direct (delivery based on a message routing key), topic (delivery based on a wildcard match between the routing key and the topic), fanout (delivery to all of the queues tied to the exchange) and headers (delivery based on header attributes).

NATS [NAT23] is an open-source messaging platform built to meet the distributed computing needs of modern applications. At its core, NATS enables applications to exchange data in the form of messages. These messages are addressed by subjects and do not depend on the underlying network. All the different messaging models (subject-based, publish-subscribe, request-reply and queue groups) are built on top of the fundamental publish and subscribe messaging model, which is asynchronous. NATS has a built-in distributed persistence system called **Jetstream** which enables new functionalities and higher qualities of service on top of the base 'Core NATS' functionalities and qualities of service.

Apache Pulsar [APA23] is a cloud-native, distributed messaging and event-streaming platform. It has a native support for multiple clusters, with geo-replication of messages across clusters. It supports common message patterns with its diverse subscription types and modes (exclusive, failover, shared…), with a very low publish and end-to-end latency and a scalability to over a million topics. It guarantees message delivery with persistent message storage provided by Apache BookKeeper. This solution includes native, lightweight compute capabilities known as Pulsar Functions that allow you to build microservices following serverless computing framework to processes messages. These functions can consume messages from one or more topics, applies a user-defined processing logic to the messages, publishes the outputs of the messages to other topics. Pulsar has a cloud-native, layered architecture that separates compute and storage into different layers. Decoupled compute and storage allow for independent scaling and enables microservices to scale elastically on short notice.

	RabbitMQ	NATS	Pulsar
Use cases			
Message routing	Yes	Yes	Yes
Pub/Sub	Yes	Yes	Yes
Queue	Yes	Yes (Core)	Yes
Event Streaming	No	Yes (Jetstream)	Yes
Key features			
Language supported	50 client types	Core NATS: 48 clients. NATS Streaming: 7 clients	7 client languages, 5 third-party clients
Msg replay	Yes	Yes	Yes
Msg retention (time, ack, sub-based)	Yes	Yes	Yes
Built-in storage	No	No	Yes
Processing capabilities	No	No	Yes (Pulsar Functions)
Delivery guarantees	At most once, At least one	At most once, At least one, Exactly once (Jetstream)	At most once, At least one, Exactly once
Built-in georeplication	No	Yes	Yes
E2e encryption	No	No	Yes
Built-in multitenancy	No, only vhosts	Yes	Yes
Observability comparison			
Prometheus + Grafana compatibility	Yes	Yes	Yes
GUI	Yes	No, only nats-top	Yes
Deployment, performance and scalability			
Horizontally scalable	No	Yes	Yes
High availability	One-node failure	One-node failure	Many-node failure
Complexity	Low	Low	Medium/High

Table 3-4: Message broker opensource comparison

3.3.4.4 Beyond SoTA

The ambition for Enabler #3 is to provide Hexa-X-II an interconnection bus that allow liquid and frictionless interoperation between Hexa-X-II derived solutions capabilities that fall in the scope of this enabler by:

- Defining a multi-domain solution that is powered by cloud-native frameworks. The idea is to extend the concept of microservice connectivity, fully affirmed in today cloud solution to a broader perspective wherein management domains belong to different administrative domains (multi-stakeholder).

- Deploying the solution considering scenarios wherein one or more management domain operates on the extreme-edge (resource volatility).
- Bringing enhancements over service bus built-in features, for it to be used in the above scenarios. These features include **connectivity, reliability, security and observability**.

In Figure 3-6, it is shown the reference architecture aimed to achieve the objectives above mentioned. They are shown the domains that are involved, from the side of capability providers and digital service provider in addition to the positioning of the integration fabric in the overall structure.

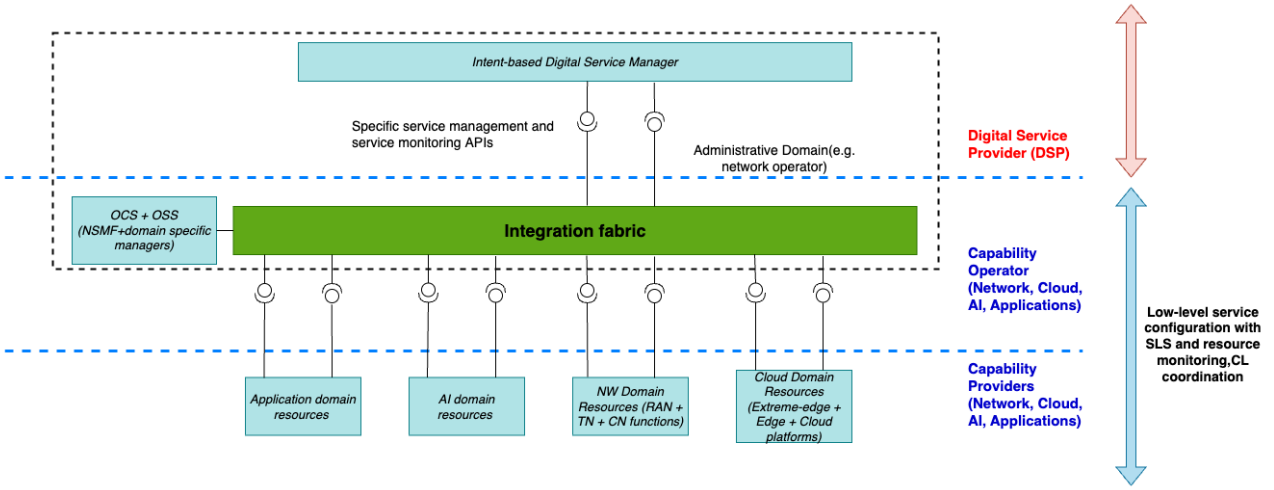


Figure 3-6: Enabler 3 Reference architecture

3.3.5 Identification of possible components and interfaces

In the development of this enabler two different architectural options are being considered: service mesh and message broker. The two designs are both valid but each of them leverages a different approach to achieve the objective. The service mesh design, manages the communication between services in a synchronous way, following a structural pattern that is managed by an orchestrator. It focuses on reliability and observability in a networked environment. The message broker approach results more lightweight, with respect to the previous one, based on an asynchronous paradigm. It creates a single communication bus to exchange information between services, consuming less computational power, and with a lower maintenance cost. In this section both proposals are analyzed.

3.3.5.1 Service mesh architecture

Figure 3-7 shows the proposed service mesh architecture. It includes the following components, which are described in Table 3-7: Services, Service registry, Security module, Monitoring module, Traffic Manager, Admin endpoint, and Communication interface.

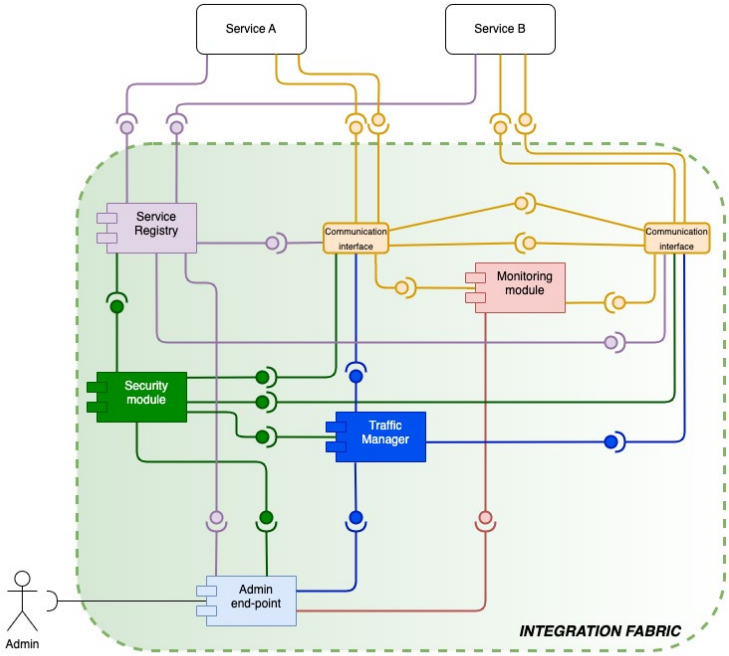


Figure 3-7:Integration fabric service mesh-based architecture.

Component	Description	Possible technologies
Services	Network functions, management systems, and external applications, which need to communicate with each other. Each of them is isolated from each other's, unaware of the topology. They can communicate only interfacing with integration fabric. To have a standardized way to connect and deploy them, each of these services is put in containers.	Container virtualization (ex. Docker)
Service registry	It has two main functions: it is the insert/delete endpoint, and it is the topology and inter-service relation manager. It is the single insertion and removal endpoint for new services. This single-entry point makes possible to better regulate the topology updates and address the horizontal scalability also in a multi-tenancy scenario. Inserting a new service means equipping it with a virtualization support (sidecar) and plugging it in the existing topology, also defining its position in term of routing and visibility w.r.t. other services. One service that wants to take advantage of integration fabric services must start a one-time onboarding procedure, that may be built as a volatile process, exploiting virtualization technologies like serverless computing. The second task shipped out by this module embraces all the operations to maintain the list of services handled by the integration fabric updated, showing the amount and types of services, and the current topology. This makes it possible to enable service discovery features.	Serverless (onboarding) Eureka server, control plane of SM solutions (discovery, registry)
Security module	Secure the access to the resources managed by the integration fabric, as well as the communication within the modules. In details, it ensures AuthN/Z mechanisms to guarantee that the resources are accessed only by accredited people. Makes possible to manage token/key to unlock different QoS or certain services. Last feature is to manage the encryption within modules (TLS, mTLS) in order to avoid leak of information in the inter-service communication.	Identity Provider, IdP (ex. Keycloak)

Monitoring module	The monitoring module oversees recording and following the flow of requests as they pass through various services. Analysis, monitoring, and troubleshooting of the system is made possible by tracing. It also offers metrics to get insight on the performance of the system. It is connected directly to all the communication interface acting as a capture filter for the request and flow of information passing through the integration fabric.	Prometheus, Zipkin, Jaeger, OpenTelemetry
Traffic Manager	It is connected to all the communication interface of the services, to make possible changes in the networking setup within the system and the resource exposed by each service. This module must allow fine-grained policies at the network layer, unlocking advanced traffic management capabilities like load balancing, routing, and traffic shaping.	Control plane of SM solutions
Admin endpoint	Observability is an important point of the integration fabric, to make the constant check of the health and key metrics of the system possible. An admin endpoint is also important to have a clear view and access at high level on the general topology, communication, and routing rules between services. This is possible because that block is directly connected to the management's blocks of the architecture, i.e., the traffic manager and service registry. The admin endpoint is secured through the security module, to ensure only authorized subjects can access this integration fabric control panel.	UI offered by commercial service meshes (ex. Kiali)
Communication interface	It is the communication core of the architecture. The set of interfaces represents the interconnection network between services that make possible to accomplish all the features designed as goals of the integration fabric. Each service, in fact, will be represented by a communication interface in the architecture, that is its "spokesperson" to interact with integration fabric key modules and the other services.	Sidecar proxy (ex. Envoy)

Table 3-5: Enabler 3 Integration Fabric Service mesh possible technologies

Figure 3-8 shows the proposed interfaces for each Service mesh component. These are detailed in Table 3-6.

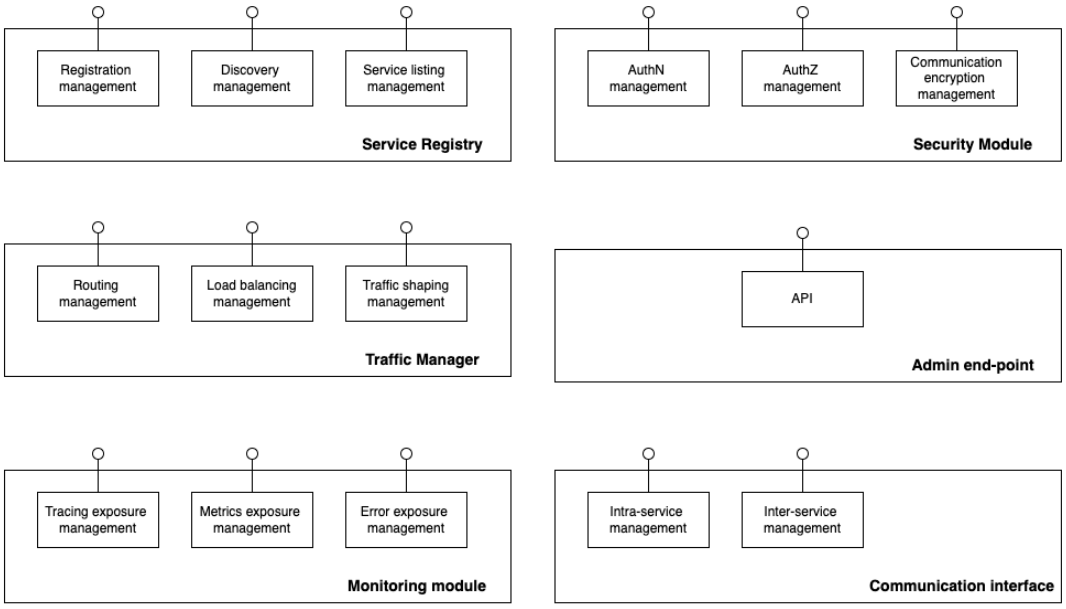


Figure 3-8: Integration fabric service mesh-based architecture interfaces

Component	Interfaces exposed	Parameters needed
Service registry	*Registration management*: exposes resources to add/remove services.	• Unique identification address (IP/FQDN). • Characterization (Name, type of service, resources exposed). • AuthN/Z parameter (auth/key/token).
	Discovery management: exposes resources to locate a certain service in the architecture.	• Identifier of service to point (ID/Name). • AuthN/Z parameter (key/token).
	Service listing management: exposes the list of all the services, or that pertain to a certain scope specified in additional parameters.	• AuthN/Z parameter (key/token). • Additional parameters related to the request characterization.
Security module	*Authentication (AuthN)*	• Unique identification address (IP/FQDN). • AuthN parameters submitted by the requesting party.
	Authorization (AuthZ)	• Unique identification address (IP/FQDN). • Resource identifier to check the access privilege of requesting party.
	Communication encryption	• Unique identification address (IP/FQDN)
Monitoring module	*Tracing:* exposes resource to check health and the status of communication of a specific service, or a group of services	• List of unique identification address (IP/FQDN) involved in the request. • AuthZ parameters submitted by the requesting party.
	Metrics: exposes metrics to get insight on the performance of the system	• List of unique identification address (IP/FQDN) involved in the request. • AuthZ parameters submitted by the requesting party.
	Error exposure: exposes a unique interface to have a global source of all errors detected in the system	• AuthZ parameters submitted by the requesting party.
Traffic Manager	*Routing:* expose resources to direct traffic between services, create/remove communication channel between services.	• Unique identification address (IP/FQDN) of requesting party. • AuthZ parameters submitted by the requesting party. • Destination service identifier
	Load balancing: expose resources in order to distribute traffic among service instances (round-robin, least connections, weighted, consistent hashing).	• List of unique identification address (IP/FQDN) involved in the request. • AuthN parameters submitted by the requesting party, to check if it is a user with the privilege to perform certain actions.
	Traffic shaping: expose resources to control and manage the flow of network traffic	• List of unique identification address (IP/FQDN) involved in the request.

	between services. It involves setting rules and policies to shape the traffic patterns according to specific requirements (includes rate limiting).	• AuthN parameters submitted by the requesting party, to check if it is a user with the privilege to perform certain actions. • Specific technical parameters that characterize the request.
Admin endpoint	*API:* exposes all the resources (or only a part) of the system to authenticated users	• AuthN parameters submitted by the requesting party, to check if it is a user with the privilege to perform certain actions. • Specific technical parameters that characterize the request.
Communication interface	*Intra service communication:* expose resources to make possible the direct communication of the service, to its interface in the integration fabric system.	• No parameters needed because it is a container-sidecar interface.
	Inter service communication: expose resources to make possible the communication with the other services communication interfaces as well as others integration fabric components	• AuthZ parameters submitted by the requesting party. • Parameters of the query to access resource/ perform actions.

Table 3-6: Enabler 3 Integration Fabric Service mesh components and interfaces

3.3.5.2 Message broker architecture

Figure 3-9 shows the proposed message broker architecture, including its components that are detailed in Table 3-7: Services, Service registry, Security module, Connectivity manager, Tracing module, Admin endpoint, and Message broker.

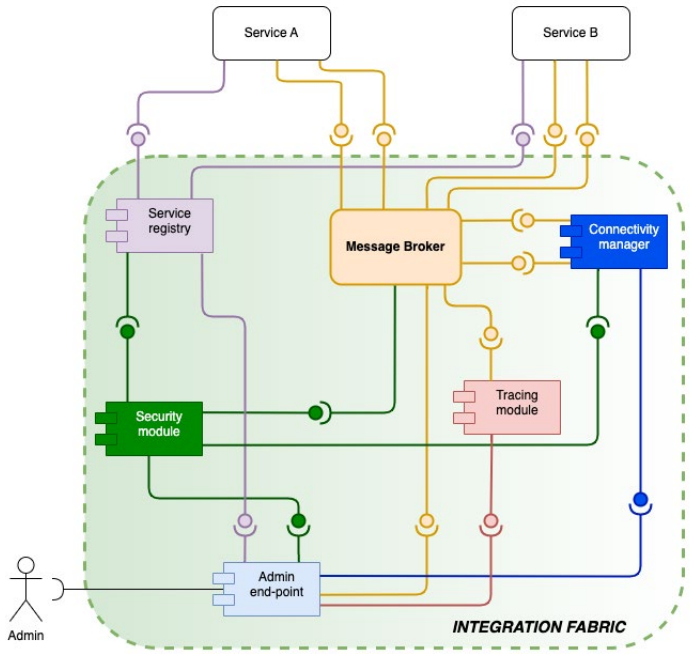


Figure 3-9: Integration fabric message broker-based architecture.

Component	Description	Possible technologies
Services	Network functions, management systems, and external applications, which need to communicate with each other. Each of them is isolated from each other's, unaware of the topology. They can communicate only interfacing with integration fabric. To have a standardized way to connect and deploy them, each of these services is put in containers. Each of them alongside the container scope is equipped with a client that makes it possible to communicate with the message broker.	Container virtualization (ex. Docker), chosen solution client
Service registry	It has two main functions: it is the create/delete services subscription and manage inter-service relation. It is the insertion and removal endpoint for new services. This single-entry point makes possible to better regulate the inter-service relations updates and address the horizontal scalability also in a multi-tenancy scenario. Insert a new service means subscribe it to the base functionalities' topic (alarm, coordination…). After this registration step the service will not be aware of the current structure around it, but it will be aware of the relevant events happening in the architecture. One service that want to take advantage of integration fabric services must start a one-time onboarding procedure, that may be built as a volatile process, exploiting virtualization technologies like serverless computing. The second task shipped out by this module embraces all the operations to keep updated the list services managed by the integration fabric. Shows the amount, the type services managed. Keep up to date the current policies used, relations within the system in terms of topic and subscriptions in the scope of the integration fabric. This makes possible to enables service discovery features, crucial in scope of the integration fabric.	Serverless (onboarding) Eureka server, (discovery, registry)
Security module	Secure the access to the resources managed by the integration fabric, as well as the communication within the modules. In details, need it ensures AuthN/Z mechanisms to guarantee that the resources are accessed only by accredited people. Makes possible to manage token/key to unlock different QoS or certain services. Last feature is to manage the encryption within modules (TLS, mTLS) in order to avoid leak of information in the inter-service communication.	Identity Provider, IdP (ex. Keycloak)
Connectivity manager	It offers advanced management in terms of subscription, retention policies and all the aspects related to the management of topic and related queue. Connectivity manager intelligently routes and distributes traffic between services based on defined policies and subscriptions. This module is connected, directly, to the message broker to modify the subscription so the engagement of a specific service and bringing to a modification of communication patterns of the system.	Custom code using chosen solution API.
Tracing module	The tracing module oversees recording and following the flow of requests as they pass through various services. Analysis, monitoring, and troubleshooting of the system is made possible by tracing	Zipkin, Jaeger,OpenTelemetry
Admin endpoint	The observability is an important point of the integration fabric, to make possible the constant check of the health and key metrics of the system. This is possible, because, this endpoint, embed a tracing and metric generator module that are connected to the topic of interest. An admin endpoint is also important to have a clear view and control at high level on the general services relations, and routing rules between services. This is possible because that block is directly connected to the management's blocks of the architecture, i.e., the traffic manager and service registry The admin endpoint is secured through the security module, to ensure only authorized subjects can access this integration fabric control panel.	UI offered by message broker

Message broker	It is the core of the architecture. It represents the central node in which all the packet is routed according to the subscription, the available topics, and the retention policies. Each service will be asynchronously connected to this module. From the broker each of them will receive all the information of the other components and will react to certain event or request by directly publishing on the broker itself (event-driven architecture).	

Table 3-7: Enabler 3 Integration Fabric Message broker possible technologies

Figure 3-10 shows the proposed interfaces for each Service mesh component. These are detailed in Table 3-8.

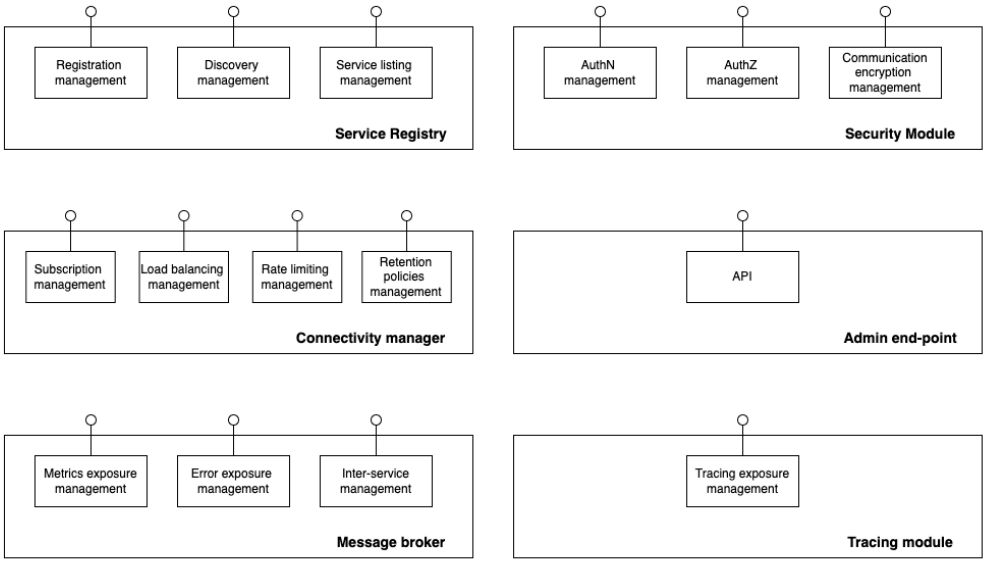


Figure 3-10: Integration fabric message broker-based architecture interfaces

Component	Interfaces exposed	Parameters needed
Service registry	*Registration management*: exposes resources to add/remove services.	• Unique identification address (IP/FQDN). • Characterization (Name, type of service, resources exposed). • AuthN/Z parameter (auth/key/token).
	Discovery management: exposes resources to locate a certain service in the architecture.	• Identifier of service to point (ID/Name). • AuthN/Z parameter (key/token).
	Service listing management: exposes the list of all the services, or that pertain to a certain scope specified in additional parameters.	• AuthN/Z parameter (key/token). • Additional parameters related to the request characterization.
Security module	*Authentication (AuthN)*	• Unique identification address (IP/FQDN). • AuthN parameters submitted by the requesting party.
	Authorization (AuthZ)	• Unique identification address (IP/FQDN). • Resource identifier to check the access privilege of requesting party.
	Communication encryption	• Unique identification address (IP/FQDN)
Connectivity Manager	*Subscription management:*	• Unique identification address (IP/FQDN) of requesting party.

	expose resources to direct traffic between services, create/remove subscription and topics	• AuthZ parameters submitted by the requesting party. • Existing/ new topic information
	Rate limiting: expose resource to manage the flux of data injected by a service to a specific topic or group of topics.	• AuthN parameters submitted by the requesting party, to check if it is a user with the privilege to perform certain actions. • Specific technical parameters that characterize the request. • Topic(s) to point
	Load balancing: expose resource to manage the topic usage in terms of message publishing, to split or balance traffic.	• List of unique identification address (IP/FQDN) involved in the request. • AuthN parameters submitted by the requesting party, to check if it is a user with the privilege to perform certain actions. • Specific technical parameters that characterize the request. • Topic(s) to point
	Retention policies: expose resources to manage the policies that manage the duration and type of management that is used for the packets of a certain topic of group of topics.	• AuthN parameters submitted by the requesting party, to check if it is a user with the privilege to perform certain actions. • Specific technical parameters that characterize the request. • Topic(s) to point.
Tracing module	*Tracing:* exposes resource to check health and the status of communication of a specific service, or a group of services	• List of unique identification address (IP/FQDN) involved in the request. • AuthZ parameters submitted by the requesting party.
Admin endpoint	*API:* exposes all the resources (or only a part) of the system to authenticated users	• AuthN parameters submitted by the requesting party, to check if it is a user with the privilege to perform certain actions. • Specific technical parameters that characterize the request.
Message brokers	*Metrics:* exposes metrics to get insight on the performance of the system	• List of unique identification address (IP/FQDN) involved in the request. • AuthZ parameters submitted by the requesting party.
	Error exposure: exposes a unique interface to have a global source of all errors detected in the system	• AuthZ parameters submitted by the requesting party.
	Inter service communication: expose resources to make possible the communication with the other services communication interfaces as well as others integration fabric components	• AuthZ parameters submitted by the requesting party. • Parameters of the query to access resource/ perform actions.

Table 3-8: Enabler 3 Integration Fabric Message Broker components and interfaces

3.3.6 Relationship with other Enablers

The integration fabric, being thought of as a layer of connections between the various capabilities offered by the solutions that are within the scope of Hexa-X-II, is connected to different enablers. More specifically, the

enablers 1, 2, 4, 5, 10 and 11. Regarding the **Enabler 1** (Programmable flexible network configuration), the integration fabric will serve as a mean to expose the capabilities and northbound interfaces, of the above-mentioned enabler towards OSS and BSS systems. The relationship with the **Enabler 2** (Programmable network monitoring and telemetry) is related to the fact that this layer of interconnection will be "developed in such a way as to provide traceability", therefore closely connected to the requirements of the Enabler 2.

The second link concerns the **Enabler 4** (Trustworthy 3rd party management). This enabler must be developed in such a way as to ensure the performance, management and security requirements of the Enabler 4.

An additional relation can be found with **Enabler 5** (Multi-cloud management mechanisms) because the integration fabric will serve as a communication bus for deploying and managing applications across multiple cloud environments.

As mentioned above, there is also a correlation with the **Enabler 10** (Zero-touch closed loop governance) and the **Enabler 11** (Zero-touch multiple closed loop coordination). Enabler 10 defines the CL that is related to the concepts of resource allocation, network configuration and/or service tuning requires a communication interface to connect the defined modules on different layers. Connected to this principle, is the link to the enabler 11, which promotes the coexistence of multiple CLs, that therefore need an interconnection bus.

Finally, the connection with the E2E enablers that are being defined in [HEX223-D21] should also be highlighted. The integration fabric is the only point of exposure of the functionalities developed in the e2e service layer defined in this task.

3.4 Enabler 4: Trustworthy 3rd party management

3.4.1 Motivation

Hexa-X-II aims to become a programmable service platform that can be easily accessed by 3rd parties, including enterprise/vertical customers (B2B segment) and application providers/service providers (B2B2C segment)[1]. This ambition allows releasing these 3rd parties from the constraints of traditional over-the-top, best-effort service delivery approaches, tapping into new capabilities to provide enhanced user experience and contribute to enrich digital ecosystem with new services (e.g., metaverse, extended reality, Web3, etc.).

From the standpoint of Hexa-X-II, this requires a controllable and auditable exposure of capabilities to individual 3rd parties, according to their “tenant” profiles. This profile provides a full characterization of a 3rd party, including information on i) market segment; ii) trust level; iii) contracted services and their SLAs; and iv) subscribers, which can be either MNO’s end-users or enterprise users, or any combination in between. **Controllable exposure** means that the Hexa-X-II can regulate the particular set of resources each 3rd party system is allowed to access and under which conditions. **Auditable exposure** means that every interaction between Hexa-X-II and 3rd party systems needs to be logged with accurate timestamps (for traceability) and support non-repudiation (for SLA verification).

To fulfill these requirements, the NaaS work that telco industry is conducting through GSMA Open Gateway can constitute a good starting point.

3.4.2 Objectives

The objective of this enabler is to define solutions for multi-tenancy support. In resource sharing environments with multiple tenants running atop, these solutions shall provide Hexa-X-II with the ability to ensure isolation in terms of:

[1] These 3rd parties can gain access to Hexa-X-II system directly or through channel partners (e.g., marketplaces/aggregators). These discussions are in scope of WP2.

- **Performance**, ensuring that SLA can be met on the service instances allocated for each tenant, regardless of workloads or faults from other running instances.
- **Management**, ensuring that each tenant can operate their allocated service instances independently, according to their dynamicity (e.g., time-varying traffic loads) and flexibility (e.g., re-configuration) needs. This operation scopes i) service lifecycle management, including activation/deactivation and scaling; and ii) subscriber management, with the provision of tailored service experiences.
- **Security,** ensuring that i) any type of intentional attack occurring in one service instance have no impact on any other running instance, and ii) the SLA data associated to tenant services, e.g. performance, fault, trace data, is safely stored and accessible only for that tenant.

3.4.3 Description of the solution

The solutions for multi-tenancy support will be developed in three separate tracks: resource controllability separation, user-centric network management, and SLA enforcement.

The **resource controllability separation track** focuses on the design and development of mechanisms to provide tenants with segregated yet customized management spaces. The management space defines what the tenant is authorized to do with regards to i) the operation of their allocated services, and ii) the control of their applications, including their interaction with Hexa-X-II resources, allowing for a frictionless network-application integration. To make sure this works in resource sharing environments, it is needed to ensure these management spaces are provisioned with permissions that do not conflict with each other. In this regard, the mechanisms in scope of this track will be built upon granular access control solutions.

The **user-centric network management track** aims at specifying solutions that allow Hexa-X-II system to manage tenant subscriber policies in relation to service consumption. In particular, these solutions shall cover two aspects:

- Offering subscribers with optimal and personalized experience, according to user preferences, SLAs of subscribed services, and network context (e.g., congestion scenarios).
- Assisting **subscribers** to gain access to hosted applications, while keeping their data safely stored, preventing any authorized entity to read/modify them. This needs to be compliant with EU regulation for data privacy and protection (e.g., GPDR).

The **Trust Level Agreements and SLA enforcement track** will cover activities related to SLA translation, assurance and (external) verifiability, with focus on security and privacy. The parameters forming the SLA structure may include Key Performance Indicators (KPIs), Key Value Indicators (KVIs) and Trust Level Agreements (TLAs). The KPI & KVI values will be translated into appropriate control, configuration and monitoring actions, while TLAs will impact on the security framework which is under project-level discussion.

3.4.4 SoTA and Beyond SoTA

This section reports on the SoTA and beyond for each of the tracks conforming the enabler.

3.4.4.1 Resource controllability separation track: SoTA

Reference	Name	Scope of work
3GPP TR 28.804	Study on tenant concept in 5G networks and network slicing management	This Rel-16 technical report explores how 3GPP management system can provide management capabilities for fulfilling requirements from tenancy use cases, whereby there is a need to provide management services and resource for each tenant.
3GPP TR 28.817	Study on access control for management services	This Rel-17 technical report investigates use cases related to access control, proposes requirements on 3GPP management system and possible solutions to fulfil them. These use cases include, among others, • Authorization and authentication from 3rd party consumers • Trust relationship with 3rd party consumers • Integration of access control with existing operator's AAA system.

3GPP TR 28.824	Study on network slice management capability exposure	This Rel-18 technical report describes use cases, potential requirements and solutions for exposure of management services to 3rd party service providers (e.g., verticals) acting as network slice customers. It also provides recommendations for further normative work, which touches on the need to integrate with CAPIF framework [23.222] and align with GSMA Open Gateway initiative.
3GPP TS 28.541	Management and Orchestration; 5G Network Resource Model (NRM)	This technical specification defines and maintains the information model of 3GPP 5G system. This model allows capturing all the management aspects (configurable and readable attributes, and relationships across them) of 3GPP 5G system resources, including NR functions, 5GC functions and network slices.
ETSI GR ZSM 010	General security aspects	This technical report does overall security threat and risk analysis for ZSM framework, lists key issues/risks of the framework based on use cases, proposes solutions to mitigate the risks, and raises potential requirements on ZSM framework to support the security capabilities. These capabilities include, among others: • Trust relationship between management domains • Security assurance • Multi-tenancy support • Access control
ETSI GS ZSM 014	Security aspects	This technical specification takes recommendation for normative work based on ZSM010 and elaborates on these solutions. The ongoing work cover the definition of management services with regards to authentication administration, authentication enforcement, authorization administration, authorization decision, security log collection and auditability.

Table 3-9: Resource controllability separation track – literature review

Table 3-9 represents the work that has been done in the standardization landscape. The focus has been on developing solutions ensuring isolation in terms of management and security.

In the 3GPP domain, TR 28.804 [28804] and TR 28.824 [28824] raised attention on the need to have multi-tenancy support in 3GPP management, especially with the focus on network slicing. However, neither of these studies concluded takeaways that justified going for a normative phase; of the main reasons sustaining this decision was the 3GPP community did not reach consensus on the scope and meaning of tenant concept. TR 28.817 [28817], approached from the standpoint of access control, identified three gaps for normative phase: i) the need to update Service Based Management Architecture defined in TS 28.533 [28533]) to support authentication, authorization and audit capabilities; ii) the need to update the 5G NRM defined in TS 28.541 [28541] to support authentication, authorization and audit capabilities; and iii) the need to support generic management services (defined in TS 28.532 [28532]) to support authentication and authorization capabilities. These gaps are now framed within Rel-18 onwards.

In the ZSM domain, ZSM 010 [zsm-010] explored key issues on security management, and provided recommendations for work in normative phase, which is as of today carried out in ZSM 014 (draft in progress). Unlike 3GPP, ZSM poses more advanced requirements (in terms of functionality) at the cost of having less detailed solutions (generally not tied to any specific technology or protocol).

3.4.4.2 Resource controllability separation track: beyond SoTA

The ambition is to provide Hexa-X-II with a native multi-tenancy support, by:

- defining the scope and impact of tenancy concept in the management and orchestration system, shedding light on the controversial issues that prevented consensus within 3GPP community. The aim is to i) associate tenant concept to 3rd party consuming Hexa-X-II capabilities, and ii) ensuring each tenant are provided with tailored and segregated management spaces.

- specifying granular access control solution as to authentication, authorization and auditability, leveraging the recommendations for normative work issued by TR 28.817 [28.817]. These solutions shall ensure that management spaces are provided with permissions that do not conflict with each other on resource sharing environments.

3.4.4.3 User-centric network management track: SoTA

To provide user-centric service experience, 3GPP has defined User Equipment Route Selection Policy (URSP). URSP is a 5G feature that enables mapping certain user data traffic (i.e., application traffic) to Protocol Data Unit (PDU) session connectivity parameters (e.g., S-NSSAI, DNN, Session and Service Continuity – SSC - mode), to make sure traffic flows from device applications are treated end-to-end as required, matching service expectations. The URSP is a signaling construction composed of one or more URSP rules, each consisting of i) one Rule Precedence, ii) one Traffic Descriptor, and one more Route Selection Descriptors.

- The rule precedence determines the order in which an URSP rule is enforced at the device.
- The Traffic Descriptor (TD) is used to determine when the rule is applicable. An URSP rule is determined to be applicable when every component in the Traffic Descriptor matches the corresponding information from the application.
- The Route Selection Descriptor (RSD) specifies PDU session connectivity parameters. They profile the behavior that will experience the user data traffic of matching application.

Figure 3-11 provides a illustrative conceptualization of the URSP rule. For a complete description, please refer to 3GPP TS 23.501 [23.501] and 3GPP TS 24.526 [24.526]. This figure shows how URSP rules matches client application traffic flows to connectivity requirements, with the use of Traffic Descriptors (TDs). Device internals include application, Operating System and modem.

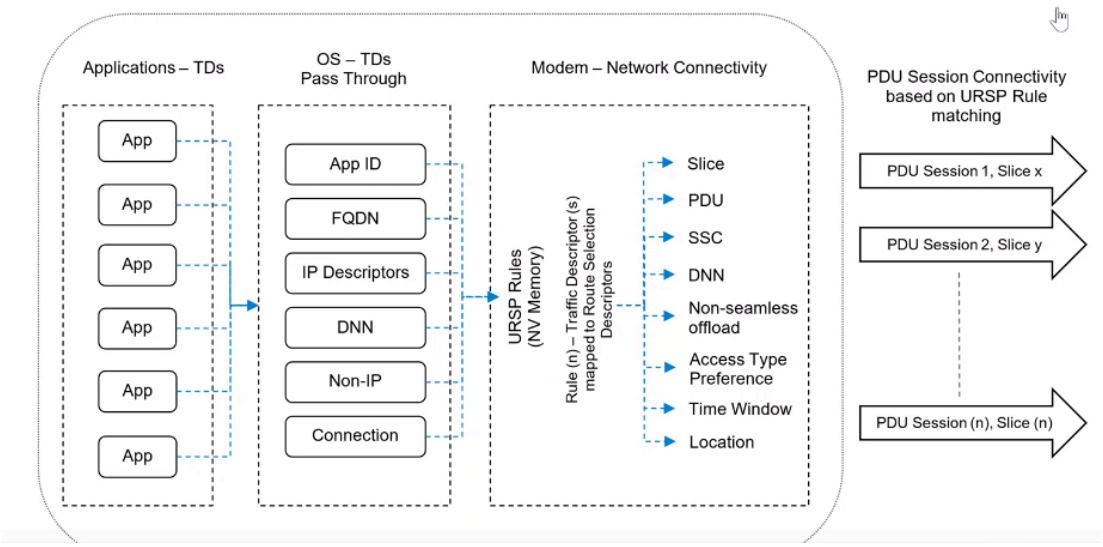


Figure 3-11: User Equipment Route Selection Policy rule construction.

The URSP is defined by the operator, and evaluated within the device, either by the Operating System (e.g., smartphones) or modem (e.g., IoT devices). This validation, referred to as URSP matching logic, is represented in Figure 3-12 and further elaborated in TS 23.503 [23.503]. It can be summarized in the points below.

- For every newly detected client application, the OS evaluates the URSP rules in the order of Rule Precedence and determines if the data traffic from triggered application matches the TD of a URSP rule.
- When a URSP rule is determined to be applicable for a given application, the OS shall select an RSD within this URSP rule in the order of the Route Selection Descriptor Precedence.
- When a valid RSD is found, the OS determines if there is an existing PDU session that matches all components in the selected RSD. If a matching PDU session exists, the OS associated the client application to the existing PDU session, i.e. route data traffic on this PDU session. Otherwise, the OS establishes a new PDU session using the values specified by the RSD.

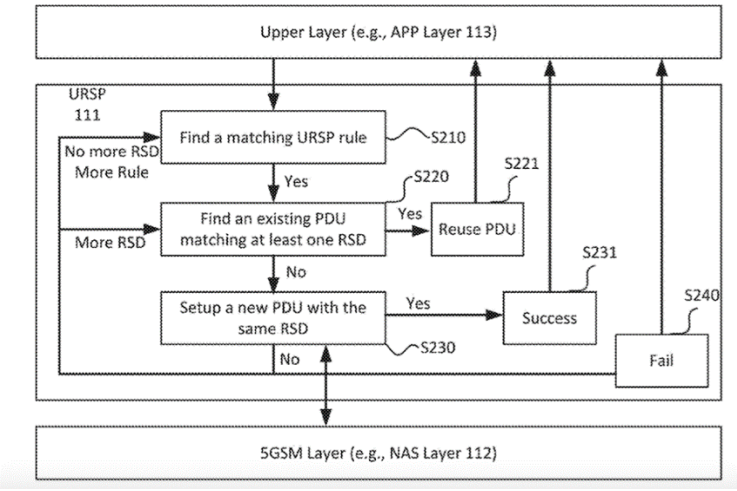


Figure 3-12: URSP rule matching logic

The Rel-18 solutions captured in these 3GPP documents provides the basis for further work in GSMA articulated into three activities: i) definition of Traffic Categories, useful to profile traffic from mass-market applications in URSP; ii) definition of OS requirements for URSP rule matching logic behaviour; and iii) handbook on URSP usage under certain scenarios. The results of i) are published in GSMA NG.135 [NG135], while the topics ii) and iii) are still ongoing on GSMA Terminal Steering Group (TSG) and Network Group (NG), respectively.

3.4.4.4 User-centric network management track: beyond SoTA

URSP solutions specified by 3GPP and GSMA allow operators to provide subscribers with optimal and personalized experiences when their device applications join the network. These solutions have focused so far on B2C services, considering applications from large-scale developers published in OS marketplaces. However, their impact on B2B, B2B2C and mission-critical services (e.g., emergencies, V2X) have not yet discussed. This is not straightforward, considering the differences between these market segments:

- B2C services will be delivered using Internet Access Services, while B2B/B2B2C/mission-critical services will be delivered using Specialized Services. This is imposed by today's network neutrality regulation.
- B2C services do not require operators to have visibility on which application is using which slice, while B2B/B2B2C/mission-critical services will do. The reason is that for these services, the operator shall be able to validate whether application qualifies to gain access to a certain slice, and if not, prevent their access.

These aspects, along with user consent management (the user shall be able to allow/disallow access, per application, to certain slices) and user data protection, will be handled to make sure Hexa-X-II tenant subscribers are provided with tailored experience while being compliant with regulation in force.

3.4.4.5 SLA enforcement track: SoTA

The SLA enforcement is a topic has been thoroughly discussed in the standardization arena (e.g., TM Forum) and in the 5GPPP community, including ICT-17 projects (e.g., 5G-VINNI, 5GEVE), ICT-19 projects (e.g., 5Growth, 5G-TOURS) and ICT-20 projects (e.g., FUDGE-5G). Despite their differences, all the solutions framed in this topic coincide on the following assumptions:

- SLA parameters only rely on network parameters that can be measured (i.e., quantitative parameters).
- SLA is rigid. SLA definition is the result of specifying threshold values for SLA parameters. Each parameter is allocated with one single value, expressed through a KPI.
- SLA is static. Once, the SLA cannot be modified throughout service lifetime. If the customer wants to modify the service conditions, a new SLA needs to be defined.
- SLA assurance is executed with threshold-based policies, typically defined with static rules.

3.4.4.6 SLA enforcement track: beyond SoTA

Hexa-X-II aims to go beyond the assumptions made in the literature review, by:

- Enriching SLA with non-quantitative parameters, including KVIs and TLAs.
- Making SLA elastic. This will be achieved by specifying range (non-threshold) values for SLA parameters.
- Making SLA dynamic. This mean SLA parameters can be modified "on-the-fly" and accommodate new requirements at operation time.
- Exploring how closed loop automation solutions can be reused and adapted to fulfil SLA assurance activities, and the advantages they exhibit against today's policy-based solutions.

3.4.5 Identification of possible components and interfaces

3.4.5.1 Resource controllability track

For this track, the solution consists in defining an information model for granular access control. This information model policies how to provision segregated yet customized management spaces to individual tenants.

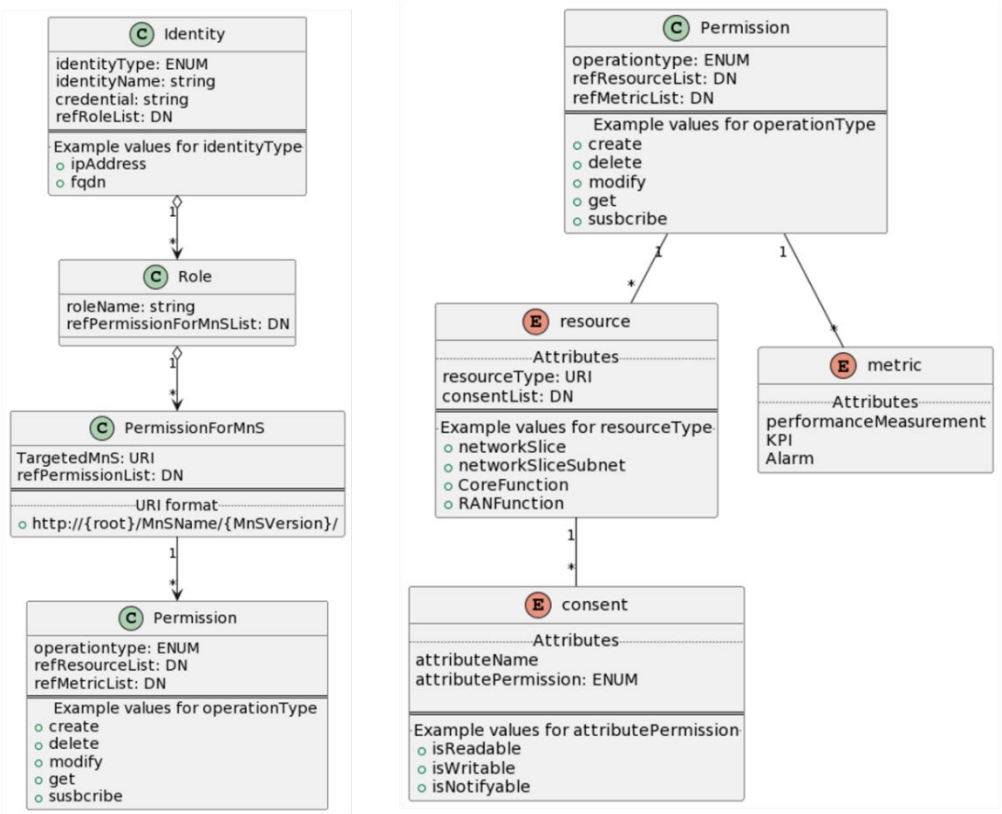


Figure 3-13: Information model for granular access control

As seen in Figure 3-13 this information model builds upon the following tenets:

- **Permissions are the most granular construction**. We have two types of permissions: i) provisioning permissions, for lifecycle management purposes; and ii) assurance permissions, for performance/fault management purposes. As to provisioning, the permissions specify which configuration actions (e.g., create, delete, modify) can be invoked over which managed resources (e.g., network slice, subnet, functions). As to assurance, the permissions specify which supervision actions (e.g., get, subscribe) can be invoked over which metrics (e.g., KPIs, alarms).
- **Roles represents a set of permissions**. The role concept enables the storage of information to what actions, resources, and metrics an authorized consumer can work upon. The roles are created by the

network operator at design time. They are typically defined according to the convenience and hierarchy of the organization. Examples of roles include “admin”, “provisioning-advanced”, “provisioning-basic”, “assurance-advanced”, “assurance-basic”, “etc.

- **Identity to role assignment**. The identity class represents the tenant identifier in the Hexa-X-II M&O system. By associating the identity with one or more roles, the management space for that tenant can be defined.

In the following, we will see the usage of this information model, at both design time and operation time.

Figure 3-14 illustrates the design time, which consists of two separate stages:

- No tenant is yet registered/onboarded in Hexa-X-II system. At this stage, the Hexa-X-II system administrator defines the different granular permissions. Once defined, they combine these permissions to create different roles.
- A tenant is registered/onboarded in Hexa-X-II system. An instance of the Identity class (see Figure 3-14) is created to represent that tenant in M&O system. This instance is assigned with one or more pre-defined roles. With this operation, the set of permissions for each tenant are specified.

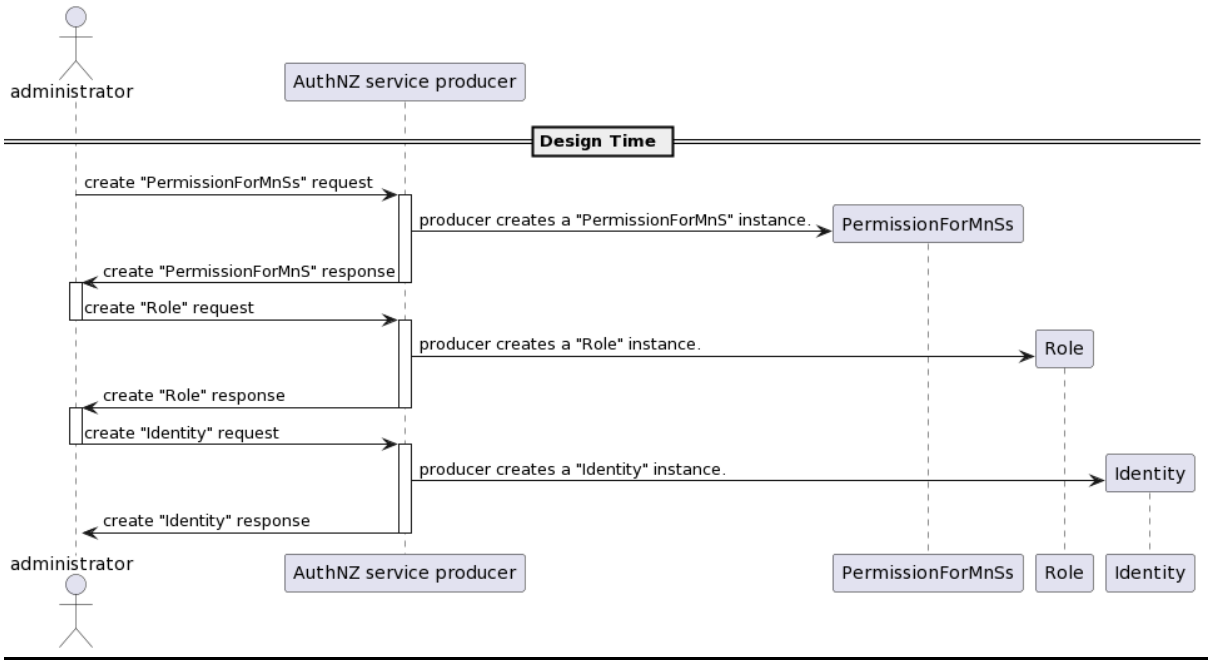


Figure 3-14: Design time

Figure 3-15 illustrates the operation time. This stage starts when the tenant wants to gain access and consume Hexa-X-II capabilities. As seen, the tenant is first authenticated. This is then followed by an authorization procedure, which returns an access token. The tenant uses issued token to request the consumption of a certain capability. The capability provider validates the token, and checks if requested actions/resources/metrics are allowed by that token. If allowed, the capability provider authorizes the request. The result of this request is returned to the tenant.

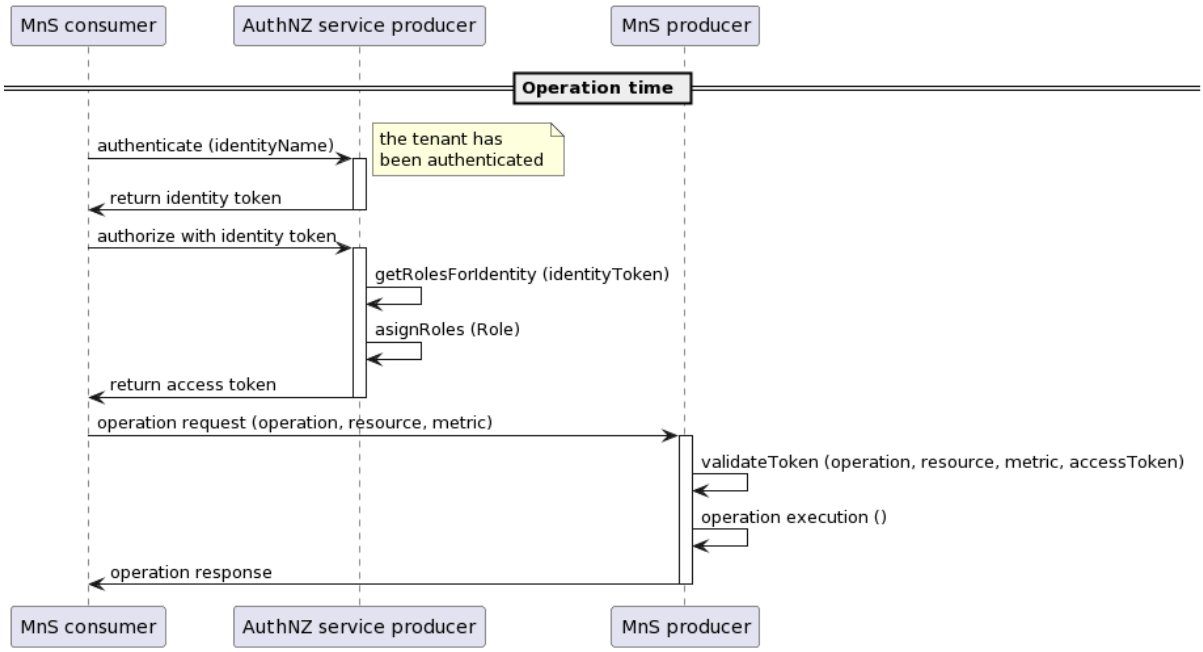


Figure 3-15: Operation time

3.4.5.2 User-centric network management track

Table 3-10 provides a summary of the most representative components of the Traffic Descriptor. From this analysis, it is worth noting that while for B2C slices (delivered as Internet Access Services) the most convenient option is to use "Connection Capabilities", for B2B/B2B2C/mission-critical slices (delivered as Specialized Service traffic), "Destination IP" and "Domain Descriptor" become plausible options. The reason is that these two TD components allow identifying destination network where the tenant application server is hosted, which is other than internet.

Traffic Descriptor component	Description	Features
Data Network Name (DNN)	**Definition:** This is matched against the DNN information provided by the application. **Scenarios:** Applies to special industry UEs with special DNN	• DNN has coarse granularity. • The large number of DNN can easily lead to difficult network operation and maintenance. • There is a security risk of DNN misused, additional security mechanisms to be considered.
Destination IP	**Definition:** Destination IP 3 tuples(s): {<IP address or IPv6 network prefix>, <port number>, <protocol ID>} **Scenarios**: Applies to special services with fixed and convergent application server IP addresses	• The IP address of the application server is variable • Some applications do not have static IP addresses and have limited application scenarios.
Domain Descriptor	**Definition**: FQDN which is used as a domain name matching criteria. **Scenarios**: Applies to applications with independent domain names, such as enterprise applications. FQDN is defined/meant to be used to represent the domain name of the server that device wants to communicate with	• Non-standard Domain Name System (DNS) is difficult to match URSP. • The device needs to monitor DNS, lead to delay problem.
Application Descriptor	**Definition**: It consists of OSId and OSAppId. This information is used to	• How to trust on the real identity of the applications? Nowadays there are no mechanisms

	uniquely identify the application that is running on the device's OS. **Scenarios**: Applies to more scenarios, easy to realize the application-level network slicing	in place to validate the identity of the application in the OS marketplace. • Operators need to trust on OS provider tagging classification, losing the e2e control.
Connection Capability	**Definition**: provide information on application expectations. **Scenarios:** Applies to when device application requests a network connection with certain capabilities	• They are used to encode GSMA Traffic Categories published in NG.135 [NG.135].

Table 3-10: Representative TD components of a URSP rule

The usage of "Destination IP" and "Domain Descriptor" is recommended because they provide operators with means to filter out the applications that can use the B2B/B2B2C/mission-critical slice, i.e., the applications that can be served by the tenant application server. This means that any rogue client trying to use the slice won't be able to gain access, as the only destinations that can be addressed and routed from the device through the slices are the IP addresses / FQDN of the network hosting the tenant application server. These IP address/FQDN are communicated by the tenant and provisioned by Hexa-X-II.

It is also important to guarantee that the IP address/FQDN of the tenant application server can be resolved in public DNS. DNS typically will be reached through the default internet access, so the complete design needs to take this into account.

For the solution of this track, we illustrate how the overall process works end-to-end.

When the tenant wants to deliver a Specialized Service to their subscribers, it needs to ask Hexa-X-II systems the provisioning of a network slice, which can be a B2B/B2B2C/mission-critical slice. To that end, it needs to communicate the following information:

- SLA of the network slice. This information is used to proceed with the instantiation and configuration of the slice.
- IP address/FQDN ranges of the application servers managing the specialized service traffics. This information is used to fill out TD components = "Destination IP" or "Domain Descriptor" in the URSP rule.

When the tenant wants to onboard their subscribers to the specialized service, the following steps will happen:

- The tenant delivers the identifiers of their subscribers to Hexa-X-II. The format of these identifiers (e.g., MSISDN, IPv4 + port, IPv6) depends on the concrete scenario.
- Hexa-X-II provisions the proper network slice on the user profile of these subscribers.
- Hexa-X-II sends the URSP rules to the tenant subscribers, via Core Network Functions.

With this solution, Hexa-X-II can ensure tenant subscribers can gain access with their applications to an optimal and tailored service experience. Once the URSP rules are provisioned on the devices, the subscribers remain in control in terms of what installed application can access to specialized services, and for which purpose. This is aligned with the user consent management propelled by regulation. The Terms & Conditions prompted/notified to the user to manage consent shall respect User eXperience (UX) guidelines.

3.4.5.3 SLA enforcement track

This track will leverage closed loop automation. The solution will consist in customizing the stages of the closed loop (monitoring, analysis, decision and execution) to the particularities of the SLA enforcement. As to the inputs/outputs of these stages, as well as their coordination, we will follow specifications in ZSM 009-1 [ZSM009-1]. Figure 3-16 pictures the ZSM view on closed loop concept.

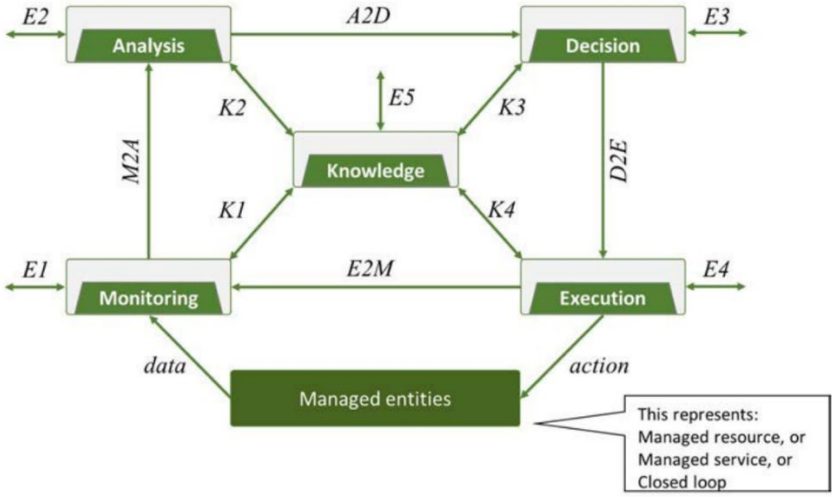


Figure 3-16: Functional view of a Closed Loop and its stages within the ZSM framework. Source: [ZSM009-1]

3.4.6 Relationship with other Enablers

Figure 3-17 pictures the tracks in scope of this enabler, and their relationship with other enablers and tasks.

User-centric network management track is related to device management, as the user profile and URSP rules will be managed by novel core network functions, including those providing data storage (in 5GC, it is UDM) and policy control (in 5GC, it is PCF).

SLA enforcement track is related to:

- Enablers #10 (closed loop governance) and #11 (closed loop control).
- Work on the translation, supervision and verifiability of the TLAs integrating the SLA, as described in [HEX223-D21].

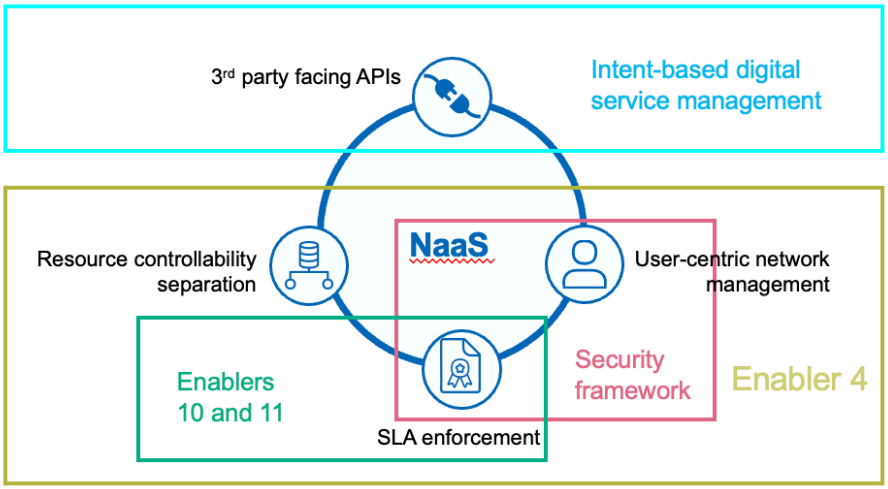


Figure 3-17:Trustworthy 3rd party management tracks

3.5 Enabler 5: Multi-cloud management mechanisms

3.5.1 Motivation

Nowadays, to satisfy both vertical customers and end users, service providers need to dimension and manage their infrastructure to fit the expected user demands. In the case of an infrastructure failure (e.g., major failure in a data centre) or a sudden increase of the number of users for a given service (e.g., concerts), service providers are often not able to react on time to keep a service operational. In these situations, thanks to

virtualization, it is less costly and more time-efficient to enable an operator to use services and/or resources provided by disaster recovery points of presence or owned by other operators/providers than to expand the local infrastructure. Furthermore, with the emergence of IoT and edge computing technologies, the need to manage dynamic workflows that are provided at both the edge and cloud part of the infrastructure is arising to achieve high performance. Distributed applications and services management has to be provided over infrastructure spanning across multiple clusters or clouds.

Distributed application deployment across a single cloud infrastructure is complex, especially when dealing with applications with external communication constraints. However, real-world application deployment adds up to that complexity when heterogeneous cloud infrastructures are necessary due to specialized needs that may arise, like the aforementioned ones. Such heterogeneity may come in the form of private or public clouds, core cloud or edge infrastructures and with requirements such as network coordination, service discovery and management of incoming computational load, making the process even more complicated. The transition towards distributed cloud computing approaches is highly relevant for the telecom industry towards 6G, considering that edge cloud computing will play a critical role in enabling emerging use cases with extreme service requirements in a variety of sectors [5GAM].

The multi-cloud management mechanism addresses the challenges of deploying a distributed application/service across multiple cloud infrastructures. Indicative examples of such applications/services regard solutions for first responders in emergency situations (real-time communication and analysis of data from IoT devices and video streams is required among persons in various locations), solutions for prompt and reliable decision making in Industry 4.0 (low-latency communication is essential for coordinating robots and machinery, monitoring processes in real-time, and ensuring safety through immediate response to anomalies), and remote healthcare (delay-intolerant communication with high bandwidth needs and data processing in various locations). Multi-cloud management mechanism provides a unified management layer that enables organizations to manage multiple cloud infrastructures from a single point of control. These mechanisms provide features such as workload management, resource orchestration, and policy enforcement that help organizations to deploy, manage, and operate distributed applications seamlessly across multiple clouds. By providing a centralized view of the cloud infrastructure, the multi-cloud management mechanism simplifies the deployment and management of distributed applications and helps organizations to meet their specialized needs.

A set of challenges exist for supporting multi-cloud management mechanisms. Given that part of the infrastructure belongs to different providers, there is a need for collaboration among them. The collaboration can be based on the conceptualization of a resources' abstraction layer and on the specification of open Application Programming Interfaces (APIs) for their interaction and the establishment of Service Level Agreements (SLAs) among them. In this way, efficient and collaborative utilization of resources among multiple cloud services may take place, considering –where applicable- resources' reservation and leasing schemes. Cloud brokerage solutions are also considered, where a broker intermediates between cloud providers and cloud consumers. Portability of distributed applications across multi-cloud infrastructures have also to be supported (e.g., by developing application in a form that can be deployed over cloud-agnostic containers).

Finally, challenges arise on the application of end-to-end monitoring platforms across multi-cloud infrastructure. Once again, the specification of open APIs can enable interoperability and interaction among monitoring mechanisms supported by the various providers.

3.5.2 Objectives

Enabler #5 pursues the following objectives:

- Implement and/or improve multi-cloud management mechanisms with the goal to achieve a unified approach to managing applications across multiple cloud environments, providing performance guarantees in multi-cluster deployments, while respecting the defined constraints (e.g., collocation of part of application components) and requirements (e.g., very low latency at the edge part, security in specific links).
- Tackle registration and management of resources across multiple clusters through unified APIs, taking advantage of proper resources abstraction.

- Manage and fuse information coming from the various clusters, providing proper abstraction levels and KPIs per layer of the infrastructure.
- Examine brokerage solutions for the collaboration among multiple providers and the enforcement of end-to-end SLAs over resources in the various clusters.
- Develop federated approaches for management of resources taking advantage of distributed ledger technologies.

3.5.3 Description of the solution

Multi-cloud management solutions are designed to address the challenges of deploying and managing applications across multiple cloud environments. These solutions provide a centralized interface for managing resources, services, and applications across various cloud providers. They offer features such as orchestration, automation, and monitoring to ensure efficient utilization of resources and seamless integration of applications. The design and implementation of centralized mechanisms to handle such challenges are necessary to guarantee the interoperability between a variety of cloud providers, the efficient utilization of the available resources and the secure computation and communication between the deployed services. These mechanisms aim at implementing and improving state-of-the-art techniques that demonstrate high potential in such complex and dynamic environments such as AI-assisted scaling and load balancing methods, while also taking advantage of existing tools [KAR23][LIQ23][KAT23] that can facilitate the efficient deployment of these techniques. This cooperation between novel algorithms and developed technologies is enhanced with observability monitoring data as input, providing a holistic view of the applications' performance and execution state at real time. Having an analytical viewpoint of the deployment can support not only immediate actuation, but also the enabling of complex decision making and thus, explore how state-of-the-art solutions such as Deep or Reinforcement Learning can be applied in real-world environments and what improvements these can offer.

One approach to multi-cloud management is to use a cloud management platform (CMP). CMPs are software tools that provide a single interface to manage multiple clouds and automate tasks such as provisioning, monitoring, and cost optimization. They enable users to deploy and manage applications across multiple cloud providers with minimal effort and reduce the complexity of managing multiple cloud environments.

Another approach is to use container orchestration tools such as Kubernetes, which provide a framework for managing containerized applications across multiple clouds. Kubernetes enables users to deploy and manage containerized applications in a scalable and efficient manner, regardless of the underlying cloud infrastructure.

Multi-cloud management solutions also incorporate AI and ML algorithms to optimize resource allocation, predict and prevent service disruptions, and improve application performance. They provide real-time visibility into the performance and health of the deployed applications, enabling users to proactively identify and resolve issues.

In this enabler, the multi-cloud management mechanisms aim at implementing and improving state-of-the-art techniques that demonstrate high potential in such complex and dynamic environments such as AI-assisted scaling and load balancing methods, while also taking advantage of existing tools that can facilitate the efficient deployment of these techniques.

The developed multi-cloud management mechanisms should provide open interfaces towards network management platforms to support the interplay between network management mechanisms and compute resources management mechanisms. Overall, the developed solutions should provide a unified approach to managing applications and services across multiple cloud environments, reducing complexity, improving efficiency, and enabling the interplay between compute and network management systems.

Multi-cloud federation in dynamic environments faces a plethora of challenges including i) the trade-off between the administrative cloud domain's openness and preserving the privacy, security, and trust; ii) robust monitoring of the availability of participating administrative domains; iii) dynamic pricing and billing for federated services with secure, trustworthy and dynamic SLA agreements; iv) guaranteeing QoS across dynamic federated cloud domains ; and v) achieving high security and privacy interconnection schema between administrative cloud domains.

We propose the application of **DLT for multidomain federation** to address the challenges above. Admission control will be dependent on the blockchain governance policy. For example, in a permissioned blockchain, a common approach is to accept members via voting. Although domains may act maliciously and reject the entry of new members, domains typically have an incentive to increase participants. The availability will be guaranteed by the incentive of each domain to maintain an active blockchain node. This improves the blockchain network security (avoiding 51 percent attacks) and increases the domain's usage budget (e.g., gas in Ethereum). Security and privacy will be established by limiting the usage budget and the use of cryptography. Newly joined domains will have a lower limited usage budget or a limited number of federation announcements, and thus are unable to spoof or spam the participating domains. Communications between domains will be recorded and validated as immutable transactions on the ledger where cryptography will be used to preserve the privacy of the data in the transactions exchanged. Finally, dynamic pricing and billing, and multi-domain QoS will be achieved by implementation of dynamic SLAs and QoS monitoring. The use of **smart contracts** is a promising solution towards the integration of both dynamic SLAs and QoS monitoring.

3.5.4 SoTA and Beyond SoTA

Considering the advances made in edge and cloud computing orchestration mechanisms and platforms, various solutions are emerging for management of multi-cloud resources for the deployment of distributed applications and services across the computing continuum [DCK23] [CBR+22] [XLC+22] [TOM+20].

The current main challenges to be tackled regard the application lifecycle management, the efficient resource allocation, interoperability aspects, SLAs management, observability in the various parts of the infrastructure, and combination of compute and network orchestration mechanisms [BAC+22].

The **application lifecycle management** involves managing applications throughout their entire lifecycle, from their composition, provisioning, deployment and operational phase, considering actions such as scaling, monitoring, and decommissioning across multiple cloud domains [GMM+21]. The challenges associated with lifecycle management include the need to ensure smooth application deployment across diverse/multiple cloud environments, the need to effectively handle dependencies and complex configurations in microservices-based software development, and the need to develop intelligent orchestration mechanisms that can support automated deployment and decision-making. CI/CD processes have to be also made available. In our work, we are going to consider the adoption of intent-driven orchestration mechanisms, combined with distributed management techniques to manage the deployment of distributed applications/services across multiple clusters in the continuum.

The task of **resource allocation** encompasses a range of operations tied to common edge/cloud orchestration issues, such as service placement, task scheduling, and virtual network embedding [BAC+22]. The placement involves the specification of the physical or virtual resources –or a part of them- that can host an application or a service, such as edge cloud infrastructure nodes or containers. The task scheduling entails the decision regarding when and where specific tasks or workloads should be executed to optimize resource usage and ensure high-quality service for applications. A growing concern related to task scheduling is compute offloading, especially in the case of IoT and edge computing applications. Virtual network embedding involves the mapping of virtualized application components or services to physical network resources, posing a set of challenges for optimal resource selection, reservation and usage. Challenges in resource allocation include efficient load balancing, taking into account any resource limitations and performance requirements, while aiming to optimize resource utilization and provide mechanisms for resource autoscaling and container migration across multiple domains to meet SLAs and performance objectives.

Interoperability challenges also arise due to the massive number of heterogeneous devices running different protocols, as well as the consideration of various types of compute and network resources across the computing continuum. There is a need for the development of components based on open technologies, open APIs and –where possible- open-source software.

With regards to **SLAs**, various works are tackling the issue of **definition and negotiation among providers** for managing SLAs that adhere to the requirements of applications and stakeholders [FOL+23]. Orchestration agents must monitor and enforce SLAs, ensuring that applications receive the agreed-upon levels of performance, availability, and resource allocation.

Modern **observability frameworks** are also emerging, considering the need for data fusion of various types of signals, such as resource usage metrics, QoS metrics, traces and logs [TZA22].

Considering the need for **network orchestration in multi-cloud environments**, this includes the development of mechanisms that support reliable network connectivity across multiple environments, while enabling the enforcement of traffic management and security policies to satisfy the deployment requirements and optimize performance [CLA+21].

3.5.5 Identification of possible components and interfaces

A multi-cluster management solution has to support proper abstraction of resources across the computing continuum (IoT, Edge, Cloud resources) and provide unified management mechanisms over compute and network resources. Each cluster may advertise its own resources and service consumption endpoints and make them available to a higher-level multi-cluster management entity. Such an approach is depicted in Figure 3-18.

The end-to-end Service/Application Orchestrator is responsible to receive and manage a deployment request for a network service or a distributed application graph. The request can be accompanied with a set of objectives that have to be fulfilled, along with constraints in terms of deployment. This information can be mapped and represented in the form of an SLA that has to be adhered. Following, the deployment request is tackled by a multi-cluster manager and a network manager. Such managers are interacting to properly facilitate the request under distributed compute and network resources. The resources may span across multiple clusters in the continuum, from IoT to edge to cloud resources. An observability stack can be considered that provides information regarding QoS metrics, compute resource usage metrics and application/service-specific metrics. These metrics can be considered by the applied orchestration mechanisms and refer to a specific hierarchy level, as depicted in Figure 3-18.

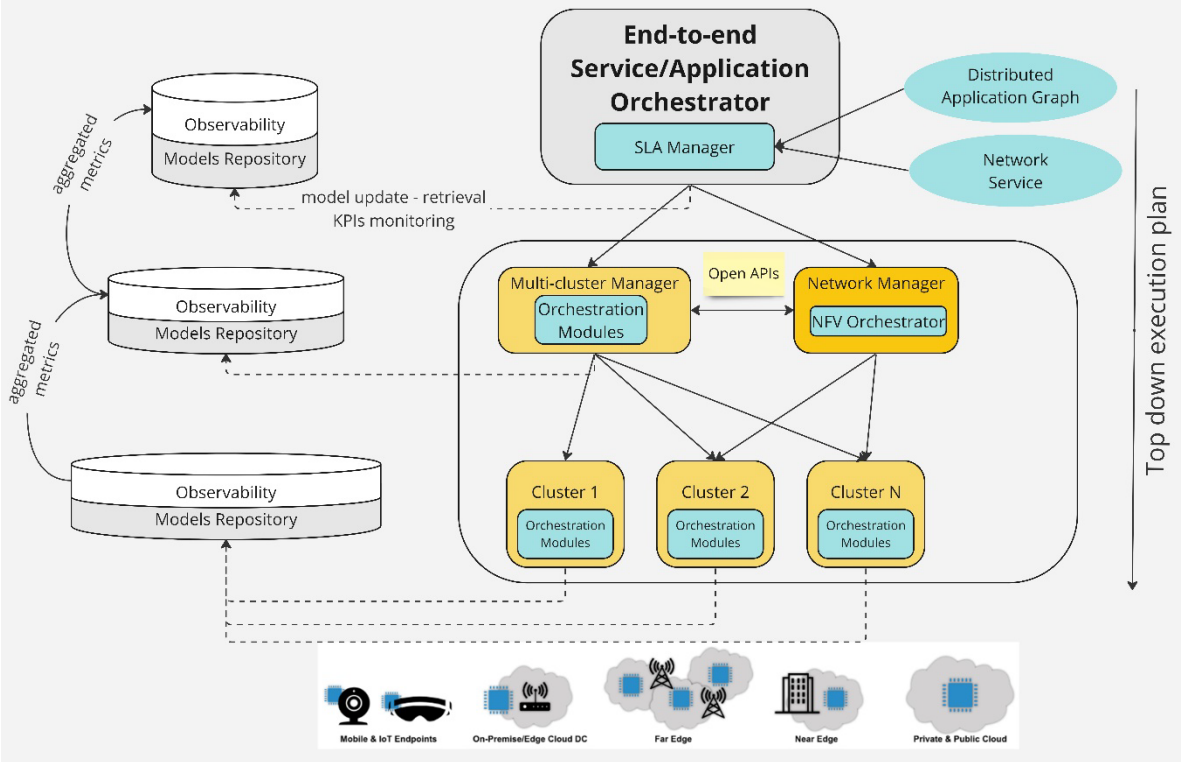


Figure 3-18: Multi-cloud infrastructure management approach

Tools to be considered:

a) **Multi-cluster compute resources management tools**:

Liqo [LIQ23]: Liqo is an open-source project that enables dynamic and seamless Kubernetes multi-cluster topologies, supporting heterogeneous on-premise, cloud and edge infrastructures.

Karmada [KAR23]: Karmada is a Kubernetes management system that enables you to run your cloud-native applications across multiple Kubernetes clusters and clouds, with no changes to your applications. By speaking Kubernetes-native APIs and providing advanced scheduling capabilities, Karmada enables truly open, multi-cloud Kubernetes. Karmada aims to provide turnkey automation for multi-cluster application management in multi-cloud and hybrid cloud scenarios, with key features such as centralized multi-cloud management, high availability, failure recovery, and traffic scheduling.

Open Cluster Management [OCM23]: Open Cluster Management is a community-driven project focused on multicluster and multicloud scenarios for Kubernetes apps. Open APIs are evolving within this project for cluster registration, work distribution, dynamic placement of policies and workloads, and much more.

b) Network services deployment over multi-cluster topologies:

Open-source MANO [OSM23]: Open-Source MANO (OSM) is an ETSI-hosted project to develop an Open-Source NFV Management and Orchestration (MANO) software stack aligned with ETSI NFV. Efforts are in progress to support the registration of multi-cluster resources to OSM.

Camara project [CAM23]: CAMARA is an open-source project within Linux Foundation to define, develop and test the APIs. CAMARA works in close collaboration with the GSMA Operator Platform Group to align API requirements and publish API definitions and APIs. Harmonization of APIs is achieved through fast and agile created working code with developer-friendly documentation. API definitions and reference implementations are free to use (Apache2.0 license). Abstraction from Network APIs to Service APIs is necessary: to simplify telco complexity making APIs easy to consume for customers with no telco expertise (user-friendly APIs); to satisfy data privacy and regulatory requirements; to facilitate application to network integration.

ONAP [ONA23]: ONAP is a comprehensive platform for orchestration, management, and automation of network and edge computing services for network operators, cloud providers, and enterprises. The ONAP MultiCloud project aims to mediate most interactions between ONAP and any underlying VIM or Cloud to enable ONAP to deploy and run on multiple infrastructure environments. Besides that, the ONAP MultiCloud project enables infrastructure providers exposing infrastructure's resources and features to ONAP for optimization of homing and placement of VNFs and supports the closed control loop remediation over infrastructure resources

Network Service Mesh [NSM23]: Network Service Mesh allows individual workloads, wherever they are running, to connect securely to Network Service(s) that are independent of where they run.

3.5.6 Relationship with other Enablers

A strong relationship of this enabler is identified with **Enabler 6** (Orchestration mechanisms for the computing continuum), since the ability to manage resources across clusters in a unified way is exploited by the orchestration mechanisms developed for the computing continuum. A relationship is also identified with **Enabler 3** (Integration fabric) since the developed mechanisms can be introduced and provided through an integrated resources-management solution.

3.6 Enabler 6: Orchestration mechanisms for the computing continuum

3.6.1 Motivation

The rise of IoT and Edge Computing has extended the deployment possibilities of microservices-based applications and services to a wider Compute Continuum including not only the Cloud layer, but also the Edge and the Extreme Edge/Device layer. It is in this sense, that multi-location computing techniques become of the essence with regard to computing continuum. The emergence of IoT technologies combined with the usage of advanced networking mechanisms and the convergence of IoT management platforms with MEC platforms introduces novel techniques for management of distributed applications and services at the extreme edge part of the infrastructure. The high volatility of the applied workloads, combined with the heterogeneity of the available IoT resources impose a set of challenges to be addressed for unified and efficient management of

applications workloads at the extreme edge and edge part of the infrastructure. The need for management of massive IoT devices in some cases introduces further complexity that has to be tackled by orchestration mechanisms in the computing continuum. The latter includes compute and network resources made available at the extreme edge, the edge and the cloud layer of the infrastructure. Indicative examples of applications that consider management of resources in the computing continuum are provided in the description of the Proof of Concepts (PoCs) in Section 4. By the term extreme edge, we refer to the outermost layer or boundary of an IoT network or architecture. It represents the devices and sensors that are situated closest to the physical world or the point of data generation, or their virtual counterparts (e.g., digital twins). The edge layer is situated between the extreme edge and the cloud. It includes more powerful devices, such as edge gateways and edge servers. Edge devices perform functionalities such as localized processing, filtering, and aggregation of data, while they help to reduce the latency, optimize the data transmission, and enhance real-time responsiveness to dynamic workloads. The cloud layer is the central part of the overall infrastructure. It encompasses cloud servers and data centers that store and process vast amounts of IoT data. The available cloud resources may handle complex analytics, long-term storage, and data visualization processes.

Orchestration of non-complex systems such as single cloud infrastructures, though, usually takes place in a central manner and can introduce major communication overhead and potentially flood the network with messages. At the same time, information to manage such heterogeneous systems may not always be available to the orchestrator due to the different authorities that may control the different infrastructures (e.g., different cloud/edge providers, telco providers etc). Thus, it is important to handle scenarios where centralized control of distributed applications is not always efficient or even possible and enable distributed management of their individual components. This, in turn, dictates the need for identification of the different entities that should be responsible for the decision making in a variety of problems (e.g., scaling, placement, migration, load balancing etc.) and the abstraction layer they exist in the continuum, i.e., the application layer as service entities, the virtualization layer as container/VM entities, the infrastructure layer as node entities, etc.

The complexity of such systems gives rise to old and new challenges regarding application orchestration in a compute continuum strongly characterized by heterogeneity. We describe some of these challenges here:

Resource provisioning

- *Service scaling*: Incoming workloads are time-variant, especially in highly dynamic environments such as hybrid edge-cloud ecosystems, so it is of vital importance to utilize the available resources efficiently, minimizing latency, without overconsuming them at the same time, maximizing resource availability. Thus, orchestrators should be able to dynamically upscale and downscale the deployed services according to the real-time incoming traffic. Scaling can be horizontal, i.e., creating multiple replicas of a service, as well as vertical, i.e., allocating additional resources to the already deployed replicas, or hybrid.
- *Service migration***:** The migration of service instances across the compute continuum is increasingly complex in comparison to monolithic infrastructures, since it concerns multiple clusters of different types and in different locations, so orchestration mechanisms should consider a variety of characteristics that may influence the resulting performance.
- *Service placement*: The initial placement of a service, which could consist of a single task, a group of independent tasks, or a group of dependent tasks in graph-based structures. Due to the variety of service structures and infrastructure architectures in the continuum, application deployment policies should consider a wide range of placement schemes.

Task scheduling

- *Task dispatching***:** In hierarchical architectures such as edge-cloud infrastructures, it is common to optimize execution by offloading tasks to higher layers to exploit more powerful resources. Especially in the IoT domain, decisions for executing workloads locally or at the edge or cloud resources are often a crucial point for the application's success.
- *Load balancing***:** Traffic balance among instances of the same service deployed across the continuum is crucial for fully exploiting the available resources when scaling is also present. According to different requirements load balancers can potentially follow increasingly complex policies, especially in multi-cloud environments.

3.6.2 Objectives

The main objectives of Enabler #6 regard:

- Enhance convergence, interoperability and openness of orchestration solutions for the computing continuum, by promoting the synergy between different providers or orchestration entities.
- Increase the distributed intelligence and the autonomy of orchestration mechanisms, considering a "system of systems" approach.
- Enable the decentralized management of applications and services, considering local and global optimization objectives.
- Incorporate new mechanisms for extreme-edge domain orchestration able to cope of huge heterogeneity and volatility of this kind of resources.

3.6.3 Description of the solution

The management of applications and services across the computing continuum requires in many cases the involvement of a set of orchestration mechanisms and platforms, given that the management of resources may be assigned to different stakeholders. Different types of synergies may be considered, such as:

- the synergy among multiple cloud providers in the deployment of an application over multi-cloud infrastructure;
- the synergy between a cloud application provider and a network provider;
- the synergy between multiple agents that support the operation of an application or service over programmable infrastructure belonging to one provider.

In the first case, multi-cloud resource management solutions have to be applied, considering resources that span across the computing continuum from IoT devices to extreme edge/edge infrastructure to cloud computing infrastructures. In the second case, open Application Programming Interfaces (APIs) have to be made available as northbound APIs on behalf of the network providers. Such APIs may be consumed on demand by the cloud application providers and can be accompanied by relevant Service Level Agreements (SLAs). In the third case, agents are assigned to application, virtualization or infrastructure elements, managing and configuring their operation. This includes enabling service scaling, migration, placement actions to the corresponding agents, them being responsible for specific parts of the continuum such as applications/services, specific physical (cloud/edge nodes, IoT devices) or virtual (containers, VMs) infrastructure elements etc.

Synergy, by definition, assumes multiple agents that interact with each other to either collaborate, compete or simply co-exist in shared environments. Multi-agent systems (MASs) are built on this logic to distribute functionality among a number of agents in environments where individual agents do not have enough information or resources to achieve their objectives. Instead, agents must cooperate on their individual objectives and collaborate on shared ones, communicating their understanding of the environment and their progress towards the objectives. Thus, the behaviour of a MAS emerges through the actions and interactions of autonomous or partially autonomous individual agents, with the guidance of an orchestrator or through a choreography of the autonomous participants.

In complex systems, such as the spectrum of the compute continuum, a MAS often needs to autonomically reorganize itself to adapt and evolve, in response to changes in the participating agents or in the external environment. AI and ML have shown significant results in building autonomy based on collected real-time and historical information from their environment guiding decision making. AI agents can learn to be reactive, proactive and collaborative based on what they observe in previous experiences and thus, have the potential to demonstrate significant performance in heterogeneous and continuously changing environments.

A popular paradigm is that of Reinforcement Learning which has been proved very effective in building autonomous agents for various tasks, such as request dispatching with Multi-Agent RL [HSW+21] or fog cluster scaling [SMO+20], due to its ability to relate state-action pairs with their corresponding rewards. Such agents often demonstrate codependent requirements due to their complex relations, so agents that manage microservices in microservice networks can interact with each other to satisfy requirements that arise from their exchanges, while agents managing computing nodes in a network in the continuum can distribute incoming workloads.

In order to successfully manage such systems in the compute continuum, it is important to create mechanisms flexible enough to enable the exploitation of inter-agent relations across the continuum and economic enough to minimize network requirements for orchestrating services at real-time. Thus, information extraction of these relations is crucial to the formation of optimal service orchestration strategies, while keeping the process at a local level.

We consider the application graph defined by such relations, consisting of static and dynamic information collected by the observability monitoring mechanisms (see Figure 3-19). The graph structure of these applications and the corresponding infrastructure can be analysed and provide this information with the help of hand-crafted features, as well as automatically extracted ones. For instance, Graph Neural Networks [SGT+09] (Graph Convolutional Networks -GCN-, Graph Attention Networks – GAT - [VCC+17] etc.) have been proved very efficient in extracting such features from graph structures, especially in cases of communication networks [ZYF+22][TSK+22]. It is difficult to apply them in huge graphs because of their depth constraints (their complexity increases exponentially for each depth layer), but they can provide useful insights for shallow networks such as those of microservices or edge/cloud clusters.

The work that the enabler will attempt to bring into the project is comprised of three main directions:

- Application graph analysis and application embeddings extraction. The application graph will originate from the observability data fusion mechanisms [HEX223-D21] and will hold information from distributed tracing, to resource metrics, to logging metrics. This direction is to be exploited by service workflow orchestration mechanisms, capturing complex qualitative as well as quantitative dependencies (e.g., via Graph Neural Network – GNNs -) between services that can indicate how service interplay can influence decisions.
- Exploitation of the application characteristics and real-time performance for enabling timely and accurate orchestration actions, such as autoscaling and scheduling decisions. Focus will be given in distributed management of applications either by assigning autonomous agents (e.g., using MADRL) to applications or services for self-management actuation (e.g., scaling, migration) and/or by placing autonomous agents across the infrastructure for decision making at a local level (e.g. request dispatching, scaling).
- Enhancement of application orchestration in synergy with network provider requirements and SLAs applied via the corresponding interfaces.

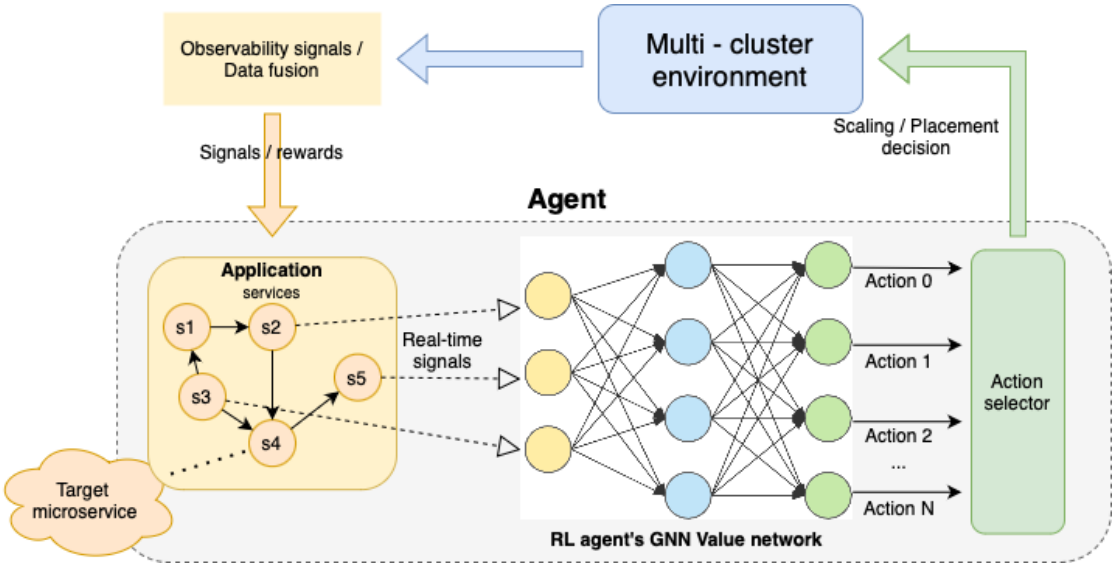


Figure 3-19: Distributed application/service graph managed by RL models in a multi-cluster environment

3.6.4 SoTA and Beyond SoTA

In modern computing environments, such as those that will be studied for Enabler 5's multi-cloud mechanisms, heterogeneous deployment will be automated, while interoperability between different types of resources will eliminate the need for cluster-specific reconfiguration. Thus, this enabler aims to exploit such environments

by moving service orchestration to a higher abstraction layer than classic computing, covering a wide range of computing resource providers and supporting a wider set of applications.

The mechanisms that will be developed will extend monolithic service orchestration by focusing on the analytically monitored relations between different services and the corresponding heterogeneous environments. Models of these relations, based on historical data collected by the application graphs, will provide generic guidelines for managing the related services, while continuous updates of the graphs with the use of real-time data will guarantee the live optimization of the application's deployment.

Additionally, the enabler will build on existing control techniques by introducing intuitive features extracted from the application graphs to optimize orchestration in a fine-grained manner. Classic approaches mainly focus on individually managing computing nodes towards specific KPIs, so the mechanisms will use autonomous but interdependent agents enabling distributed management, which is crucial for the compute continuum strongly characterized by decentralized architectures. Thus, the tools will face a series of classical orchestration challenges (e.g., scaling, migration, scheduling) in a new and complex environment, while using novel and detailed features for richer input.

3.6.5 Identification of possible components and interfaces

Based on the planned work, we envisage the following components for its implementation:

- Application graph analysis mechanisms
- Network coordination mechanisms
- Distributed service autonomous agents
- Distributed node autonomous agents

The observability mechanism [HEX223-D21] offers a variety of monitoring signals that can be used for modelling the application. One of the possible components of the enabler is a modelling and analysis tool of the application graph. The tool should be able to debrief the characteristics of the application's services (average execution time, resource consumption etc.) and capture their inter-dependencies (latency, request data loads etc.) based on historical and real-time data obtained by the observability mechanism. Similar work should be considered for the network metrics and requirements in order to coordinate the compute continuum deployments with the corresponding network providers.

In order to manage the distributed components of the applications as well as the equally distributed resources, a set of autonomous agents should be put in place, having access to the compute nodes' orchestration interfaces. These agents can manage the resources allocated to individual application services or can be responsible for the resources of whole computing nodes, and thus, they should be structured to receive the corresponding feedback signals. Theoretical approaches derived from the area of multi-agent systems for complex systems management will be exploited. The adoption of ML techniques is considered as granted for the guidance of the multi-agent systems. A high-level view of such an approach is illustrated in Figure 3-20.

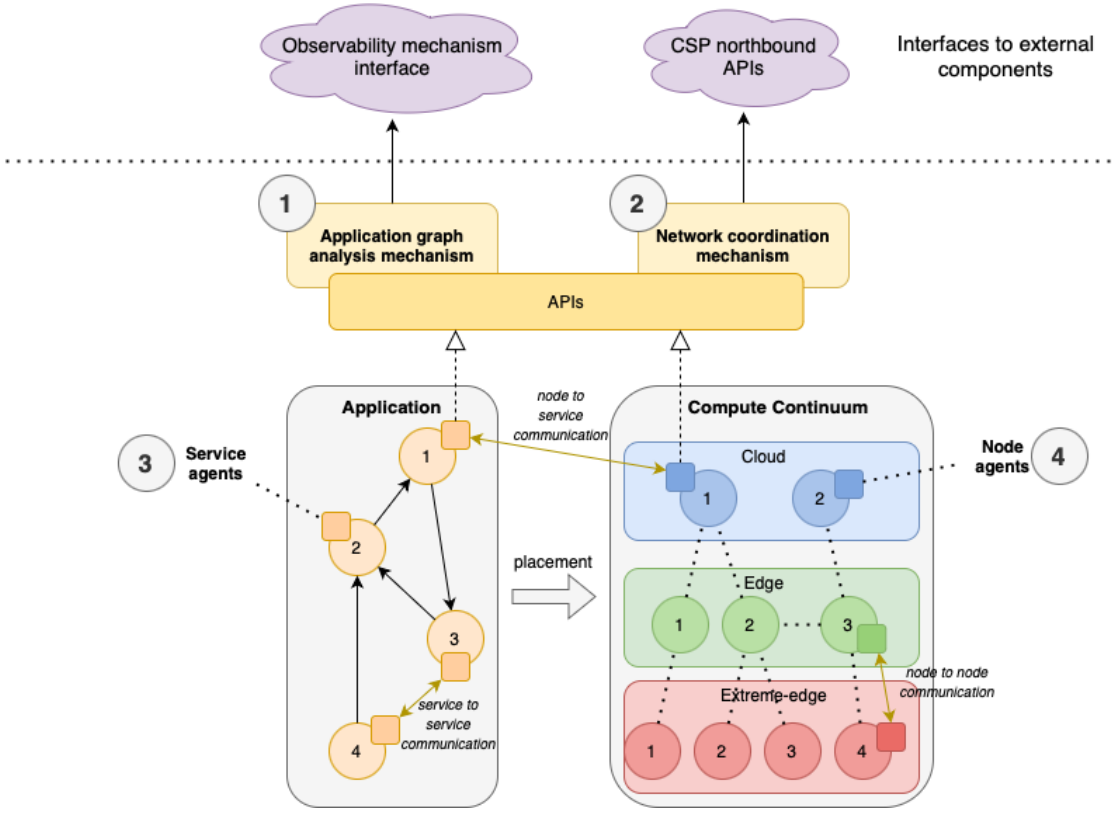


Figure 3-20: High level view of a synergetic orchestration approach in the computing continuum

Since this enabler is based on a distributed logic, its design is also based on a distributed architecture on each deployable node of the continuum. As shown in Figure 3-21, each node deployed should consist of the corresponding offerings such as interfaces for accessing the selected orchestration technologies for each node (e.g., Kubernetes, k3s, KubeEdge etc.), the observability mechanism's local deployment and the local versions of the developed components.

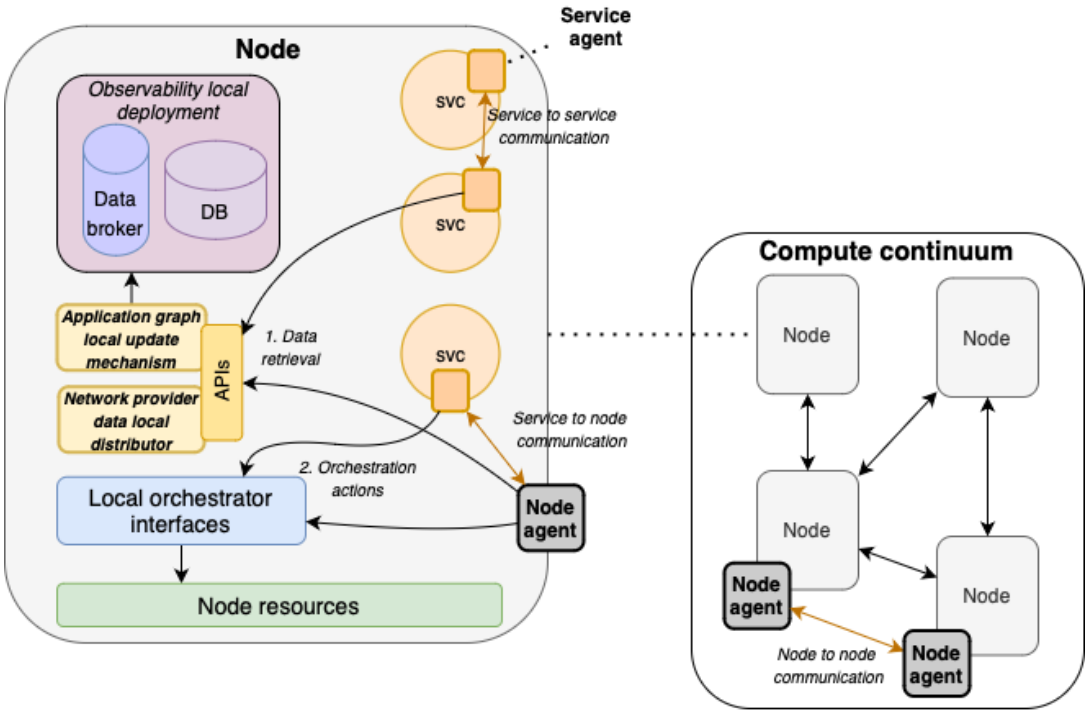


Figure 3-21: Distributed interaction logic between multiple agents

3.6.6 Relationship with other Enablers

The enabler will build on synergies with external tools such as multi-cloud management (**Enabler 5**) and observability data fusion tools for obtaining data or applying orchestration actions. Additional synergies will come from the network domain with which the developed mechanisms will try to coordinate to satisfy shared requirements. Furthermore, PoC-B.1 (see section 4.2) will take up to deploy a set of latency-dependent components on the compute continuum, so the developed orchestration mechanisms will be called upon to optimize the PoC's performance and make guarantees for its requirements.

3.7 Enabler 7: Sustainable AI/ML-based control

3.7.1 Motivation

Operators of 5G networks are actively seeking to reduce their operational expenses by implementing energy-saving management solutions to decrease power consumption. To achieve energy savings, they execute specific actions with optimized parameter configurations, such as energy-saving state switches. However, the numerous combinations of energy-saving actions can pose challenges for operators in determining the most effective actions to take. For instance, different energy-saving actions may conflict with each other, or they may clash with other network performance targets. Furthermore, aside from the possible energy-saving actions and their impact on service performance, the rise of edge computing adds another layer of complexity in sustainable network management. The addition of extremely distributed possible locations for network functions (i.e. the UPF, being the 5G version of packet forwarder) makes it impossible for current levels of automation to cope with at scale, with many hundreds of services running on the network. When combining all these perspectives, it quickly becomes difficult to calculate or predict what any action would do to service performance and energy efficiency.

Besides the above, 6G is posing very critical requirements regarding environmental sustainability (see section 2.1.1), since it is not only a matter of improving energy efficiency in order to reduce energy costs but also about the associated overall carbon footprint. Reducing energy consumption is a way to achieve decrease of carbon footprint but this may also be the case with involving greener energy sources and make decisions based on these sources.

By using intelligent mechanisms, implemented with AI/ML algorithms, these problems become more tractable. 6G is expected to incorporate AI/ML mechanisms on a massive scale as part of a distributed data driven network architecture with the goal to take the network automation to an unprecedented level [YAX+20]. However, considering the KVI on environmental sustainability, AI/ML may be adding its own carbon footprint (especially on the training procedure), therefore a necessity arises to actually design and implement ML algorithms while keeping their energy consumption at the minimum possible level in the trade-off with other KPI performance levels [VHI+21].

3.7.2 Objectives

The main objective of this enabler is to improve the process(es) of network management from automated operations towards full autonomy and include environmental sustainability as another objective next to performance. The addition of objectives to the management process even further increases the complexity of the problems both in the amount of information incoming from the network and the possible actions to be taken. This is the main reason why AI/ML algorithms are the pillar of possible solutions.

3.7.3 Description of the solution

The concept of intent-based networking (IBN) is gaining renewed attention due to both increased demand and advancements in technology. IBN aims to enable faster and more efficient service delivery by utilizing high levels of automation that abstract and simplify network management and reduce the need for human intervention. Also, the process involves service assurance with several essential functions, including monitoring, analysis, planning and execution, that work together to ensure that the network meets the intended service requirements based on its design and purpose. In this regard, fulfilling QoS and Quality of Experience (QoE) requirements of services necessitates these functions, and it can be best implemented with the principles

of autonomous closed loop control policies. These loops can be instantiated upon the reception of new intents (such as a new QoS or QoE requirement), and each closed loop can be responsible for various Key Performance Indicator (KPI). An intent can also be 'minimize RAN energy consumption' or 'minimize energy consumption for some specific services or at some specific areas', which would result in energy-saving and can enable the network to analyse and determine the ideal trade-off between achieving energy savings and providing a satisfactory service experience.

This analysis and determination will be performed by optimisation algorithms/mechanisms which will be developed to optimally (re)allocate resources for the various functions of the system (e.g., services, workloads, tasks). In order to cope with the extreme location distribution, and determine the number, placement and resources of the edge NF instances, the determination mechanism will consider the changes in traffic and application distribution. Current solutions for this are somewhat static, with user mobility and traffic changes only sometimes taken into account on very long timescales. These changes are more dynamic, however, and require adjustments in the NF configuration by scaling, migrating, or deploying new instances. This dynamicity needs to be taken into account in a proactive manner to further increase efficiency (energy or other KPIs) possibly by predicting changes with AI/ML models and performing the actions in advance. Furthermore, different user plane functionalities (i.e. for 5G networks UPF PSA, intermediate UPF, uplink classifier, branching point) might be separated and only deployed on the instances where they are needed.

In order to make the whole management process aware of the energy consumption, metrics will need to be collected explicitly considering renewable energy contribution. They will be divided/mapped onto adaptive AI/ML-driven analytics [6GR23-D21] at the 6G control and management planes and used for optimal energy efficient placement of functions and their configuration. On top of this, sustainable MLOps tools will be incorporated to provide support to cover all phases of the ML model process, e.g. coding, training, evaluation and execution.

In addition, the energy consumption metrics can be also considered for end-to-end energy efficient federation [XB20] of network functions and orchestration tasks in a multi-domain system. Current federation solutions only consider the federation of different network functions from a network service. However, the existing AI/ML algorithms that provide orchestration and management of the network service also may be considered for federation having in mind their high energy footprint. This federation of AI/ML mechanisms needs to tackle multiple aspects that are not only limited to energy efficiency but also include security, latency and other KPIs in order to achieve optimal multi-domain network performance. Some of these concepts will be integrated into PoC#A.1 which is described in section 4.1, where co-bots are utilised for conducting manufacturing or rescue tasks depending on the use case of interest. The necessary ICT features (functionality, data, maps) to support the various co-bots' capabilities (e.g., camera, mobility, wheels, propellers, etc.) will be additionally considered in the resource allocation problems. The general methods and solution concepts of Enabler 7 are depicted on Figure 3-22.

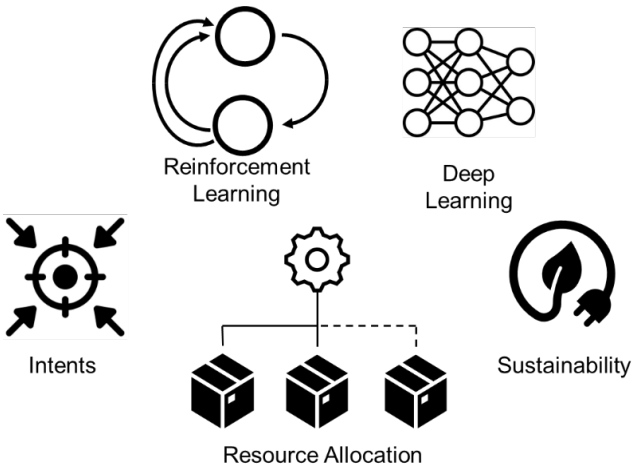


Figure 3-22: Enabler 7 depiction

3.7.4 SoTA and Beyond SoTA

General aspects of environmental sustainability

As mentioned in the motivation section, environmental sustainability is quickly becoming one of the main goals of telecommunication companies. Many in the industry are taking on the challenge and pulling their entire ecosystem of suppliers, partners, and regulators along with them. Many also expect their customers to start requiring product and services with a smaller footprint while even paying more for them. At the same time, regulations such as the European Green Deal are providing another perspective to the push in that direction.

There are three components that can be considered for environmental sustainability:

- Circularity (sustainable model, process, or economic system focused on re-use and waste elimination) considered for fixed and mobile devices as well as for network equipment.
- Energy use and efficiency need to deal with the expansion of the energy-hungry infrastructure for increased demand services and needs for each one.
- The origin of the energy that fuel the different systems will change with the use of renewable energies.

These three components of an environmental sustainability model should be compared among the different actors, operators, customers and governments for evaluation purposes, and a clear definition of these aspects should be offered by the standardization bodies for a valid comparison. Related to this, the GSMA elaborated a report about Environmental, social and governance (ESG) Metrics [GSM23] which defines four categories, including the environmental one:

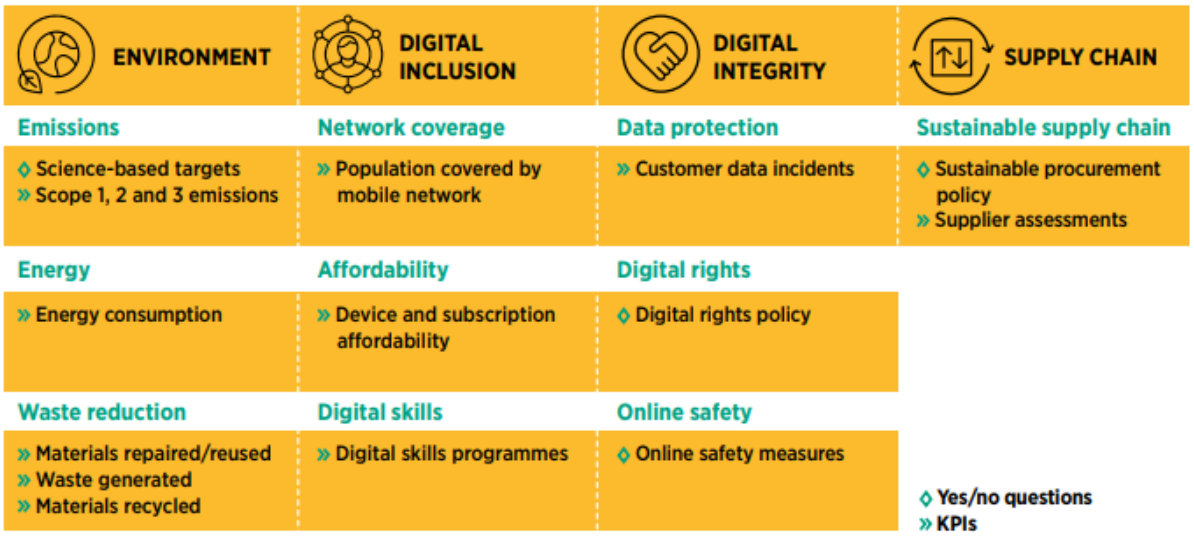


Figure 3-23: Mobile industry KPIs

Attending to environmental aspects, which are the main relation to our work here, we can distinguish the energy related indicators and more concretely, the energy consumption ones:

Energy	Energy consumption	GSMA-ENV-03
	1.3a Total energy consumption i. Total energy consumed (MWh) ii. Total energy consumed (MWh) per 1GB of data	
	1.3b Network energy consumption i. Total network energy consumed (MWh) ii. Total network energy consumed (MWh) per 1GB of data	
	1.3c Network energy mix i. Percentage grid renewable ii. Percentage grid non-renewable iii. Percentage off-grid renewable iv. Percentage off-grid non-renewable	

Figure 3-24: KPIs in the Energy section

The analysis of the competitors and the internal metrics achieved for each company could propose a KPI goal. By examining the practices and strategies adopted by industry rivals, an organization can gain invaluable insights into prevailing standards and potential areas for improvement. Furthermore, a detailed scrutiny of internal metrics, such as energy usage data, operational efficiency, and sustainability initiatives, provides a clear picture of the company's current standing. This self-assessment aids in identifying strengths to leverage and weaknesses to address, thereby facilitating the setting of KPIs. The different levels proposed in the KPIs could be information or constraints to be considered in the optimization processes of the networks and digital services. Further analysis should be performed for scenarios where an operator uses services and resources from a partner. At this stage the sustainability information is something internal for the companies but when a resource from other partners is used, the sustainability information should be included for a complete overview of the values for a specific service. The modification or extension of data models to include information about energy consumption, for example in the capabilities exposure related APIs, will allow the complete overview of the related costs of a service or network reconfiguration, information that should be taken into consideration in the automation closed loops.

On the other side, we also need to consider the relation between the programmability of the network and the agile reconfiguration processes including AI/ML techniques. Related to this, the ITU has a specific group on Environmental Efficiency for AI and other Emerging Technologies [AIET23] which evaluates [TRW23] the energy demand for data analytics considering data collection, data storage, data transmission and data processing:

$$E_{data} = \int_0^T DC_i(t)dt + \int_0^T DT_{IoT(i)}(t)dt + \int_0^T DT_{mobile(i)}(t)dt + \int_0^T DT_{cable(i)}(t)dt + \int_0^T DT_{dataCenters(i)}(t)dt + \int_0^T DT_{cloud(i)}(t)dt + \int_0^T DP_i(t)dt \quad (15)$$

Equation 3-1 Energy demand

The previous formula (Equation 3-1) shows how the Energy demand is obtained as the addition of the energy consumed for data collection (DT), data processing (DP) and data transmission (DT), considering IoT, mobile, cable, datacenters and cloud networks.

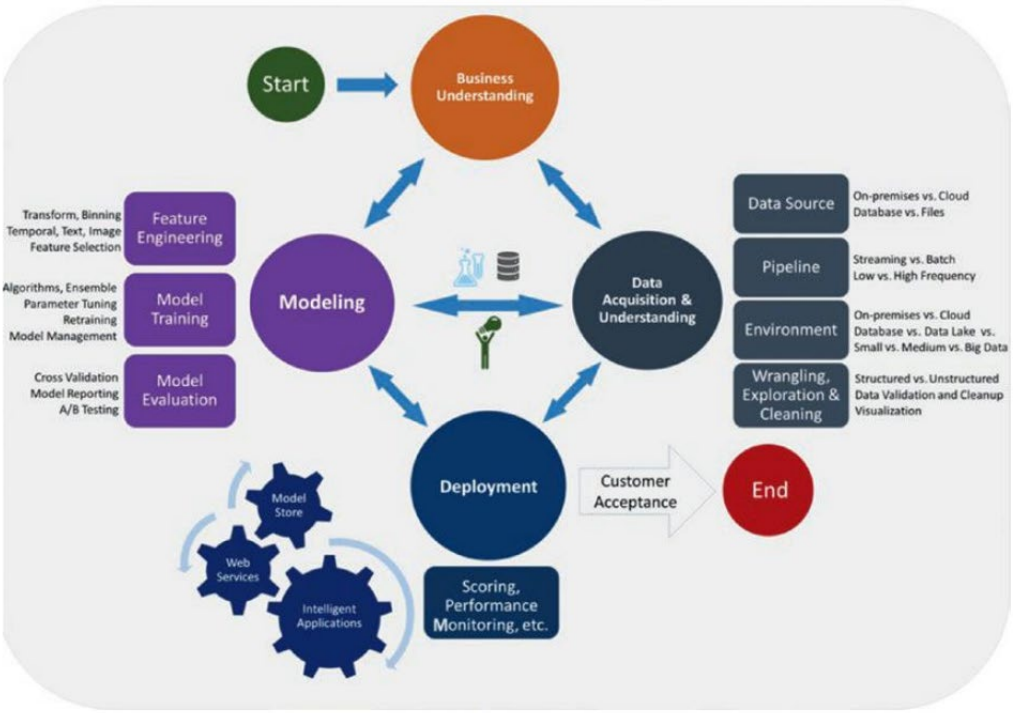


Figure 3-25: Data science lifecycle [Eta19]

The previous figure shows the overall process for big data used by AI systems including: understand business problem, ingest data, modelling and deployment.

The energy used in the management actions need to be evaluated together with the cost of getting the service, and a balance among a good analytic model must be reached with an adequate ratio of energy consumption per service or its use. The training process could consume a lot of energy but sometimes this process is outsourced, making the sustainability evaluation complex.

Intent-based management with energy aspects

The ultimate aim of network automation is to provide proactive and Zero-touch network and service management (ZSM) that needs no human involvement. Unlike conventional script-based automation, zero-touch automation covers multiple domains such as Radio Access Networks, core and cloud, providing end-to-end automation. It may also cover numerous operators. At the heart of ZSM lie intelligent and cognitive closed loops capable of learning optimal behaviour through interaction with the network while performing multiple functions. Modern ML methods allow closed loops to adapt to the changing conditions dynamically and thus, help maintain a robust and flexible network. In the SoTA, closed loop systems have been well-understood and established, and those systems can find be applied to various domains. It is common to assume that a closed loop structure consists of different modules where each module has different responsibility. This responsibility sharing can enable a closed loop to become more scalable and practically feasible. The use of closed loop in the telecommunication domain is not new but the emergence of this usage was not critical as the complexity of previous generation could still be managed by rule-based polices. However, rule-based automation cannot manage complex network requirements, due to which closed loop automation has emerged again.

Also, intent-based policies are the foundation of the new era of autonomous networking. Intents are declarative policies that allow network operators to provide the requirements at a high level to simplify and improve the agility of the operations. This simplification is achieved by the declarative nature of intents given in high-level objectives within an abstract form, and they are translated into requirements across layers of management and across multiple domains that compose the end-to-end (E2E) service. This new paradigm of flexibility will be governed and orchestrated by autonomous AI/ML-based decision-making execution units. For example, the expectation on energy consumption can be specified within an intent, and a closed loop mechanism can be utilized for QoS/QoE control with the consideration the energy requirement.

Due to its importance in the forthcoming 6G networks, international standardization and academic and industry research organizations have several ongoing activities to develop an architecture for autonomous networks. Among these, ETSI ZSM [zsm-002], TM Forum's Zero-touch Orchestration, Operations, and Management (ZOOM) [TM-Zoom14], ETSI ENI (Experiential Network Intelligence) ISG [ETSI-ENI19], and ITU Focus Group on Autonomous Networks (FG-AN) [ITU-AN20] bring together many different market players to achieve the goal of autonomous networks. Those organizations and studies are mostly focused on intent definition, closed loop architecture and how to utilize AI/ML methods from an architectural perspective. When it comes to implementation with the concern of energy consumption there can be different algorithmic approaches and also different types of ML models. More investigation will be needed in the design and selection of those different approaches. In addition, it can be more beneficial to take into account the energy concern with an end-to-end networking perspective which is currently lacking. As an example, realizing an intent with end-to-end energy consumption requirements may need different actions from different network domains each with their own contribution to the overall network energy consumption. This necessitates a holistic view covering different domains to the energy consumption problem. Also, the operators face challenges in determining energy-saving actions due to the numerous possible combinations. For instance, some energy-saving actions may clash with each other, or conflict with other activities such as network optimization actions. Additionally, evaluating the impact of energy-saving actions on service experience beforehand, such as UL/DL RAN throughput and latency, is not a straightforward task. This complexity makes it difficult to strike a balance between achieving energy savings and maintaining satisfactory service experience, as some energy-saving actions may actually degrade the overall service quality. This conflict is expected to be resolved autonomously without human intervention.

Energy efficient resource allocation

Allocating resources to UPFs is usually seen as a UPF placement problem and thus it is mostly performed in a static manner at service deployment, while user and application mobility, as well as traffic volume might change over time. Recently, however, there has been some attention on reacting to some of these dynamics and perform continuous adaptations. For example, the authors of [LCA+22] developed a heuristic algorithm called Dynamic Priority and Cautious UPF Placement and Chaining Reconfiguration (DPC-UPCR) which efficiently remaps Service Function Chaining (SFC) requests (SFCRs) and readjusts UPF placements in online scenarios. However, they explicitly exclude the service functions in the Data Network (DN) of the edge UPF and assume that the service function will use a different resource pool - "our model does not include the destination nodes of an SFCR since we assume that they are DNs co-located with A-UPFs". This assumption makes the solution not suitable for use in common cloud deployments because it requires different resource pools. Another limitation is that the resource allocation for a Virtual Network Function (VNF) is in number of units (e.g. vCPUs) usually only linearly related to the number of sessions (SFCs) each VNF needs to process. These units stack up to a node cap C_c. This implies reserving resource units to VNFs, which diminishes resource sharing capabilities, or otherwise ignores the fact that resource sharing comes with a price of multiplexing, which increases when the number of VNFs increase because more VNFs have to be scheduled on the same machine.

The dynamic changes mentioned previously may be tackled by adjustments in the Edge UPF configuration, which as an isolated problem can be solved by scaling, as described in the first part of [SHR20]. However, this is done in a narrow context without taking into account the location and taking only (prediction of) traffic input to the deployed UPF. Although this type of scaling is shown to work well in many cases due to its simplicity, it may also lead to inefficient distribution of resources since the scaling is purely based on incoming UPF traffic and omits any other constraints like used resources on the host or other hosts around. Thus, a better solution is required to trigger resource allocation changes and add more context to the scaling since it may involve allocating resources at another location, possibly in a proactive manner to further increase efficiency. This may still of course include predicting future user and/or application mobility or changes in traffic volume.

The second part of [SHR20] also deals with the problem of SFC placement (architecture illustrated on Figure 3-26), similar to the previous one, by using Integer Linear Programming (ILP) and heuristics. The main differences are the inclusion of radio resources as Physical Resource Blocks (PRBs) assigned to an SFC, as well as the inclusion of a virtualized MEC application function (VMAF) at the end of the service chain, albeit only as another VxF with no distinction to VNFs which forward packets. The resource requirements for both are expressed in number of Central processing unit (CPU) units. This assumption also omits the fact that there

are multiple types of resources (Network Interface Card – NIC -, CPU, Mem, etc) and VNFs that forward packets are heavy on NIC resources and less on CPU, while for VMAF(s) the opposite is valid. Hence, it is necessary to add this complexity to the problem and make it more applicable for real cloud environments.

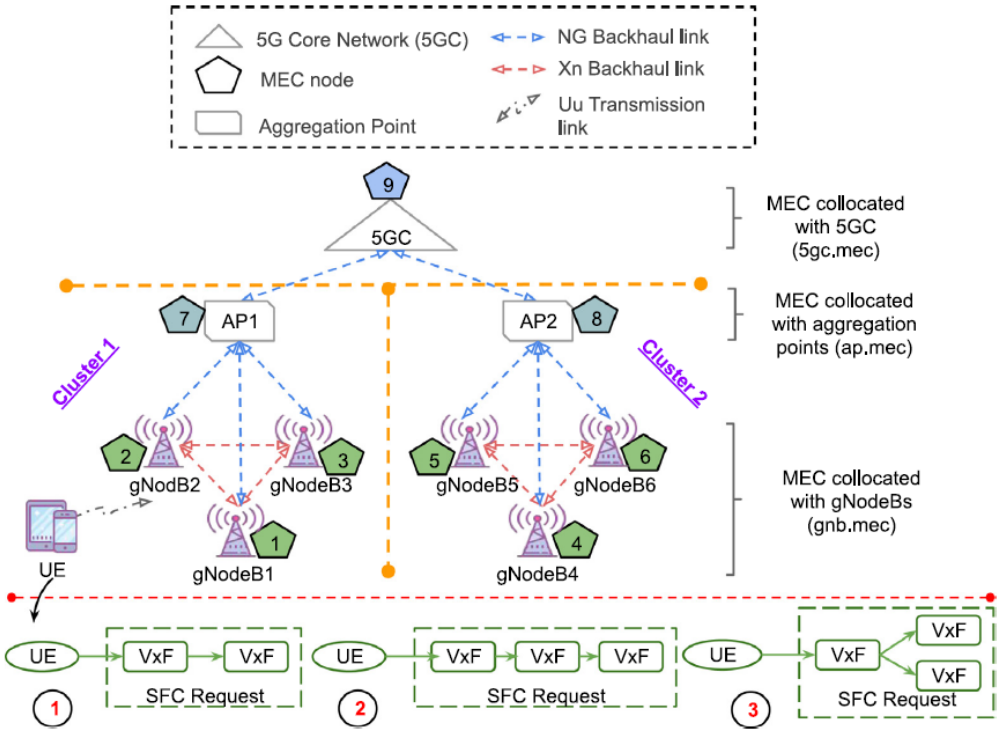


Figure 3-26: The infrastructure and SFC concepts analysed in [SHR20]

Finally, most solutions in the literature are reactive and use ILP or heuristic-based algorithms (including the two cited here) that are computationally demanding, in particular when requirement complexity increases. In order to quickly enforce necessary changes, faster decision-making is required. Introducing a more realistic multi-resource sharing will increase the complexity dramatically, so ML-based solutions might be beneficial for such cases.

The main goal of networks is to transport information related to services, so it is of value to also consider the problem of service workloads placement perhaps together with network function workloads. To take this into account, we also consider the problem of optimal placement of computational workloads (e.g., services, AI workloads, tasks), to the available compute nodes of the system (e.g., robotic units, servers) by ensuring minimum energy utilisation, and maximum trustworthiness. This mechanism will be integrated with the various components of PoC#A.1 (see section 4.1) to provide maximum management intelligence and ensure sustainability.

Various optimisation placement algorithms have been developed for energy efficient 6G networks lately. In [LLY+23], the problem of deploying AI tasks on a network of edge computing nodes that can communicate with a cloud server is studied. This paper aims to minimize the total computing time and energy consumption of all task nodes and maximize the inference accuracy of AI tasks by jointly optimizing the resource allocation and computing offloading decision of each node. The paper proposes an alternating direction multiplier method-based algorithm that can efficiently solve this problem by decomposing it into easy-to-handle subproblems. In [MGK+22], the joint problem of real-time user association, traffic routing and VNF placement is studied towards maximizing the mobile network's energy efficiency and user acceptance ratio. This paper proposes an energy-efficient real-time heuristic called ONE that leverages online convex optimization techniques to solve the problem in a distributed manner. In [TST+22], a dynamic practical model that enables the efficient management of power resources in IoT networks is proposed. It presents two user-scheduling algorithms, namely, minimum distance scheduling (MDS) and maximum channel gain scheduling (MCS) and compares their performance with different power-allocation methods and precoding schemes.

However, the studies cited above do not take into account the possibility of having robotic nodes among all of the compute nodes. In such cases, the various robots' capabilities (e.g., camera, arm, wheels) should be additionally considered at the extreme-edge domain as part of the compute continuum in the resource allocation problems as well as the battery level of these nodes. An initial study on this topic has been performed in [HEX23-D73] and will be extended here. There are also limited studies of (re)allocating resources for the various functions of the system with the aim of both minimising the power consumption and maximising the trustworthiness of the system by placing the functions and services on trustworthy nodes. This is also something planned to be studied.

Sustainable AI/ML

On one hand, AI/ML can be used to incorporate prediction capabilities when implementing energy-aware network management in order to achieve further energy savings by means of taking proactive decisions rather than reactive ones. Furthermore, it is nowadays well expected that an AI-native approach can bring great benefits into telco networks in coping with complexity management and OPEX reduction in general. Native AI, as defined a "*having intrinsic trustworthy AI capabilities, where AI is a natural part of the functionality, in terms of design, deployment, operation, and maintenance*" [IJR+23] will really mean going for a data-driven architecture as it is being designed in [HEX223-D32] in which AI/ML would be applied at all layers and through all compute and network infrastructure domains [HEX22-D62]. On the other hand, considering the environmental aspect of the sustainability target/KVI, we can state that energy consumption reduction, and more importantly the carbon footprint associated to it, will be a great challenge not only due to the more a distributed and pervasive approach of 6G networks and associated increase of computing due to the network itself, but also due to AI/ML contributing to increase the energy consumption if we do not perform countermeasures to mitigate it [6GR23-D21].

As it is documented in [LLS+19] the most consuming ML phase is the training process when the significant computing process takes place. Once the ML models are trained there is no clear difference in incorporating ML as part of the decisions loop or not from the energy consumption perspective. We find some numbers in this regard:

- NVIDIA trained MegatronLM (smaller than GPT-3) over 9 days consuming ~27648 kWh, almost 3 times the average consumption of US homes/year [Lab21]
- University of Massachusetts Amherst: "training a single AI model can emit as much carbon as five cars in their lifetimes." [SGC19]

Therefore, the first thought to consider is that it is not always the best approach to use the most powerful ML model if we consider the environmental sustainability KPI. There can be cases in which it is not worth to use it versus a more traditional algorithm when traditional algorithms may be enough to reach an acceptable performance according to the established SLA. So the optimal balance between reached performance and energy consumption should be found.

Second, it is not only about the energy-efficiency or energy consumption reduction, but also about reducing the actual carbon footprint and CO2 emissions. These reductions are what makes us contribute positively to the environmental sustainability KVI, so the use of green energy should be prioritized.

As stated in [Pat22] there are some main aspects that need to be considered if the reduction of the ML carbon footprint needs to be prioritized:

1) Selecting efficient ML model architectures, such as sparse models (this measure can reduce energy consumption by a factor between 3x–10x)
2) Using processors and systems optimized for ML training, versus general-purpose processors, (reduction by 2x–5x)
3) Computing in the Cloud rather than on premise reduces energy usage and therefore emissions (reduction by 1.4x–2x).
4) Location with the cleanest energy (reduction by 5x–10x)

With the current envisioned Enabler #7 in Hexa-X-II we will go a step further in the environmental sustainability target when incorporating ML by means of designing and implementing the following features beyond the SoTA to further improve carbon footprint reduction in AI-based network management:

- Improved energy observability as part of the MLOps cycle to be aware and estimate training energy consumption in a real time basis.
- Estimation of the expected energy saving from concrete algorithms that will be compared with the actual expected increase of energy consumption due to the associated model training, in order to take final decision whether implementing them or not, and in case of positive decision, evaluate how, when and where.
- Design and implementation of dynamic mechanisms that will be moving the ML models training phase in time and location to greener energy sources during runtime for carbon footprint optimization.

Multi-domain federated learning

Nowadays, integrating different sources of data for training ML models is a hot topic, and has been applied in multiple fields such as healthcare [MWW+17], networking [RRA+21] and industry [PUL22]. Paradigms as Federated Learning (FL) or Distributed Learning (DL) play a key role in the integration, and have been widely studied in the literature [MCM+21][BHV+23]. In theory, the paradigms are designed to handle scenarios in which sources belong to different administrative domains (problem already considered by standards [mec-040]). In this case the challenges of the multi-domain scenario such as data privacy and synchronization arise. This problem is also related to resource federation but approaches in the literature do not seek to optimize the ML model training process itself [IMT+23]. Due to the importance of ML in the forthcoming 6G networks it is crucial to develop a multi-domain system for federating raw data, weights or models used in ML training that jointly optimizes both the training process and resources used (w.r.t. energy and compute).

3.7.5 Identification of possible components and interfaces

Intent-based management with energy aspects

We envision that intent and closed loops are two main elements of this solution. Closed loop can be composed of different modules - monitoring, analysis, decision and execution (see Enabler #10 description in 3.10.3). Communication between them is performed through specified interfaces. It is also possible that some modules of a closed loop can be in an interaction with other networking functions such as Management Data Analytics Function (MDAF) for the usage of AI/ML model.

An analytics module can provide insights in the energy performance, and the energy optimization module can recommend different actions to optimize the energy consumption. However, those recommended actions may need to be predicted rather than enforced on the real network directly so that the impact on other services can be measured. After the final decision, closed loop action can be applied to real network.

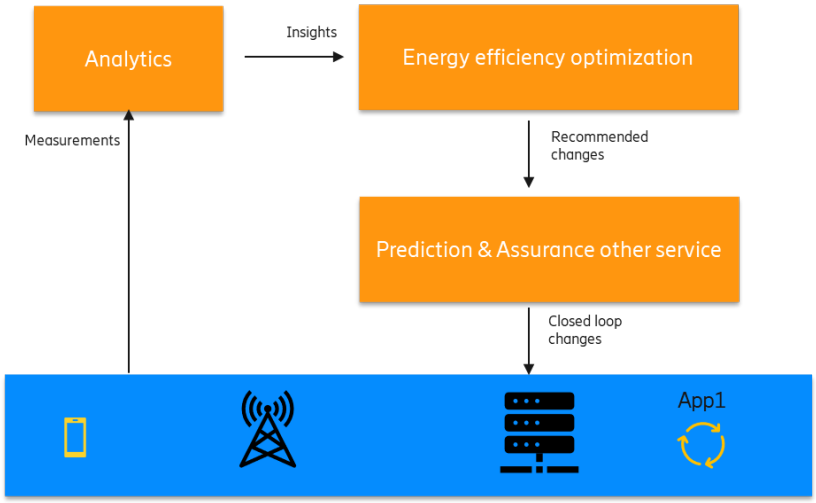


Figure 3-27: Energy Optimization and Assurance with a closed loop structure

Energy efficient resource allocation

The optimisation algorithm/mechanism planned to be developed for workload placement will be closely connected with a system monitoring component which will monitor the system and trigger the algorithm for (re)allocating the functionality when there is a need (e.g., decreased performance, trust risk, etc). Also, it should be closely connected with a service and network registry component to receive the data related to compute nodes status and computational workloads' requirements. The optimisation will be towards energy efficiency with the use of the energy efficiency optimisation module developed within this enabler and towards maximum trustworthiness with the utilisation of the trust manager component from Enabler 4. Finally, the output of this component should be handled by an orchestrator for enforcing the proposed (re)allocations to the system.

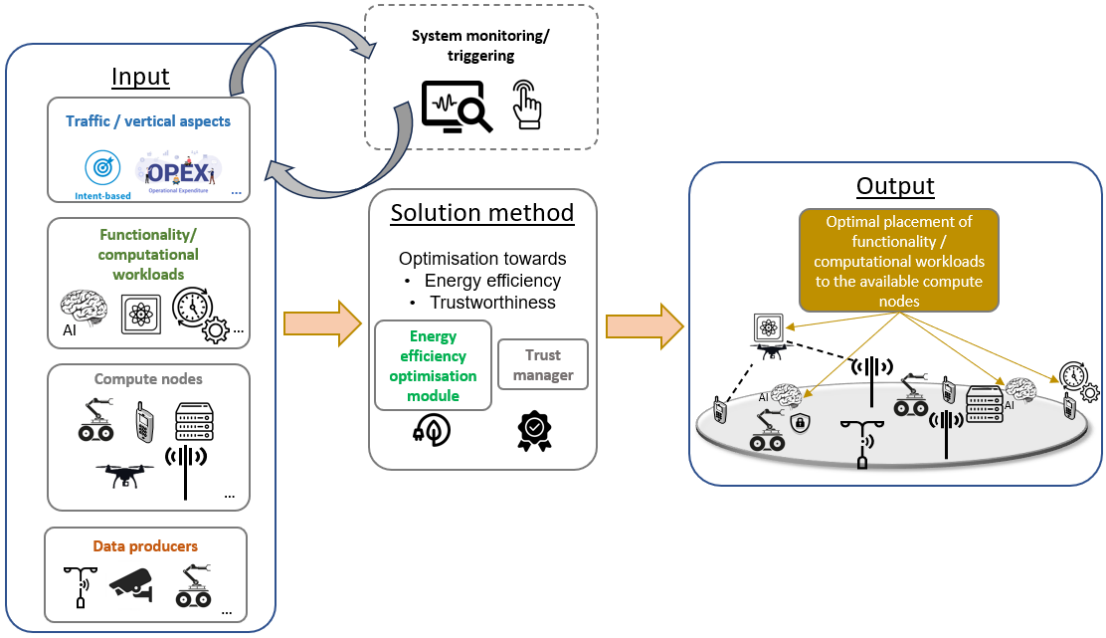


Figure 3-28: Optimisation problem of placing functionality to the available compute nodes towards energy efficiency and trustworthiness, including Enabler 4

Sustainable AI/ML

As explained previously we envision to incorporate extra energy monitoring capabilities in real time as part of the MLOps cycle, which is represented in Figure 3-29 This will allow us to continuously compare actual ML training energy consumption versus energy saving brought by each ML algorithm. Besides, based on a database regarding energy available resources the objective will be to move ML training to available greener energy sources or postponing when possible, so carbon footprint will be optimized.

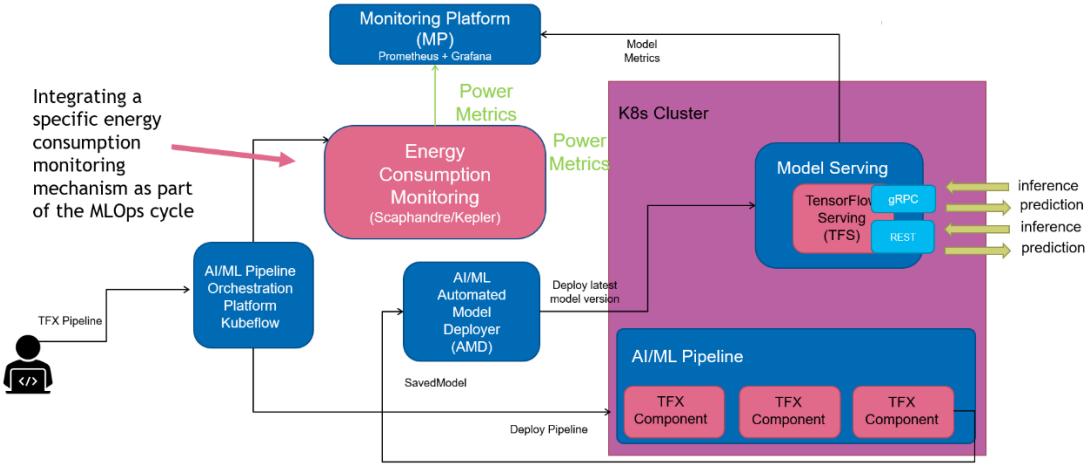


Figure 3-29: MLOps cycle including extra energy metrics

Multi-domain federated learning

Figure 3-30 illustrates an example architecture depicting the interactions between the training and federating orchestrators to fulfil the joint optimization of energy consumption and that of compute resources.

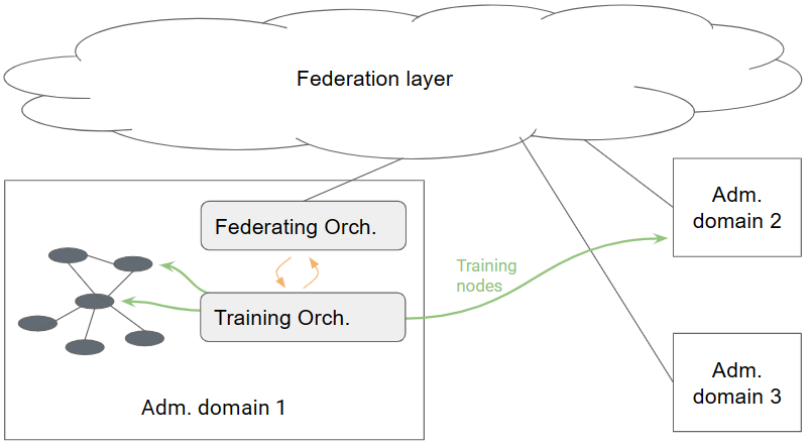


Figure 3-30: Training and federating orchestrators

3.7.6 Relationship with other Enablers

This enabler's focus is the development of AI/ML algorithms which use data from monitoring systems. These algorithms are intended as the implementation of core intelligence in closed management loops developed in **Enabler 10 and 11**. In this sense, this enabler is related to these two enablers but also to aspects contributing to the enabler E3.1.4 Intent-based Management described in [HEX223-D3.1]. In [HEX223-D2.2] intent interfaces will be identified, and this enabler will also need these interfaces.

3.8 Enabler 8: Trustworthy AI/ML-based control

3.8.1 Motivation

The goal of trustworthy AI/ML is to ensure that AI/ML systems are developed and used in a way that is transparent, accountable, reliable, safe and ethical. There are different key aspects that are associated with trustworthy AI/ML [ABB+20], including: security, privacy, and explainability. The security aspect of trustworthy AI/ML considers the design of the AI/ML system to be robust, reliable, and safe for the individual and society. The privacy aspect of trusted AI/ML aims to develop the AI/ML system in accordance with data protection laws and regulations to respect the privacy of the individual and protect personal data disclosure. The third key aspect associated with trustworthiness is explainability. ML models which handle complicated computational tasks with essentially no human participation are inherently complex, black boxes. This has increased the need for accountability and raised concerns about the decision-making processes used by AI/ML systems. Explainability aims to enhance transparency of black-box ML models in order to enable humans to understand the decisions made by the AI/ML systems.

In case of network management and orchestration, AI/ML can be used in several areas such as network optimization, fault detection, and service orchestration automation. ETSI Industry Specification Group (ISG) defines different use cases that cover four categories, including Infrastructure Management, Network Operations, Service Orchestration and Management, and Network Assurance, showing the potential benefits of Enhanced Network Infrastructure (ENI) and the use of AI/ML in networks [WFC+18]. Due to the significant role of AI/ML on management and orchestration of the network in 6G, attacks against AI/ML system can affect any system that relies on these technologies. In ML, the data can be used in the training and inference phases of AI/ML systems. Studies show that other than data, the model parameters also can provide information about the processed data, making them just as valuable as the data themselves [SKK+22]. There are two kinds of attacks against ML, adversarial attacks and privacy attacks. Adversarial attacks, including poisoning and evasion attacks, refer to a set of techniques used to deceive ML algorithms by adding purposefully crafted perturbations to the input data. Although these disturbances are intended to be invisible to humans, they may lead the model to predict or categorize data incorrectly [CAD+18]. Privacy attacks include

membership inference and model inversion attacks, which refers to those techniques that aim to compromise the confidentiality of sensitive data used to train ML models [PMS+18]. In addition, since it is difficult to test AI tools for their correctness and explain their behaviours, it is crucial for network administrators and operators to utilize explainability tools to understand, interpret, and trust the decisions and actions made by AI systems. In this regard, trustworthy AI/ML systems should make their decision-making transparent to human users and provide clear and concise explanations of their decisions tailored to the expertise of the target users.

3.8.2 Objectives

The objectives of the trustworthy AI/ML-based control enabler is to provide more robust models by protecting against adversarial attacks, minimizing leakage of personal and sensitive data, and providing clear and concise explanations and justifications of the AI's reasoning or decision-making process.

3.8.3 Description of the solution

With the increased adoption of AI/ML in network management and orchestration to automate and optimize the operation of networks, trustworthy AI/ML will gain more attention to provide more robust, reliable, and trustable networks. Trustworthy AI/ML includes several concepts where we will mostly focus on security, privacy, and explainabilty concepts.

Trustworthy AI/ML security focuses on the concept of ensuring the security, integrity, and reliability of AI/ML systems throughout their lifecycle which include data security, model security, adversarial attack (such as poisoning and evasion attacks) resilience, etc. There are different mitigation studies to prevent security risks such as model regularization, which prevents overfitting and improves model robustness, input sanitization and pre-processing, defensive distillation [PMW+16], adversarial training, regular model updates, etc. which can be used to increase the robustness of ML models.

Trustworthy AI/ML privacy focuses on the concept of ensuring the privacy and protection of sensitive information in AI/ML systems. It consists of practices that prioritize user privacy, data protection, and compliance with privacy regulations. An example of privacy risks to AI/ML systems includes privacy attacks such as membership inference attack and model extraction attack. To mitigate privacy risks against AI/ML, several techniques can be used, including data minimization, anonymization and De-Identification, differential Privacy (DP), homomorphic encryption (HE), secure multi-party computation (MPC), federated learning, etc.

To improve the **explainability aspect of AI/ML** methods for smart network management, solutions will be taken from the active field of study on explainable artificial intelligence (XAI) which has evolved in order to make the behaviour and predictions of AI/ML systems intelligible to people and offer transparent decision-making processes of complex AI systems, [VFM99], [GA+19]. The goal will be to find answers to the question of "why is this happening?" by enabling a human to understand the structure and behaviour of the model by finding meaning between the attributions of the input data and the model outputs. There are two main types of explainability solutions, post-hoc and intrinsic [PV20]. Post-hoc explainability can be applied to an existing black-box ML model to extract explanations from it and can be achieved with several techniques, including feature attribution [LL17], counterfactuals explanations [DMB+20], and anchors [RSG+18]. Intrinsic explainability, instead, consists of designing ML models which are intrinsically interpretable. This idea can be realized in many ways, such as limiting the complexity of the model [LSZ+19], decomposing the model structure [JKF+19], and adding constraints during the training [AJ+18].

The trustworthy AI/ML-based control enabler will take advantage of the solutions illustrated in Figure 3-31 to enhance security, privacy, and explainability of AI/ML, and will make an effort to integrate following primary areas of work into the project:

- Show the vulnerability of AI/ML model to adversarial attacks and feasibility of such attacks by carrying out experiments in intent-based cognitive closed loop management use case and enhance the security of the system using mentioned solutions.
- Investigate the vulnerability of AI/ML model to privacy risks by carrying out experiments in intent-based cognitive closed loop management use case and enhance the privacy of the system using mentioned solutions.

- Utilizing explainability tools to test AI/ML models for their correctness and to explain their behaviours to network administrators and operators. This is important to understand, interpret, and trust the decisions and actions made by AI systems.

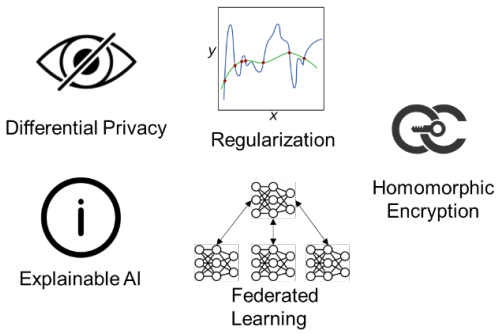


Figure 3-31: Enabler 8 depiction

3.8.4 SoTA and Beyond SoTA

Leveraging AI and ML techniques, AI-driven network management and orchestration can enhance the automation and efficiency of the network operations. A trustworthy AI/ML-based control ensures high level of reliability, security, and efficiency for network operations and services to control and optimize network resources. In the literature it has been established that ML techniques are vulnerable to several attacks [BNS+06] that target both the training phase (i.e., poisoning attacks) and inference phase (i.e., evasion attacks). Introducing carefully crafted perturbations to training or test samples can cause integrity, availability, or privacy violation. Thus, if security and privacy concerns respecting to AI/ML are not addressed, enthusiasm to use AI/ML may decrease. In recent years, the vulnerability, and risks of AI-driven systems in several domains such as closed loop automation, intent-based networking, AI-driven zero-touch network, etc. has been the subject of research studies. As an example, in [BT+2020], they mentioned that in zero-touch service management, the end-to-end intelligence service which drive the closed loop, can encompass both making decisions and initiating their execution (e.g., decisions to optimize the E2E service) as well as making forecasts and suggestions related to a given service (e.g., predict service demand). Since the decision-making process is dependent on information obtained from domain data collection services, an attacker may design inputs that would cause the ML model used by the E2E intelligence services to make incorrect predictions and decisions; thereby harming performance and resulting in financial loss. Figure 3-32 shows an illustrative example of adversarial attacks on VNF auto-scaling scenario where scaling decisions are made by the ML model used by domain intelligence. When attacker manipulates metric data, he/she can fool the ML model to take scale-out decision instead of performing scale-in operation, which will result in adding new VNF and increasing the costs. In addition, they underlined the importance of collaboration and data sharing among various Management Domains (MDs) to increase accuracy and speed up the learning of ML models utilized by the various MDs. Concerns about trust and privacy arise when shared learning is empowered.

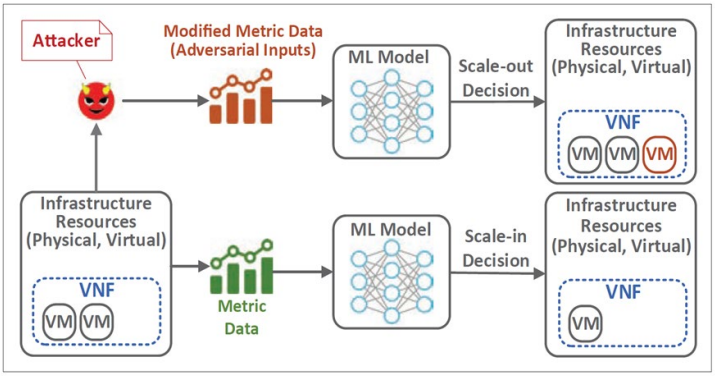


Figure 3-32: Adversarial attack illustrative example from [BT+2020]

In another study [UQA+2018], they show the effect of adversarial attacks on cognitive self-organizing network. Evasion attacks against convolutional neural networks are designed to show how much a malware classifier can be evaded. They show that when adversarial training is applied as defence mechanism, the malware classifier is more successful on classifying adversarial examples. In [BK+20], they mentioned that monitoring and predicting the performances of the running network slices or their key performance indicators are crucial to achieve zero touch management of network slices in 5G. Not all KPIs, but those related to service can arise privacy concerns as they can disclose critical information on the performance of services. Thus, as it is illustrated in Figure 3-33, to tackle with privacy issue, an ML technique, known as federated learning, is used where each slice trains its own local model and sends them to slice orchestrator without sharing local dataset. The slice orchestrator aggregates the result and generates the global model which is shared between slices. Thus, using FL, the data in each network slice is kept local, and only local learning models are aggregated.

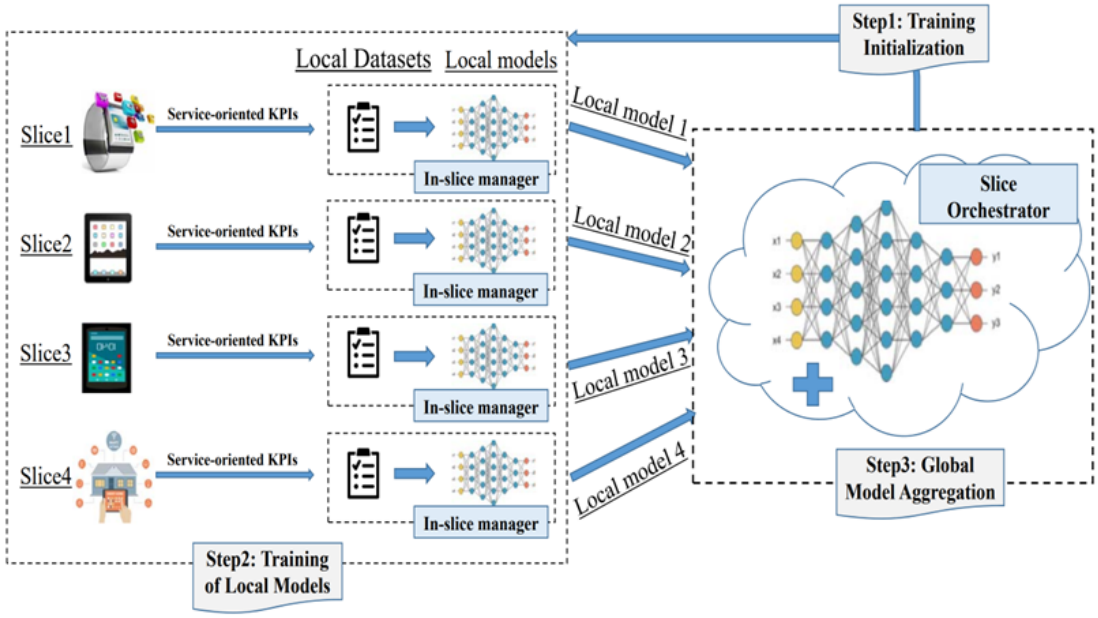


Figure 3-33: Training Process of Federated Learning from [UQA+2018]

In [RCV22], they elaborate on challenges of managing and orchestrating the massive number of slices in 6G network using AI-driven zero-touch management and orchestration. In their work, to ensure the transparency among the interacting actors in the slicing ecosystem such as tenants and operators, the importance of explainable AI (XAI) tool is highlighted. In the literature, few works have applied explainability techniques to telecom use cases with the goal of enhancing transparency and interpretability. [TIB+20] considered the problem of identifying the root cause of SLA violation predictions with ML in 5G networks. Several feature attribution methods have been applied and compared (e.g., SHAP, LIME, Eli5) and the experimental results found SHAP to be the most accurate. [TIF22] designed a new eXplainable Reinforcement Learning (XRL) method, Both-End Explanations for Reinforcement Learning, to provide explanations that correlate inputs and decomposed output of the model. Their results empirically showed that the extracted explanations can not only identify whether the model is biased toward certain input features but also allow to correct and remove the bias by applying carefully crafted weights in the output layer.

From SoTA we can observe that the security and privacy risks such as adversarial attacks and disclosure of sensitive data in AI-based network management and orchestration are possible. This fact poses a danger as AI is expected to play a key role in the 6G communication systems. According to prior studies, adversarial attacks with optimized perturbations can impair the functionality of a telecommunication networks or services. Therefore, effective defensive strategies are needed to counteract the consequences of such attack threats. For us to preserve the trustworthiness of future telecom networks and services, we need to pinpoint all conceivable AI-driven management tasks and use cases. The probable shortcomings of these AI systems must then be evaluated, and we should concentrate our efforts on providing the required level of robustness. In line with our goal, we plan to consider intent-based cognitive closed loop management system as a use case to show the vulnerability of AI/ML model to adversarial attacks, investigate privacy risks, utilize explainability to interpret

the decisions made by AI/ML, and enhance the security and privacy of the system using mentioned solutions. On transparency, the literature highlighted the need of more effort in designing and evaluating reliable explainability tools for telecom. For example, no existing methods can explain a sequence of decisions, but only individual predictions. Furthermore, there is still a lack of techniques that are capable of tailoring the explanations to the technical expertise of the target user.

3.8.5 Identification of possible components and interfaces

In Figure 3-34, you can see an example of interaction between a network management and orchestration and trustworthy AI/ML function in the management plane. The trustworthy AI/ML function includes two sub-components which are "Privacy & Security Guard" and "XAI". The knowledge data, including information on the managed environment as well as domain models and expert knowledge, will be the inputs to trustworthy AI/ML function. The data needs to be security and privacy processed before it is used in model training. The privacy & security guard component is responsible for processing data and guarantee security and privacy. The processes may include, inserting adversarial examples to training data, regularization to prevent overfitting in data, add noises to data, or any other processes. Then, the processed data can be used during training phase to generate models. To provide transparency, it is required for the model to be explainable for those who use the models. This feature is provided by the XAI component.

Trustworthy AI/ML function can be either integrated inside network management and orchestration or outside of it, according to the scenario in which AI/ML is used. If AI/ML is going to be used to provide security for the network, such as detection of DDoS attacks, then it is better to be integrated separately from network management. But if AI/ML function is responsible for a task inside network management, then it is better to be integrated inside.

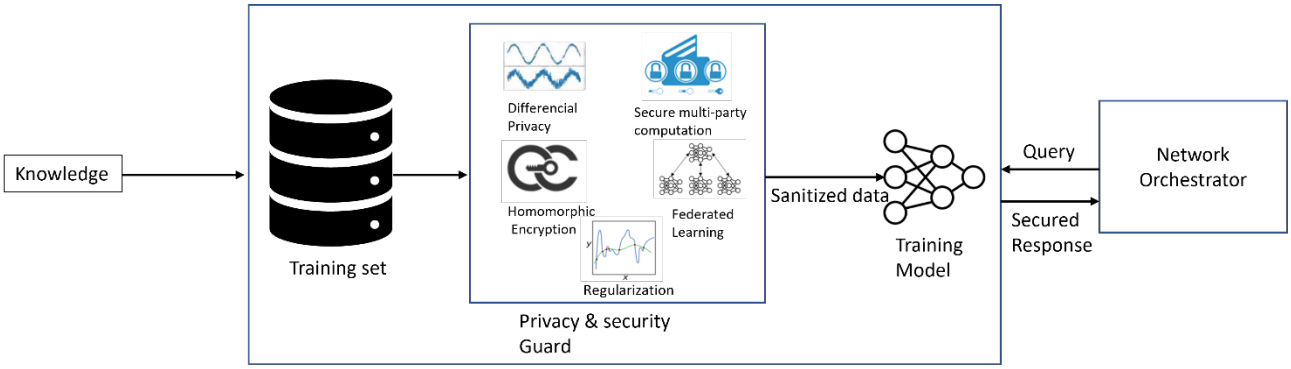


Figure 3-34: Trustworthy AI/ML-based control enabler components and interfaces

3.8.6 Relationship with other Enablers

This enabler can have relation with **Enabler 7** (Sustainable AI/ML-based control) where the aim is to improve the process(es) of network management from automated operations towards full autonomy using AI/ML. Enabler 8 can be used to improve the reliability and trustworthiness of the Enabler 7.

3.9 Enabler 9: Network Digital Twins

3.9.1 Motivation

Efficient training of AI/ML models requires access to high-quality training data-sets comprising a sufficient number of data-points, which is difficult in a network resource management context due to user data privacy concerns. Moreover, Reinforcement Learning models are trained through direct interactions with the environment, by exploring a set of actions, and observing their effect on the state of the environment to use as feedback to improve the decision policy. However, in a network management context, performing exploration of the action space is not possible, due to the possible impact of poorly chosen actions. Therefore, it would be beneficial to train models based on simulated environments, which should have similar characteristics and behaviour as the network infrastructure.

To this end, this enabler will develop a Network Digital Twin that provides a near real-time representation of the network with different degrees of granularity depending on the use case, to mimic the behaviour of the network and provide the necessary feedback for model training. It will then enable the effect of configuration changes to be checked on the digital twin before being applied to the real network, thereby ensuring that no unintended adverse effects are produced. This is especially important given the nature of modern networks comprising layers of virtualised components with complex interactions. Indeed, VNFs and CNFs are expected to share physical resources on top of a flat cloud computing infrastructure, where in some scenarios hardware acceleration or even dedicated hardware is possible. This flat infrastructure has a very high level of location distribution, which increases the complexity of management decisions. The decisions of when and how to deploy and configure NFs are made in the orchestration layer on top of this flat infrastructure. If we assume that this layer implements intelligence with algorithms that require solution space exploration, performing these actions to find the best solution on the actual production infrastructure is not feasible to explore the intended what-if scenarios. There is a need for models which represent the infrastructure well so they can use them to learn how to reach a specific goal, or to test and verify the impact of taken actions in case of uncertainty. The Artificial General Intelligence field heavily uses games and world models in order to achieve this [BCP+16]. An analogous solution is needed for network management, where Digital Twin technologies can generate models that have the same high-level properties as real network deployments (depending on the specific management problem at hand). Having in mind that the current network deployments are being extended towards the far and extreme edge of the network where heterogenous, volatile and mobile network resources are part of network topologies, the creation of such realistic Digital Twin models is very challenging. Network simulators have been used for this, but they sometimes have very detailed implementations which makes them computationally unfeasible or are too simplistic and don't resemble real deployments. Models generated from real-world data can fix these problems.

3.9.2 Objectives

The overall objective of this enabler is to use Digital Twinning technologies to generate network/infrastructure models with the objective of integrating those models in closed management loops as copies of their real counterparts. This would facilitate AI/ML algorithms to be able to estimate certain properties of the networks they are managing without actually enforcing actions on them.

3.9.3 Description of the solution

This enabler will take as input monitoring data from enabler 2 in order to continuously update the state of twins representing VNFs/CNFs. It will also incorporate an AI engine that will facilitate network management by predicting the effect of intended configurations on network performance. Taking the latest SoTA on Graph Neural Networks (GNNs) as a launchpad, the engine will introduce innovations that will enable it to make more accurate predictions on virtual network functions running on shared physical infrastructure.

Indeed, graphs have been used for a very long time to represent networks of all kinds. They are a natural counterpart of networks in the conceptual 'world' since they can have all or most of the characteristics of their 'real' things. Communication networks have also long been modelled and analyzed with methods developed for graph analysis. In recent times, there were some big developments in Graph Neural Networks, which are a class of Neural Networks that can operate on graph structured data. This opens up lots of opportunities for network modelling [AFP+22]. GNNs main idea comes from applying convolutions to graphs, and there are multiple variants like GCNs, GATs, etc. They have been successfully applied to areas like traffic forecasting [RBO21], ETA prediction [DSW+21], molecular chemistry [WJH+21], in social networks to make better recommendations systems [WLT+21] or even for skeleton-based action recognition [PSX+20]. In the networking world, GNNs are applied as promising solutions for mobile traffic forecasting [FEP22], malware detection [BKT+21] or network modelling [FPS+23]. In fact, GNNs have recently also been used as Network Digital Twins of a simulated network [FSP+22]. The GNNs operate by performing a series of message passing operations between nodes in the graph. Each node aggregates information from its neighbours and updates its

own representation based on this aggregated information. As such GNNs are very suitable for creating these virtual representations of the exiting evolving networks and they open a whole new field for applying predictability when solving exiting networking problems. This concept is well aligned with the Internet Engineering Task Force (IETF) view of Network Digital Twins concepts and Reference architecture [ZYD+23]. Figure 3-35 is an illustration of the connections between the 'real' entities in the network and their counterpart models or 'twins' in the virtual world. A more detailed description of interfaces and the overall architecture of the solution will be provided in following deliverables in the project.

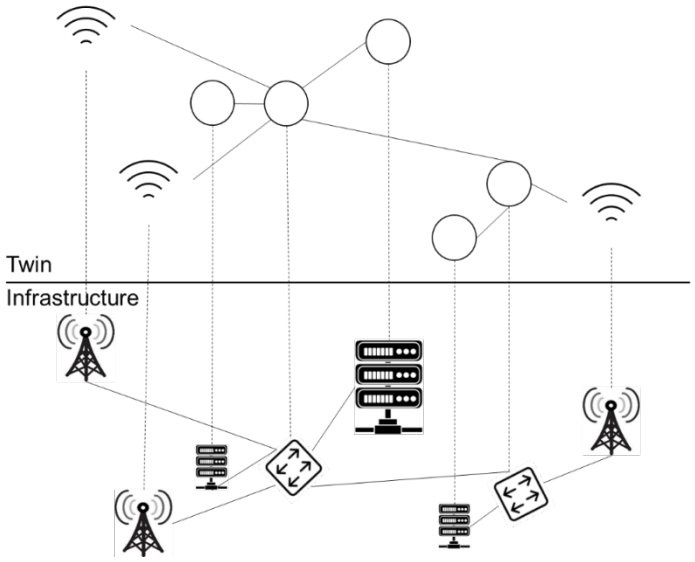


Figure 3-35: Enabler 9 depiction

3.9.4 SoTA and Beyond SoTA

The concept of network digital twins for network management has recently been introduced in literature and is a novel research topic. For example, in [FSP+22] the authors propose a Network Digital Twin (NDT) based on GNNs to determine flow-level QoS metrics depending on the traffic volume, the used topology, and the routing and queuing policies being used. It is similar to an earlier publication [FRS+22] aimed at estimating performance of networks, where the authors used simulated data of several topologies of networks, and created a model which can predict per flow performance KPIs (delay, loss, jitter). The simulated network is represented by queueing at each network node, where flows have different entry and exit with a specified routing configuration (illustrated in Figure 3-36).

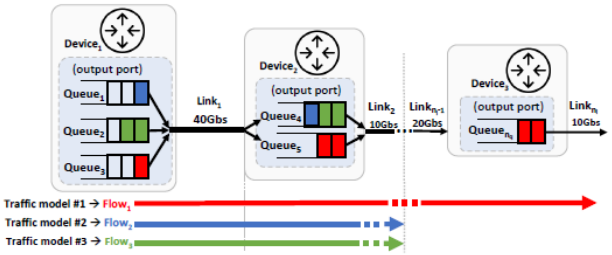


Figure 3-36: Schematic representation of the network model used in [FSP+22]

But representing networks as graphs (and subsequently GNNs) is not the main contribution of the paper, its actually the generalization capabilities to larger networks. The GNN is learned together with two Recurring Neural Networks (RNNs) which produce the latent variables that encode flow (with state of links and queues it traverses as input) and link states (with state of queues on link). The link throughput generalization is done by representing it as a constant and multiplier (i.e. 100Mbps x10) which is learned from the data. The flow length generalization is done by setting a maximum flow length that the RNN encodes, and in case of longer

flows, representing them as chains from the maximum flow lengths. As showed by the results, this allows the GNN to achieve very good performance on unseen topologies and configurations.

Out of the few ideas presented, none of them deal with the aspect of infrastructure resource sharing across multiple virtual networks. Since networks are becoming completely virtual, real models that capture the features of cloud environments and resource sharing are necessary in order for management processes to use them and estimate performance of these virtual networks in different scenarios. This will allow the management processes to reconfigure or even redeploy networks with minimal impact and without hard-resource reservations whenever they are not necessary.

Recently, the suitability of GNNs for network modelling has also been exploited in network digital twins where they are used are for slicing [WWM+20]. However, the GNN-based digital twins in that work do not consider Graph Convolutional Networks (GCN). Recent advances demonstrating that convolution can also be computed for dynamic graphs [MRM20] means that new feature sets can be leveraged to enhance Network Digital Twins beyond the SoTA. GCNs enhance the modelling of graph structures, capturing dependencies and relationships between nodes in complex network configurations. They excel at handling dynamic graphs, accommodating changes in network structures and providing up-to-date insights. GCNs enable the extraction of rich feature representations from graph data, capturing local and global information to enhance forecasting and decision-making. By leveraging GCNs, digital twins can improve prediction accuracy, optimize network operations, and make better-informed decisions.

3.9.5 Identification of possible components and interfaces

Creating a general twin model that implements all possible features of a complete end-to-end network seems like a daunting task. This model would contain many parameters and require lots of processing. Hence, it can be envisioned that the division of network management domains can be applied to these twin models. Management domains will thus be modelled as distinct digital twins with interfaces to other domain twins. Furthermore, it is also possible that within a network management domain there could be separate smaller management processes that can use a twin model of a smaller entity, be it a sub-network, network node, radio propagation environment, etc. In order to cope with this, a specific pipeline is envisioned, similar to standard ML operational pipelines but with a possible different final step in case multiple models are necessary. This final step is the assembly of the final twin to be used in the specific network management problem. Further details of interfaces and interactions envisioned between the various digital twins will be provided in following deliverables in the project.

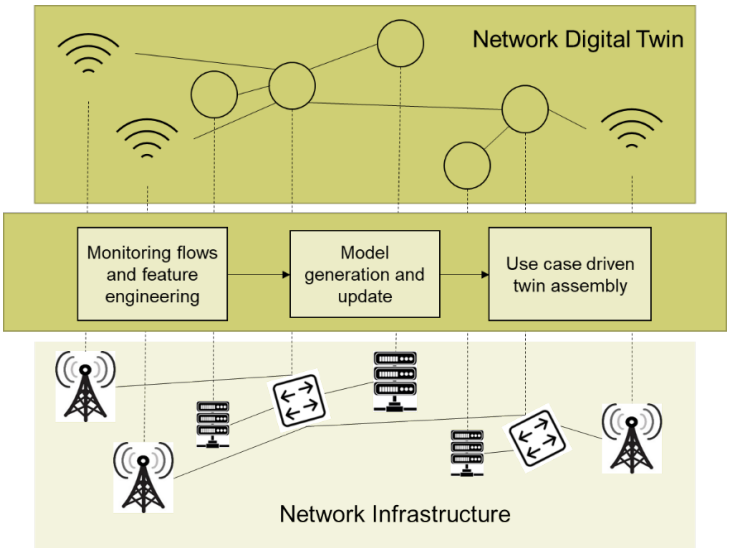


Figure 3-37: Pipeline steps for Network Digital Twin in Smart Network Management

3.9.6 Relationship with other Enablers

As discussed in the enabler's objectives, Network Digital Twins are aimed at improving the governance of closed loop management loops (**Enabler 10**) in the sense that they could be used in well-designed procedures

in order to estimate performance of networks without actually taking the intended management actions. This estimation can then be used in order to automatically check whether to enforce this action.

Furthermore, the concurrent execution of multiple closed loops to automate the continuous optimization of resource allocation, network configuration and service deployments may lead to conflicting decisions. The related execution stage, especially when executed over target infrastructures composed of multiple domains running fully independent closed loops, often leads to unstable conditions or poor performances in the end-to-end connectivity and resource utilization. Techniques based on network digital twins allow to create emulated environments that can realistically replicate the possible reactions of the network to variable sets of concurrent configurations. Such environments can also be used to assess the results of the execution of concurrent closed loops on the global behavior of the network before their actual enforcement. The closed loop coordination procedure **(Enabler 11**) may thus include an intermediate evaluation step, performed on top of network digital twins, to early identify potential conflicts and mitigate the related decisions selecting the proper subset of re-configuration actions to be executed.

3.10 Enabler 10: Zero-touch closed loop governance

3.10.1 Motivation

6G networks are expected to integrate multiple technologies organized in complex, multi-domain infrastructures and highly variable network topologies, combined with various levels of virtualization. The network integrates and interconnects heterogeneous devices, up to the extreme edge, characterized by extreme volatility and a variety of computing capabilities and resource constraints, which need to be managed jointly with the network functions in an end-to-end manner. A variety of services, for final users and verticals, runs on top of the shared infrastructure, requiring different QoS/QoE guarantees and various degrees of isolation, generating highly dynamic network traffic and computational load spanning across the extreme edge, edge, and cloud continuum. This increases the complexity of the network management, requiring the adoption of automation mechanisms that help to reduce the burden of network and service operation, optimizing the usage of the infrastructure resources, while dealing with the dynamicity and variability of traffic, devices, and applications during the service runtime. The concept of zero-touch closed loops (CL) allows to automatically update the resource allocation, the network configuration and/or the service tuning following a data-driven approach that continuously re-optimizes the target infrastructure based on the monitored data and a mix of target criteria and goals. CLs can be formally modelled as a set of functions (monitoring, analysis, decision and execution) which can be virtualized and operated dynamically following different scopes, time scales and domains. Their management needs to be combined with the global orchestration processes of networks, integrating their cycles in the mobile network management system and adopting the same principles used for the other management functions defined by 3GPP.

3.10.2 Objectives

The main objective of CL Governance is to enable the integrated management of CL procedures as part of the mobile network management operations. This objective is addressed by introducing an additional functional entity responsible to coordinate the different stages of various closed loops, allowing the full automation in terms of CL functions provisioning, configuration, runtime update, and termination. The CL itself would significantly reduce (or even remove) any human intervention to adapt configurations and services to the network dynamic conditions, speeding up operations while optimizing resources usage and reducing operational costs, in line with Smart Network Management objective discussed in Section 2.2.5. The Governance of CLs extends this concept, further automating the configuration and provisioning of CLs according to the target objectives, e.g., SLA management, network performance optimization, service resiliency, etc.

3.10.3 Description of the solution

The closed loop (CL) approach, shown in Figure 3-38 is based on a chain of functionalities (called stages) that starts from the monitoring of relevant metrics and KPIs, and proceeds with their analysis, often based on AI/ML techniques. This analysis generates a detailed insight of the target managed entities allowing to take

decisions on corrective actions or re-optimization/re-planning strategies, which are then enforced through the execution of re-configuration or re-allocation commands. This loop can be effectively represented through the four stages of monitoring, analysis, decision, and execution.

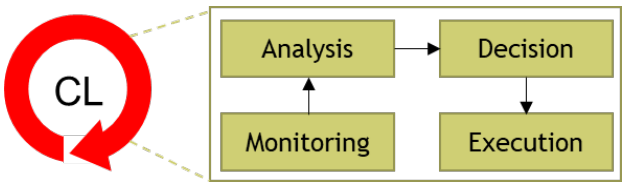


Figure 3-38: Closed loop stages

Hexa-X-II will investigate multi-dimensional CLs for mobile network automation, operating with:

- different time granularities, for real-time reactions and short-term control actions to deal with failures or suddenly underperformance, as well as for medium- and long-term reconfiguration, e.g., to adjust to traffic and service dynamicity and for autonomous re-optimization.
- different domain scopes, applying closed loops at specific technological domains or network segments, e.g., for management of radio resources, migration and scaling of core network functions, re-configuration of transport networks, or addressing TSN/DETNET control flows; closed loops can be also applied at different resource domains across the extreme edge, edge and cloud continuum, e.g., applying the CL mechanisms to Smart VIM (Virtual Infrastructure Manager) features.
- different architecture layers, potentially involving different actors, starting from the physical infrastructure (considering both computing and networking resources), and going up to the network operation and service orchestration layers.

The CL Governance enabler will target the automation for the various CL approaches that will be designed and implemented in Hexa-X-II, coordinating the instantiation and configuration of the CL stages taking into account their particular nature and implementation. Another key point is how to enable the interaction of the CL functions with the rest of the network management system and their collaboration with other enablers (e.g., AI/ML, network monitoring and telemetry, transport network programmability, resource control in the continuum, etc.). For example, for the decision stage, reactive, proactive, and predictive models will be applied, depending on the goals of the given CL, as well as its time scaling. The adoption of AI/ML techniques at the analysis stage would allow to automatically and continuously tailor the decisions to the complexity and heterogeneity of the target network and the dynamicity of the upper layer services. In this area, a key point is the access to high-quality training datasets comprising a sufficient and diverse number of data points to enable the training of ML models with the required level of accuracy. Approaches based on Reinforcement Learning would allow to execute the training through direct interactions with the environment, exploring the effects of the actions taken. However, applying this approach over a running network would be unfeasible because of potential degradations. The implementation of a smart network management sandbox (an example of which is illustrated on Figure 3-39, used for ML model optimization as described in [VSD+21]) integrated in a CL management would allow to validate the decisions before committing the execution stage, relying on the integration of network digital twins as a solution to facilitate the exploration and validation of intelligent algorithms.

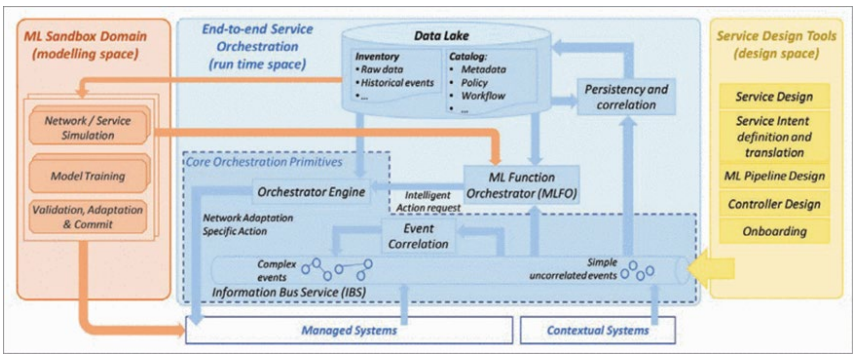


Figure 3-39: Smart network management sandbox for CL decisions based on intelligent algorithms [VSD+21]

Moreover, to enable a successful integration with the overall management system of the mobile network, the CL stages need to be implemented through network functions (NFs) fitting the Service Based Architecture model and their management and orchestration should be handled jointly with the provisioning and operation procedures adopted for the other NFs. As part of this logic, function splitting, resource allocation, energy-efficient placement, migration, and scaling actions of CL functions should be handled as part of the M&O process, potentially in combination with communication service, network slice or slice subnets lifecycle management depending on the scope of the target CL. Finally, a key point is related to the interfaces and capabilities exposed by CL functions, among them and towards the rest of the management system, and their interaction with the other NFs. This is clearly applicable to the collection of monitoring data (e.g., from NWDAF or MDAF) and the execution of configuration commands, exploiting the programmability offered by the various domains and network technologies. However, additional interfaces exposed by the intermediate levels (analysis and decision) are also required to handle the coordination of multiple loops, control their procedures, and share knowledge, models or information produced during the execution of the various stages.

3.10.4 SoTA and Beyond SoTA

Several research work and standardization activities are currently investigating the adoption of CL mechanisms in network management. For example, the **ETSI ENI** architecture [eni-005] introduces the concept of cognitive network management through closed loops supported by AI techniques, context awareness and metadata-driven policies that automate management and operation actions in several areas and layers. Services are adjusted based on users' needs, context conditions and business objectives; mobile networks are automatically provisioned and operated for self-assurance and self-optimization in the areas of network slicing, network service management and resource orchestration [eni-008][eni-010]. In this context, ETSI GR ENI 017 [eni-017] specifically provides an overview of CL architectures for Experiential Network Intelligence.

ETSI ZSM targets CL automation in the ETSI ZSM 009 [zsm-009-1][zsm-009-2][zsm-009-3] specifications, which analyse enablers and solutions for automation in end-to-end service and network management. ETSI GS ZSM 009-2 [zsm-009-2] specifically introduces the concept of CL Governance, with an analysis of challenges and architectural options to automate CLs' operation. **3GPP** introduces the topic of CL as part of the management services for Communication Service Assurance [28.535] [28.536] and architectures based on multi-layer closed loops operating at the communication service, network slice, slice subnet and network function layers.

TMForum ([IG1219A], [IG1219F]) has defined an AI Closed Loop Automation Management Platform (AI-CLAMP), where CLs at various levels (e.g., autonomic systems, autonomous networks, self-driving cars, etc.) consume data or events to develop and operate "automated solutions" able to adjust and optimize their performance towards meeting user and business expectations.

With this in mind, Hexa-X-II aims to design and implement a CL architecture that, following the SBA (Service Based Architecture) paradigm, integrates seamlessly with the 6G system. The CL will be orchestrable through the CL Governance service, and such orchestration will be part of the orchestration process for service and network (including transport). Each CL function will be deployed through the CL Governance service into the existing 6G management system, jointly with the other management functions, selecting their optimal placement in the cloud/edge continuum (function placement), in a dynamic and scalable manner, enabling their seamless interaction and cooperation with other 6G functions. Standard interfaces will enable the programmability of the various CL functions, consumed by a specific Governance function in charge of managing orchestration requests from the network and service orchestrators and triggering in turn the orchestration of a proper CL. In this regard, specific closed loop service models will be defined, along with a logic to translate such models in operation for the management of the closed loop functions. CL analysis and/or decision process can be AI-based and will support the interaction with possible existing AI/ML process already running in 6G system. In this sense, advanced topics such as the concept of cognitive closed loop, will be investigated. An initial design of the CL governance architecture, including the CL Governance service and its interaction with the rest of the 6G system, is described in the following section.

3.10.5 Identification of possible components and interfaces

Figure 3-40 shows a CL deployment inside an example of a 6G system implementation, coordinated through the CL Governance service. The components in green, with the explicit indication of an API, are CL-related functions while the elements in blue are existing and generalized functions belonging to the 6G system. The CL Governance service is the centralized entity in charge of instantiation, life-cycle management and operation of the components representing the stages of various CLs, i.e., Monitoring, Analysis, Decision, and Execution, with the additional Knowledge function where data produced and consumed by the other elements are stored and shared. The interface exposed by the CL Governance function can be exploited for the coordination of multiple and concurrent CLs, as described in Section 3.11.5. It should be noted that the knowledge element does not implement a stage itself, but it is rather an entity that stores and provides information, i.e., configurations, data, trained models, etc. In this sense, the persistence and the capability to securely expose the data is required, in order to be consumed by other elements of the loop and beyond.

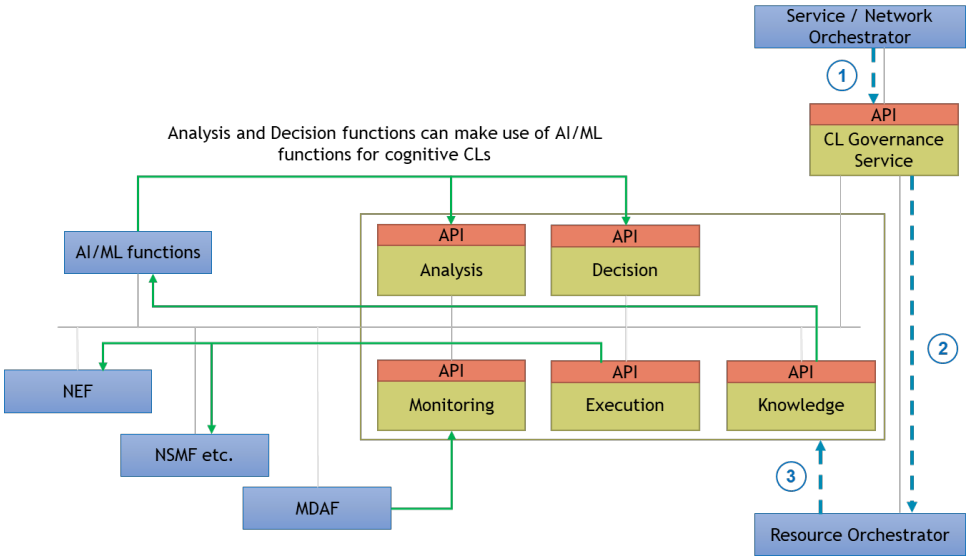


Figure 3-40: Closed loop components and Governance service

The creation of a CL is triggered during the orchestration of service or a network slice and involves also functions belonging to the 6G system. Following the figure, the Service/Network orchestrator requests the initiation of a CL through the interface exposed by the CL Governance Service (1) which in turn requests the Resource Manager the provisioning of the CL functions (2), that are in fact virtual elements, and their deployment in the 6G system (3). At runtime, the CL functions will implement the circular interaction monitoring-analysis-decision-execution with the support of the Knowledge function, interacting at the same time with existing 6G functions, as indicated by the green arrows in the figure. Figure 3-41 describes an example of workflow where the CL Governance function coordinated the provisioning of a CL operating at the service layer while, Table 3-11, provides a short description of each CL functions together with a set of possible technologies that may be exploited for their implementations. Sections 3.10.5.1-3.10.5.6 provide an initial definition of the abstract interfaces exposed by the CL functions and the CL Governance service.

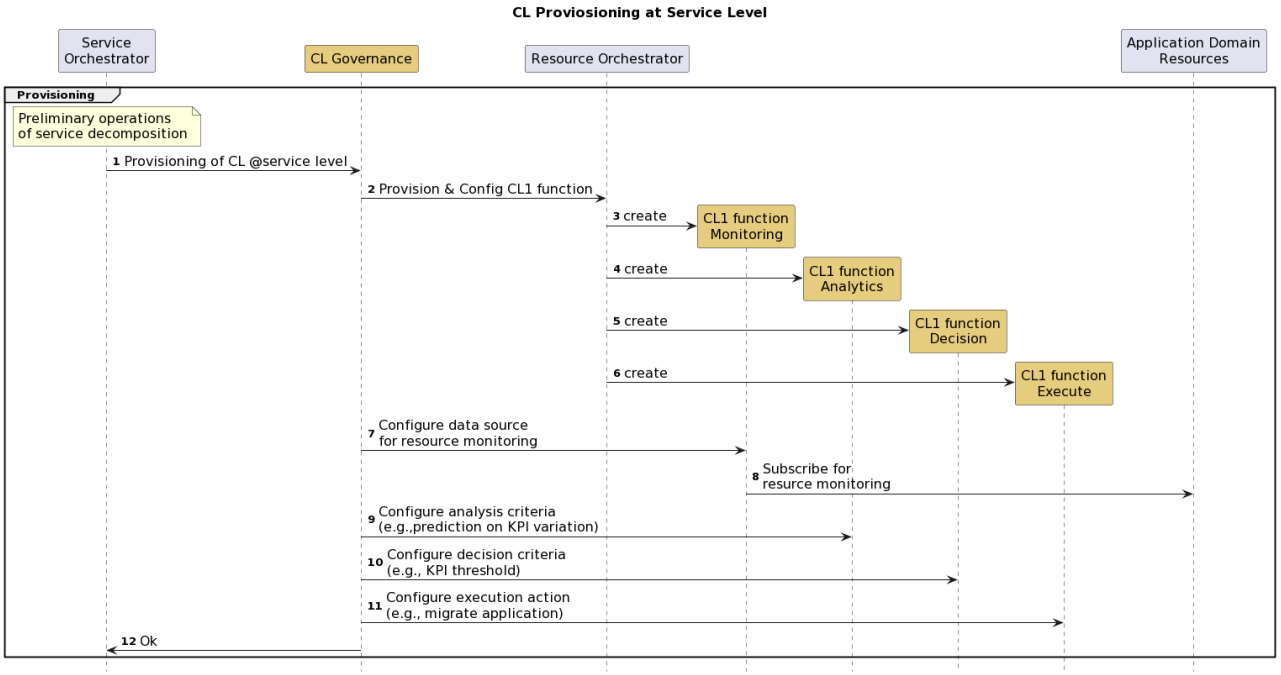


Figure 3-41: Closed loop provisioning workflow at Service Layer

Component	Description	Possible technologies
CL – Monitoring	Represents the first stage of the CL and collects data required for the system automation from system itself (e.g., MDAF) or external sources.	Prometheus, ELK, Grafana, Monitoring Platform from ANChOR
CL – Analysis	Analyses the data collected by the Monitoring stage and derives information on what is happening and or happened in the monitored systems. Can make use of information resulting from external AI/ML process e.g., belonging to the 6G system	TensorFlow, KubeFlow, ….
CL – Decision	Takes decisions on the basis of the information derived by the Analysis service. Such decision aims to produce possible corrective actions that would project the system towards a desired state. Can make use of information resulting from external AI/ML process e.g., belonging to the 6G system	
CL – Execution	Enforces on the system the decision(s) taken at the decision state, if any. This can include interactions with different management functions (e.g., CSMF, NSMF), NEF, etc.	Orchestrators, SDN controllers, RAN controllers, any other entity that can accept and apply a configuration in the target domain
CL – Knowledge	Stores data (e.g., configuration) used by the other stages of the CL or by other elements such as 6G system's AI/ML functions. Information stored can also include AI/ML models to be employed at Analysis/Decision stage.	For ML models: TF Serving, Model Registry in MLFlow For monitoring data: InfluxDB, MongoDB, MINIO, …

CL Governance Service	Allows the management of a CL by external entities such as Orchestrators and CL Coordinators (see Section 3.11). In this regard, it exposes a number of interfaces for CL life-cycle management, configuration, operations, status, and information retrieval, as reported in Table 3-12	OSM, Service Orchestration Platforms [HEX23-D63]

Table 3-11: List of CL functions and possible implementing technologies

3.10.5.1 CL – Monitoring

Interface name	Service descriptions	Potential consumers	Possible Standards or Technologies
Monitoring_config	Allows to set the configuration of the monitoring functions, e.g., activating new monitoring jobs or setting new data sources.	CL Governance Service	REST
Monitoring_data	Allows to retrieve the real-time or monitoring data through queries or subscribe/notify mechanisms	CL - Analysis CL - Decision CL - Knowledge	REST Message Queuing Telemetry Transport (MQTT)

Table 3-12: Interfaces exposed by CL Monitoring component

3.10.5.2 CL – Analysis

Interface name	Service descriptions	Potential consumers	Possible Standards or Technologies
Analysis_config	Allows to set configurations to the function, such as the type of analytics to be used or the AI/ML model to be employed in the loop	CL Governance Service	REST
Analysis_data	Allows to retrieve insights derived by the analysis process	CL - Decision CL - Knowledge	REST MQTT

Table 3-13: Interfaces exposed by CL Analysis component

3.10.5.3 CL – Decision

Interface name	Service descriptions	Potential consumers	Possible Standards or Technologies
Decision_config	Allows to set configurations to the function, such as the policies to be used in the infer the decision or the AI/ML model to be employed in the loop, in the case of AI-based decision process	CL Governance Service	REST

Table 3-14: Interfaces exposed by CL Decision component

3.10.5.4 CL – Execution

Interface name	Service descriptions	Potential consumers	Possible Standards or Technologies
Execution_config	Allows to set configurations to the function, such as the target functions for the configuration (NEF, NSMF, etc)	CL Governance Service	REST
Execution_operation	Allows to enforce on the system the decision(s) taken by the Decision function	CL - Decision CL - Knowledge	REST MQTT

Table 3-15: Interfaces exposed by CL Analysis component

3.10.5.5 CL – Knowledge

Interface name	Service descriptions	Potential consumers	Possible Standards or Technologies
Knowledge_config	Allows to store and retrieve information useful for the execution of the loop. It may include configurations, data, AI/ML models, etc.	CL Governance Service CL – Monitoring CL – Analysis CL – Decision CL - Execution	REST

Table 3-16: Interfaces exposed by CL Decision component

3.10.5.6 CL Governance Service

Interface name	Service descriptions	Potential consumers	Possible Standards or technologies
CL_LCM	Handles the provisioning, scaling, termination and, in general, Life Cycle Management (LCM) actions on Closed Loop services.	Service/Network Orchestrator	REST
CL_Operation	Allows to operate the various stages (as a whole or each of them) and related functions of the CL (e.g., to activate and stop the execution of each stage)	Service/Network Orchestrator CL Coordination Service CL Governance Service of other CLs (in case of direct interactions between peer CLs)	REST

CL_Config	Allows to configure objectives, targets, internal policies, and in general configuration attributes of the CL.	Service/Network Orchestrator CL Coordination Service CL Governance Service of other CLs (in case of direct interactions between peer CLs)	REST
CL_Status	Allows to retrieve the internal status of the CL, info about potential failures, or CL statistics (via queries or subscribe/notify mechanisms).	Service/Network Orchestrator CL Coordination Service CL Governance Service of other CLs (in case of direct interactions between peer CLs)	REST MQTT
CL_Info	Allows to retrieve information from the internal stages of the CL (via queries or subscribe/notify mechanisms).	Service/Network Orchestrator CL Coordination Service CL Governance Service of other CLs (in case of direct interactions between peer CLs)	REST MQTT
Delegation_Management	Allows to receive delegations from external CLs (see section 3.11.3).	CL Coordination Service	REST

Table 3-17: Interfaces exposed by CL Governance component

3.10.6 Relationship with other Enablers

Enabler 10 maintains relationships with several enablers, in particular with Enablers 1, 2, 3, 4, 6 and 11. **Enabler 2**, Programmable Network Monitoring, is one of the entities to feed the CL *monitoring* stage. A CL consists indeed of several sequential stages including the Monitoring, as explained in Section 3.10.1, and network KPIs can be used as input for the CL logic. Similarly, **Enabler 1** (Programmable Flexible Network Configuration) provides an interface for the *execute* stage to request the modification of the network configuration, as elaborated by the analysis and decision stages of the CL. Collaborations with **Enabler 4** (Trustworthy 3rd party management) can allow to apply the CL concept to SLA management. The Integration Fabric (**Enabler 3**) provides the interfaces for the communication between different modules at different layers, exploited by the Governance to communicate with Resource Orchestrator, CL Coordination layer (**Enabler 11**), and Continuum Orchestration (**Enabler 6**). In addition, the Integration Fabric can also be used to mediate the interaction among the CL functions. **Enabler 6** (Orchestration mechanisms for the continuum) allows the CL Governance service to provision the CL functions in the extreme edge/edge/cloud continuum, at orchestration time and under explicit request of the orchestrator (Service, Network, etc.). **Enabler 11** is in charge of coordinating multiple coexisting CLs, by exploiting interfaces exposed by the related Governance entities.

3.11 Enabler 11: Zero-touch multiple closed loop coordination

3.11.1 Motivation

The increasing level of network automation enabled through the adoption of multiple CL-based strategies may bring some undesired effects and negative consequences such as reduced network stability, inconsistencies in the end-to-end and multi-domain configuration or contrasting decisions at the service and network layer orchestration.

This is usually caused by concurrent closed loops that work in stand-alone manner, each of them with their own objectives and goals, and operating with restricted scopes that limit their visibility to small sets of managed entities (e.g., services, slices, slice subnets, groups of edge nodes, etc.), single network domains and single tenants. The combination of the automated actions of these CLs may lead to unstable configurations and conflicting decisions that need to be regulated and mitigated through a coordination of the various CLs.

Coordination of CLs is therefore essential to ensure that their decisions and actions are aligned and complementary, leading to improved network performance, stability, and consistency. By coordinating closed loops, they can share information and feedback, avoid conflicts, and ensure that their objectives and goals are aligned with the overall network objectives and goals.

3.11.2 Objectives

Similar to Enabler 10, the goal of the Zero-touch multiple closed loop coordination is to reduce the OPEX, in this case, going beyond the concept of single closed loop governance. 6G systems are characterised by the coexistence of multiple closed loops that may run in different scopes (i.e., per domain or network segments, per infrastructure layer, per stakeholder) and at different time scale (see Section 3.10.3). In this regard, the aim of this enabler is to provide a multiple-closed loop coordination in an automatic manner in order to:

- Maximize the control effect through inter-CL collaboration (e.g., exploiting delegation mechanisms).
- Mitigate conflicts, limiting contrasting decisions and race conditions that may lead to unstable network states.

3.11.3 Description of the solution

As analysed in the previous section, CLs can operate within different architecture layers, domains, tenant's scopes and timeframes. However, the execution of their decisions has an impact on several managed entities, even beyond the ones under direct scope and control of the given CL. Moreover, in order to take effective decisions applicable to wider scopes and infrastructures, the collaboration among multiple CLs is fundamental. In this sense, CLs can cooperate following a peer-to-peer or a hierarchical approach. The first model is based on horizontal interactions where each CL is responsible for its own domain but through the exchange of information, with different levels of exposure, they reach decisions which are globally consistent. The hierarchical model is based on the concept of a "parent" CL with a global point of view or E2E perspective, which drives the actions of multiple "child" CLs, acting as a centralized point. The hierarchical model is associated to the "delegation" concept, where the parent CL *delegates* some actions to one or more child CLs. It should be noted that a mix of cooperation models can be adopted at the different stages of coordinated CLs. For example, the monitoring and the analysis stages can operate with peer-to-peer interactions, while the decision and execution stage can operate with hierarchical interactions. A critical aspect is related to CLs operating at different administrative domains, where the cooperation models should take into account constraints related to data ownership, level of exposure of CL capabilities and information, as well as access control policies for committing commands at the execution stage.

Hexa-X-II will address the coordination of concurrent CLs with multi-dimensional scopes, as follows:

- Short vs. medium-/long-term decisions.
- CLs operating at the service, network and infrastructure layers.
- CLs operating at different domains, e.g., at the cloud, edge and extreme edge domain, or at the access, core or transport network domain, or in different technological domains within the transport network.
- CLs operating at different tenants' levels.

- CLs operating at different network slices' levels.

Figure 3-42 provides an example of peer-to-peer coordination among CLs operating at the service layer in different domains, while a delegation-based approach is adopted to coordinate the multi-layer CLs within each domain. This approach is particularly suitable for multi-operator scenarios, e.g., in case of federation. An alternative approach, which reduces the complexity but is mostly applicable in single-operator scenarios is represented in Figure 3-43, where the delegation approach is adopted among per-layer CLs operating with a multi-domain scope at the service and network-layer, while a peer-to-peer cooperation regulates the interaction of the CLs at the infrastructure layers operating with per-domain scope.

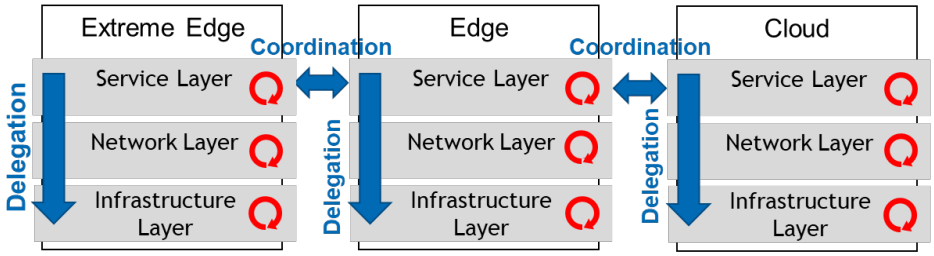


Figure 3-42: Example of peer-to-peer coordination among CLs at extreme edge, edge and cloud domains combined with hierarchical delegation of per-layer CLs

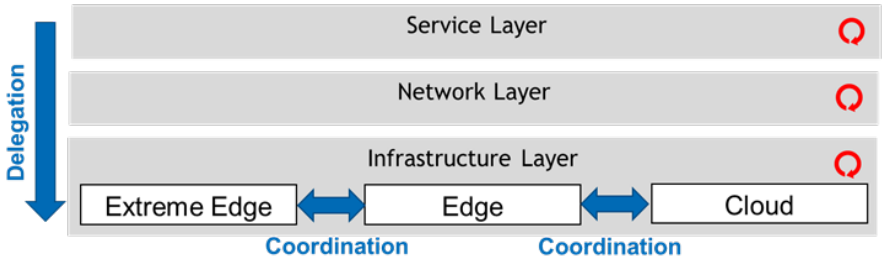


Figure 3-43: Example of delegation among per-layer, multi-domain CLs combined with peer-to-peer coordination of per-domain CLs at the infrastructure layer

The solution will propose workflows and interfaces, at the functions implementing the various CL stages, to enable delegation and escalation strategies for the coordination of concurrent CLs, evaluating the pros and cons of the various interaction models when applied to the different scenarios mentioned above. This enabler is closely related with other project outcomes as network digital twins (Enabler 9) may be applied for early detection of conflict decisions, while explainable AI techniques may help in the manual operations for conflict mitigation and resolution.

3.11.4 SoTA and Beyond SoTA

The Closed loop coordination is a step beyond the concept of Closed loop Governance and is debated in the same documents reported in Section 3.10.4 for Enabler 10, produced by ETSI ZSM, 3GPP, in particular in SA2 for network data analytics and closed loops for SLA management (see section 7.1.1.1), and TMForum for autonomous networks (see section 7.2.1.1). ETSI ENI does not discuss the topic of the coordination of multiple and concurrent loops.

Hexa-X-II aims at building a Closed loop Coordination (CLC) framework capable of dealing with multiple CL in different scopes, such as per-tenant CLs, per-layer (infrastructure, network, applications), per and inter-domain. The CLC will consist of multiple functions specialized in different aspects of the coordination and, for the most relevant, specific workflows and interfaces will be defined with the aim of implementing concrete examples of conflict detection between concurrent CLs and subsequent mitigation actions, along with coordination models such as Delegation (see Section 3.11.3) and, on the opposite direction, Escalation. The coordination of different loops requires that such loops must be somehow exposed, i.e., the coordinator should be able of consume specific APIs exposed by the different loops to enforce the coordination actions. In this regard, the coordination will be able to consume specific interfaces exposed by the CL Governance service for

each managed CL (see Table 3-18). Another point that will be investigated is the cross-CL Knowledge sharing, in terms of inter-CL visibility, secure access control, and efficient data distribution across multiple domains.

3.11.5 Identification of possible components and interfaces

An initial architecture of CLC is shown in Figure 3-44, where the set of coordination functions operates on top of a set of different running CLs. In this architecture, the CL Governance service is the only entity at the CL level that exposes an external interface towards the different element of the CLC, wrapping the internal mechanisms of the various CLs. Similarly, the CL Coordination service offers the API to register new CLs, to request coordination actions or to receive notifications related to existing CLs, wrapping the interaction with the other "internal" CLC functions. All of them can interact with each other, as well as with functions belonging to both 6G system and CLs, in read-only mode, in order to gather information relevant for coordination, arbitration, conflict detection, mitigation or resolution decisions. The interfaces exposed by the CL Coordination service are reported in Table 3-18.

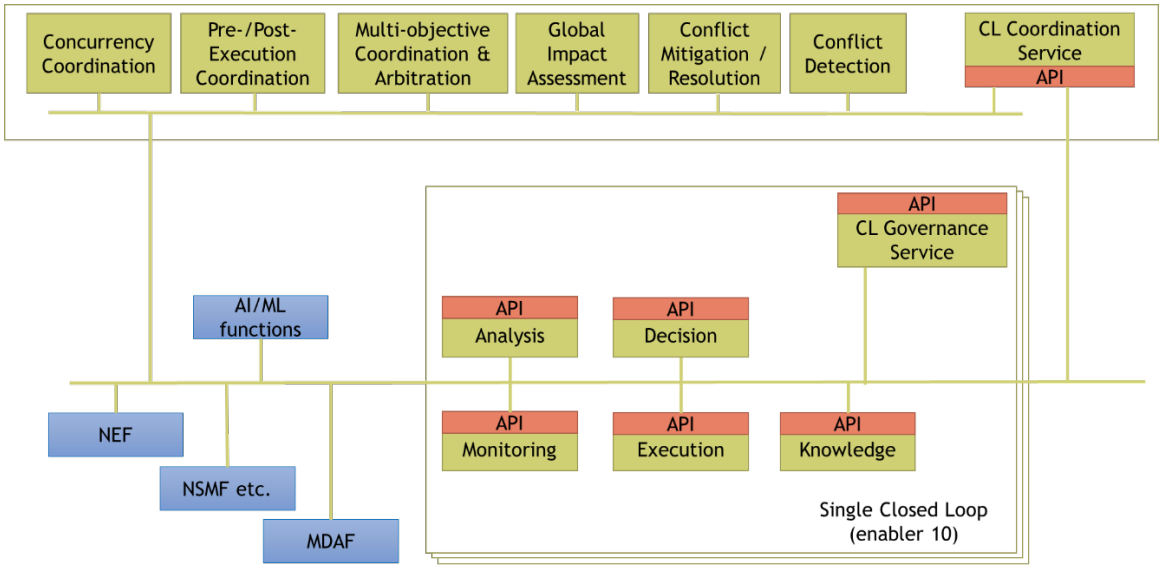


Figure 3-44: Closed loop Coordination architecture

3.11.5.1 CL Coordination Service

Interface name	Service descriptions	Potential consumers	Possible Standards or Technologies
CL_LCM_Notifications	Allows to receive information on changes on the LC of CLs, including registration of new CL instances.	CL Governance Service	REST
Multi-CL_Operation	Allows to operate over multiple CLs providing global commands or objectives.	Service/Network Orchestrator External CL Coordination Service (in case of hierarchical CLs with multiple layers)	REST
Multi-CL_Config	Allows to configure objectives, targets, internal policies, and in	Service/Network Orchestrator	REST

	general configuration attributes for clusters of CLs.	External CL Coordination Service (in case of hierarchical CLs with multiple layers)	
Multi-CL_Conflict	Provides information on managed conflicts (via queries or subscribe/notify mechanisms).	Service/Network Orchestrator External CL Coordination Service (in case of hierarchical CLs with multiple layers)	REST MQTT
Multi-CL_Impact	Allows to request the evaluation of the impact of stages from a set of CLs	Service/Network Orchestrator External CL Coordination Service (in case of hierarchical CLs with multiple layers)	REST
Escalation_Notification	Allows to receive notifications from CL Governance about the need to escalate for a given CL objective to external CLs.	CL Governance Service	MQTT

Table 3-18: Interfaces exposed by CL Coordination Service

3.11.6 Relationship with other Enablers

CL Coordination exploits interfaces provide by CL Governance entities (**Enabler 10**) to deal with multiple and concurrent closed loops in order to enable a system-wide control automation. Furthermore, CL Coordination exploits the Integration Fabric (**Enabler 3**) to interact with the CL Governance and other coordination entities.

4 Planned Proof of Concepts

The enablers presented in Section 3 will be validated in Proofs of Concept (PoC) integrating a subset of them to demonstrate a concrete use case. The planned PoCs involving the smart network management enablers presented in this document are described in this section. Those two PoCs will be later integrated in the project to build a full end-to-end system PoC that is targeted to be released by the end of the project (June 2025).

4.1 Component-PoC#A.1. Sustainability and trustworthy-oriented orchestration in 6G

This PoC will demonstrate AI mechanisms for control and programmability of 6G network elements focusing on energy consumption aspects. Given the AI mechanisms that need to be executed and the system capabilities, resources will be assigned in a manner that optimizes the energy consumption while considering performance and security requirements (e.g., whether a deployed node should be used to run a certain service).

The solution will use network programmability and consider a zero-touch approach to automate the network reconfiguration and energy-aware self-optimization at runtime. Since AI takes a critical role in configuring the network for optimization from energy saving and security perspectives, the AI models providing these functionalities should also be hardened-by-design to protect against possible security and privacy attacks targeting AI model itself. Figure 4-1 provides a high level representation of PoC#A.1 flow. The status, the capabilities, the energy consumption, and trustworthiness level of the elements of the environment as well as the service and workload requirements will be collected to ensure energy efficiency, as well as high security, trust and performance. AI mechanisms will be explored to control and programming the network elements as well as for decision enforcement.

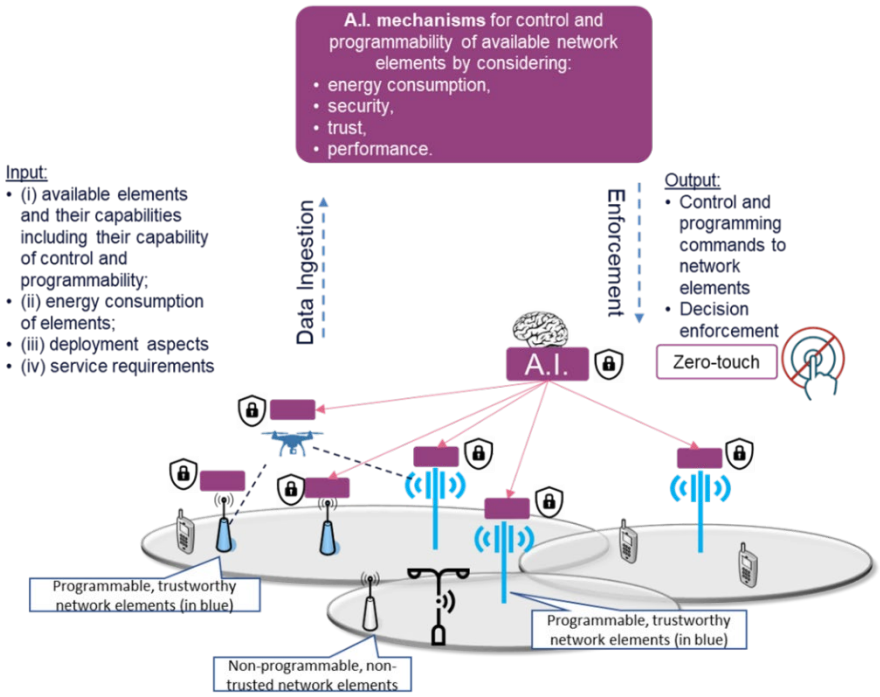


Figure 4-1: High level view of the PoC#A.1

4.1.1 Description of the functionalities

The inputs to the implemented control mechanism encompass various crucial aspects, including (i) the identification of available network elements and their respective trust levels and capabilities, including their control and programmability; (ii) meticulous consideration of energy consumption; (iii) the intricacies of deployment, particularly the communication between these network elements; (iv) the precise delineation of service requirements; and (v) the secure collection of data from heterogeneous sources, which serves as input

for distributed AI mechanisms. The solution is designed to incorporate robust and privacy-enhanced AI mechanisms that effectively manage the control and programmability of these network elements, while balancing energy consumption, security, trust, and performance considerations. Additionally, it includes mechanisms for the dynamic deployment and configuration of monitoring probes, as well as network monitoring tailored to AI-based automated traffic profiling, distinguishing between local and centralized traffic management for optimal network operation.

The output of this Proof of Concept (PoC) will comprise control and programming commands directed towards the selected network elements, facilitating the enforcement of critical decisions. A central tenet of this endeavour is the pursuit of trustworthy zero-touch operations, pivotal for ensuring the requisite quality of services and energy efficiency. Key functional components integral to this PoC encompass collaborative robots (cobots), a versatile multi-platform resource orchestrator, and sophisticated intent-based mechanisms. Cobots will play a pivotal role, serving to inspect, maintain, and, when necessary, repair components within smart networks, including themselves. The management and orchestration frameworks under scrutiny span various facets of the computing continuum, addressing network services provisioning, 6G applications provisioning, and the intricate interplay between network and application providers. Importantly, the utilization of intent-based approaches will enable seamless integration of Hexa-X-II managed resources and third-party applications, fostering direct and open communication between them, with due consideration given to intent-based conflict resolution mechanisms.

4.1.2 Benefits

Zero-touch control and programming of 6G network elements is enhanced with trustworthiness and security (including security and privacy for AI against attack targeted to AI pipeline) in order to ensure reliability in addition to energy efficiency and performance towards sustainability goal.

4.1.3 Enablers contributing to this PoC

Enabler 1 - Programmable flexible network configuration: specific network element control and management will be considered together with cloud-native SDN controller.

Enabler 2 - Programmable network monitoring and telemetry: monitoring framework will be provided and integrated with other enablers.

Enabler 3 - Integration fabric: communication bus that eases the multi-domain solutions integration ensuring secured data exchange.

Enabler 4 - Trustworthy 3rd party management: Trust manager component for evaluating the trust level of each network/compute node (e.g., servers, robotic nodes) available in the system with the use of AI/ML.

Enabler 6 - Orchestration mechanisms for the computing continuum: deployment and runtime management mechanisms for the developed service/application chain, considering resources in the edge and cloud part of the continuum.

Enabler 7 – Sustainable AI-based control: Optimisation mechanisms for functionality placement (including various computational workloads, services etc.) to the available compute nodes (e.g., servers, robotic nodes) towards energy efficiency and trustworthiness.

Enabler 8 – Trustworthy AI/ML-based control: exploration of how AI/ML-based control can be protected against adversarial attacks and privacy attacks to provide a more robust model and to prevent sensitive data to be disclosed.

Enabler 10 - Zero-touch closed loop governance: provisioning of CL functions for automated migration of application components among cobots, based on battery level.

4.2 Component-PoC#B.1. AI-assisted end-to-end lifecycle management of a 6G latency-sensitive service across the compute continuum

This PoC regards the validation of the main processes for the end-to-end lifecycle management of a 6G latency-sensitive service. It includes the onboarding, deployment, and operational phase of the service over programmable infrastructure across the compute continuum, considering device and cloud computing resources (see Figure 4-2). It aims to highlight aspects related to automated service onboarding (e.g., by third parties without explicit knowledge of infrastructure management processes), synergetic orchestration of the service over multiple network and edge/cloud computing resources (dynamic usage of multi-cluster resources and exploitation of edge computing functionalities) and the intelligence/automation features that can be introduced through the exploitation of AI technologies.

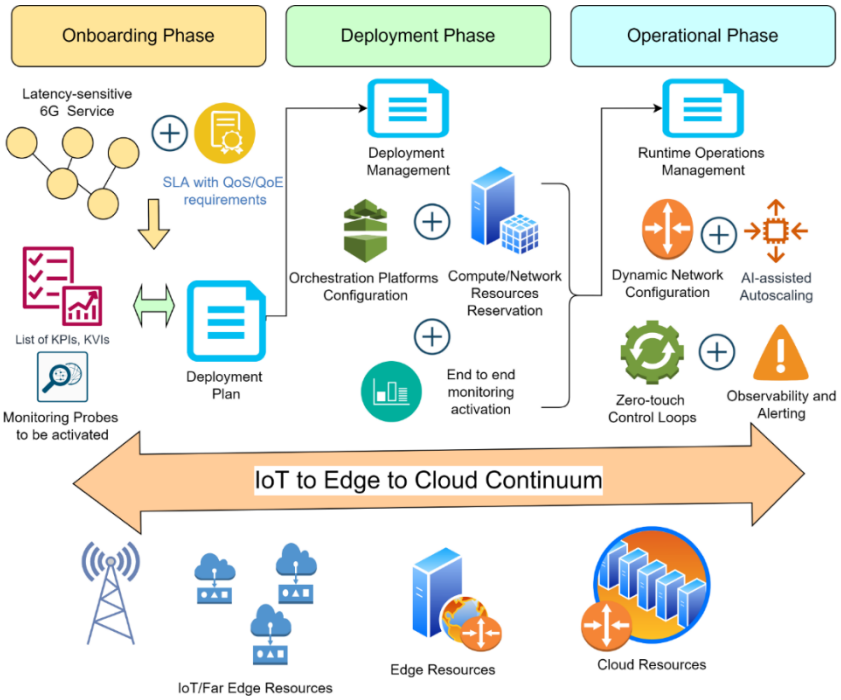


Figure 4-2: High level view of the PoC#B.1

Some of the key features that can be validated for serving a 6G latency-sensitive services include the quick and reactive management of bursty workloads (e.g., need for automated enforcement of scaling policies, agility in end-to-end network configuration), the data fusion and prompt analysis of telemetry data for bottlenecks identification, anomalies detection and undertaking of corrective actions.

4.2.1 Description of the functionalities

During the **onboarding phase**, critical steps are undertaken to ensure the successful deployment of a 6G service. This encompasses defining deployment types and meticulously outlining privacy and quality of service/experience (QoS/QoE) requirements. It involves the intricate task of translating Service Level Agreement (SLA) prerequisites, including Key Value Indicators (KVIs), Key Performance Indicators (KPIs), and Trust Level Agreements (TLAs), into precise deployment and configuration demands across the entire compute continuum, spanning from IoT devices to edge computing and cloud infrastructure. Additionally, a robust licensing strategy is articulated for operation across multi-domain infrastructure, utilizing mechanisms such as OAuth and Distributed Ledger Technologies (DLTs). Furthermore, the onboarding phase entails the configuration of a management fabric setup tailored to the unique needs of the service, establishing a distinct and customized management space that grants customers control over their services while maintaining the provider's resource controllability, all underpinned by end-to-end trust guarantees.

During the inter-computing **deployment phase**, a comprehensive approach is taken to seamlessly integrate various computing elements. This involves crafting a deployment plan that spans multiple network management and edge/cloud orchestration platforms, meticulously specifying orchestration mechanisms tailored to each segment of the compute continuum while harnessing the synergy of these mechanisms. The deployment of the service itself is executed with precision, accompanied by the activation of end-to-end monitoring systems and the seamless integration of telemetry component placement algorithms within the orchestration process. Furthermore, network equipment is configured via a cloud-native Software-Defined Networking (SDN) controller, ensuring the efficient and agile management of network resources throughout the deployment process.

In the AI-assisted **operational phase**, a multifaceted approach is employed to ensure the optimal performance and responsiveness of the deployed systems. Continuous monitoring and alerting mechanisms are in place, aligning with the stipulated Service Level Agreements (SLA) and Quality of Experience (QoE) requirements. Dynamic and AI-assisted enforcement mechanisms govern scaling, live migration, and compute offloading policies, adapting in real-time to the evolving workload demands. Special emphasis is placed on optimizing the performance of edge computing services, crucial for achieving low end-to-end latency, and this is facilitated through dynamic resource management at the edge and ensuring business continuity. The network undergoes dynamic reconfigurations as directed by AI-assisted control loops, promoting agility and adaptability in response to changing conditions. Furthermore, zero-touch control loops are seamlessly integrated to support timely telemetry for time-sensitive network services. This intricate orchestration extends across multiple domains, from IoT to edge to cloud, fostering "horizontal coordination," all facilitated by distributed AI agents and functions that are provisioned and reconfigured dynamically to ensure efficient and effective system operation.

4.2.2 Benefits

This PoC offers a comprehensive validation of critical processes in the end-to-end lifecycle management of latency-sensitive 6G services, bringing forth a multitude of benefits. It encompasses the entire journey, from onboarding through deployment to AI-assisted operational phases, ensuring a seamless and agile service delivery. Notable advantages include the ability to facilitate automated service onboarding, even by third parties without in-depth infrastructure knowledge, synergetic orchestration across diverse network and edge/cloud resources, and the infusion of intelligence and automation through AI technologies. Key functionalities validated encompass agile management of bursty workloads, swift enforcement of scaling policies, dynamic end-to-end network configuration, prompt analysis of telemetry data for bottleneck identification and anomaly detection, as well as corrective action implementation. This PoC thus promises to enhance the efficiency, agility, and intelligence of 6G services, laying the foundation for advanced, low-latency communication and service provision in the evolving digital landscape.

4.2.3 Enablers contributing to this PoC

Enabler 1 - Programmable flexible network configuration: specific network element control and management will be considered together with cloud-native SDN controller.

Enabler 2 - Programmable network monitoring and telemetry: Monitoring framework will be provided to acquire, process and export multiple data sources. Provision of QoS and telemetry data to the activated orchestration mechanisms.

Enabler 3 - Integration fabric: ease the onboarding and operational phases, ensuring a multi-domain communication bus, that enables liquid and frictionless interoperation between services.

Enabler 4 - Trustworthy 3rd party management: Trust manager component for evaluating the trust level of each network/compute node (e.g., servers, robotic nodes) available in the system with the use of AI/ML.

Enabler 5 - Multi-cloud management mechanisms: Management of the deployment of the service over multi-cluster infrastructure. Activation of scaling policies per part of the infrastructure, management of live migration mechanisms and management of interconnection of the clusters.

Enabler 6 - Orchestration mechanisms for the computing continuum: Enforcement of intelligent orchestration mechanisms, considering intent-driven approaches for deployment of distributed application/service components, as well as technologies that are based on multi-agent systems and reinforcement learning techniques.

Enabler 10 - Zero-touch closed loop governance: Provisioning of multiple CLs exploiting AI techniques at the analysis stage and with CL functions deployed in the continuum.

Enabler 11 - Zero-touch multiple closed loop coordination: Coordination of multiple CLs operating in different domains (e.g., for resource orchestration over edge and cloud domains), with conflict resolution mechanisms.

5 Conclusions

This deliverable reports on the initial work that has been performed in Hexa-X-II project regarding Smart Network Management. The document has set the foundations of the smart network management enablers that will be fully designed and implemented in later phases of the project.

An overview on the drivers behind the development of 6G technology has been provided. Firstly, identifying and analysing the environmental, social, and economic factors influencing the adoption and deployment of 6G technology. Secondly, the potential benefits and challenges associated with these drivers have been explored. Finally, the preliminary introduction of key performance indicators (KPIs) has been related to Hexa-X KPI. Hexa-X-II KPIs and requirements will be considered in the next phases of the design. The overarching objectives and goals for the management and orchestration in the Hexa-X-II project have been elaborated as well as covering how they relate with the identified KPIs so far, highlighting the crucial role played by effective management strategies in realizing the full potential of 6G technology.

Section 3 is the main section in this deliverable, which has provided an overview of the work done so far regarding the enablers. 11 enablers have been identified for Management and Orchestration of 6G networks based on the settled objectives. The described M&O enablers are the following:

- Enabler 1: Programmable flexible network configuration
- Enabler 2: Programmable network monitoring and telemetry
- Enabler 3: Integration fabric
- Enabler 4: Trustworthy 3rd party management
- Enabler 5: Multi-cloud management mechanisms
- Enabler 6: Orchestration mechanisms for the computing continuum
- Enabler 7: Sustainable AI/ML-based control
- Enabler 8: Trustworthy AI/ML-based control
- Enabler 9: Network Digital Twins
- Enabler 10: Zero-touch closed loop governance
- Enabler 11: Zero-touch multiple closed loop coordination

For each enabler, the document presents: i) motivation, this is why the enabler is needed and the problem it solves, ii) main overall objective, iii) description overview, iv) key references in the SoTA used as starting point and how the enabler goes beyond that SoTA, v) identification of possible components and interfaces, and vi) relationship with other enablers.

The identified list of enablers are expected to have significant positive impact towards 6G regarding the following aspects: i) Increase of network automation and network autonomy and consequently also reducing operational expenses (OPEX) (Enablers 1, 3, 7, 8, 9, 10, 11), ii) Environmental sustainability, network efficiency and decarbonization (Enabler 7) iii) Trustworthiness (Enablers 3, 8) iv) Improved performance in terms of zero-perceived latency and higher speed (Enablers 2, 5, 6).

Selected enablers will be validated in Proof of Concepts (PoC) being integrated with other enablers to build a concrete use case. The planned PoCs involving the smart network management enablers presented are: PoC#A.1. Sustainability and trustworthy-oriented orchestration in 6G and PoC#B.1. AI-assisted end-to-end lifecycle management of a 6G latency-sensitive service across the compute continuum. Table 5-1 provides a roadmap for enabler integration in the proposed PoCs.

	PoC#A.1	PoC#B.1
Enabler 1: Programmable flexible network configuration	X	X
Enabler 2: Programmable network monitoring and telemetry	X	X
Enabler 3: Integration fabric	X	X
Enabler 4: Trustworthy 3rd party management	X	X
Enabler 5: Multi-cloud management mechanisms		X

Enabler 6: Orchestration mechanisms for the computing continuum	X	X
Enabler 7: Sustainable AI/ML-based control	X	
Enabler 8: Trustworthy AI/ML-based control	X	
Enabler 9: Network Digital Twins		
Enabler 10: Zero-touch closed loop governance	X	X
Enabler 11: Zero-touch multiple closed loop coordination		X

Table 5-1: Summary of PoC and M&O Enablers

The updated SoTA overview relevant for Management and Orchestration (M&O) is provided in the Annex of this document. It covers SDO (Standard Development Organizations) and Open-Source Software, Industry fora and Research Projects. First, the standards and advancements proposed by SDOs and contributions from the open-source software community have been collected. Second, the subsection on Industry Fora has covered the contributions and developments from various fora, exploring the ongoing efforts and collaborations within the industry to advance 6G technology and identifying key trends and innovations. The subsection on Research Projects has discussed noteworthy research projects in the EC SNS domain, examining their methodologies and potential implications for the Hexa-X-II project.

This initial identification and foundations design of the 11 enablers presented in this document have been performed based on the high-level drivers [HEX223-D11], objectives analysis established for smart network management, the initial end-to-end Hexa-X-II system blueprint provided in [HEX223-D21] and SoTA analysis. The work presented in this deliverable sets up a framework for further research and development within the project:

- Further requirements will be incorporated in the following design iteration, once that work done in the project regarding use cases analysis released by Dec'23.
- The overall Hexa-X-II system blueprint regarding M&O functionalities will also be considered for the update of the current enablers in following iterations in the project in a top-down approach.
- The current enablers will be considered also in a bottom-up approach to design the 6G platform blueprint that the whole Hexa-X-II project aims to build.
- Next deliverable [HEX224-D63] will cover the initial Design of 6G Smart Network Management Framework. This report will provide early results on the 6G smart network management enablers, contributing to the 2nd iteration of the overall Hexa-X-II system blueprint.

6 References

[23.222] 3GPP TS 23.222, Functional architecture and information flows to support Common API Framework for 3GPP Northbound APIs; Stage 2, v16.9.0, October 2020.

[23.288] 3GPP TS 23.288, Architecture enhancements for 5G System (5GS) to support network data analytics services, v17.04.0, May 2022.

[23.436] 3GPP TS 23.436, Procedures for Application Data Analytics Enablement Service, v2.0, June 2023.

[23.501] 3GPP TS 23.501, "System architecture for the 5G System (5GS).", v17.10.0, September 2023.

[23.503] 3GPP TS 23.503, "Policy and charging control framework for the 5G System (5GS); Stage 2", v17.10.0, September 2023.

[23.558] 3GPP TS 23.558, Architecture for enabling Edge Applications., v17.9.0, September 2023.

[23.700-80] 3GPP T.R. 23.700-80, Study on 5G System Support for AI/ML-based Services, v18.0.0, December 2022.

[23.700-81] 3GPP T.R. 23.700-81, Study of Enablers for Network Automation for the 5G System (5GS); Phase 3, v18.0.0, December 2022.

[24.526] 3GPP TS 24.526, "User Equipment (UE) policies for 5G System (5GS);. Stage 3", v17.8.0, March 2023.

[28.533] 3GPP TS 28.533, "Management and orchestration; Architecture framework.", v17.2.0, March 2023.

[28.535] 3GPP TS 28.535 "Management and orchestration; Management services for communication service assurance; Requirements", v17.7.0., June 2023.

[28.536] 3GPP TS 28.536 "Management and orchestration; Management services for communication service assurance; Stage 2 and stage 3", v17.5.0, March 2023.

[28.541] 3GPP TS 28.541, "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3", v17.12.0, September 2023.

[28.804] 3GPP TR 28.804, "Telecommunication management; Study on tenancy concept in 5G networks and network slicing management", v16.0.1, October 2019.

[28.817] 3GPP TR 28.817, "Management and orchestration; Study on access control for management service", v17.0.0, December 2021.

[28.824] 3GPP TR 28.824, "Study on network slice management capability exposure. Status: Under change control", v0.12.0, June 2023.

[5GAM] 5G Americas white paper, Distributed Compute and Communications in 5G, https://www.5gamericas.org/distributed-compute-and-communication-in-5g/

[6GR23-D21] Use and Business Cases, Design and Technology Requirements, and Architecture Specification. May 2023.

[ABB+20] P. Ala-Pietilä, Y. Bonnet, U. Bergmann, M. Bielikova, C. Bonefeld-Dahl, W. Bauer, and A. Van Wynsberghe, "The assessment list for trustworthy artificial intelligence (ALTAI)," European Commission, 2020.

[AFP+22] P. Almasan M. Ferriol-Galmés, J. Paillisse, J. Suárez-Varela, D. Perino, D. Lopéz, A. A. Pastor, Perales, P. Harvey, L. Ciavaglia, L. Wong, V. Ram, S. Xiao, X. Shi, X. Cheng, Al. Cabellos-Aparicio, and P. Barlet-Ros, "Network Digital Twin: Context, Enabling Technologies, and Opportunities", IEEE Comsoc magazine, vol. 60, no. 11, pp. 22-27, November 2022.

[AIA23] AI@Edge project. https://aiatedge.eu/

[AIA-D2.2] AI@Edge, "Preliminary assessment of system architecture, interfaces specifications, and techno-economic analysis," 2022.

[AIET23] ITU-T, "Focus Group on Environmental Efficiency for Artificial Intelligence and other Emerging Technologies (FG-AI4EE)," [Online]. Available: https://www.itu.int/en/ITU-T/focusgroups/ai4ee/Pages/default.aspx

[airflow] Apache Airflow. 2023. Available at: https://airflow.apache.org/ [Accessed 05 April 2023].

[AJ+18] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in Advances in neural information processing systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018.

[APA23] Apache Pulsar, 2023. [online]. Available: https://pulsar.apache.org/docs/3.0.x/concepts-architecture-overview/

[BAC+22] J. Bachiega, B. Costa, L. Rebouças de Carvalho, M. Rosa, and A. Araujo, "Computational Resource Allocation in Fog Computing: A Comprehensive Survey," in ACM Computing Surveys, 2022.

[BCP+16] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, ad W. Zaremba, "OpenAI Gym", 2016. [Online]. Available: https://arxiv.org/abs/1606.01540

[BHV+23] H. Bousbiat, Y. Himeur, I. Varlamis, F. Bensaali, and A, Amira, "Neural Load Disaggregation: Meta-Analysis," Federated Learning and Beyond. Energies, 2023;

[BK+20] B. Brik, and A. Ksentini, "On predicting service-oriented network slices performances in 5G: A federated learning approach," In 2020 IEEE 45th Conference on Local Computer Networks (LCN), pp. 164-171, 2020.

[BKT+21] J. Busch, A. Kocheturov, V. Tresp, and T. Seidl, "Nf-gnn: Network flow graph neural networks for malware detection and classification," In 33rd International Conference on Scientific and Statistical Database Management, pp. 121-132, July 2021.

[BM23] C. J. Bernardos and A. Mourad, "MIPv6 RAW mobility", Internet-Draft, 2023, IETF

[BNS+06] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?." Proceedings of the 2006 ACM Symposium on Information, computer and communications security. 2006.

[BT+20] C. Benzaid, and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," IEEE Network, vol. 34, no. 2, pp. 186-194, 2020.

[CAD+18] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, "Adversarial attacks and defences: A survey," 2018, arXiv:1810.00069.

[CAM23] CAMARA GitHub, API backlog Working Group [Online]. Link: https://github.com/CAM23b/WorkingGroups/blob/main/APIBacklog/documentation/APIBacklog.md

[CAM23b] Linux Foundation's CAMARA website [Online]. Link: https://CAM23b.org

[CBR+22] B. Costa, J. Bachiega, L. Rebouças, and A. Araujo, "Orchestration in Fog Computing: A Comprehensive Survey," in ACM Computing Surveys, vol 55, no.2, pp. 1-34, 2022.

[CLA+21] S. Clayman, A. Neto, F. Verdi, S. Correa, S. Sampaio, I. Sakelariou, L. Mamatas, R. Pasquini, K. Cardoso, F. Tusa, C. Rothenberg, and J. Serrat, Joan "The NECOS Approach to End-to-End Cloud-Network Slicing as a Service," in IEEE Communications Magazine, vol. 59, no. 3, pp. 91-97, 2021.

[COI23] Computing in the Network Research Group (coinrg), [Online]. Available: https://datatracker.ietf.org/rg/coinrg/about/

[CON23] CONFIDENTIAL6G project, SNS. [Online]. Available: https://confidential6g.eu

[DCA20] Acumos DCAE integrating in ONAP wiki, [Online]. Available at: https://wiki.onap.org/display/DW/Acumos+DCAE+Integration (Accessed 17 July 2023).

[DCA23] Git dcaegen2 ml-prediction-ms [Online], Available at: https://git.onap.org/dcaegen2/services/tree/components/ml-prediction-ms (Accessed 18 July 2023).

[DCK23] S. Dustdar, V. Casamayor, and P. Kumar, "On Distributed Computing Continuum Systems," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp: 4092-4105, 2023.

[DED-D2.4] Dedicat6G Deliverable D2.4 "Revised System Architecture," June 2022.

[DMB+20] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-Objective Counterfactual Explanations," in Parallel Problem Solving from Nature – PPSN XVI, T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, and H. Trautmann, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 448–469. doi: 10.1007/978-3-030-58112-1_31.

[DSW+21] A. Derrow-Pinion, J. She, D. Wong, et al., "ETA Prediction with Graph Neural Networks in Google Maps", https://arxiv.org/abs/2108.11482

[EMC23] Project EMCO, https://project-emco.io/

[eni-005] ETSI GR ENI 005, "Experiential Networked Intelligence (ENI); System Architecture", v2.1.1, December 2021

[eni-008] ETSI GR ENI 008, “Experiential Networked Intelligence (ENI); InTent Aware Network Autonomicity (ITANA)”, v2.1.1, March 2021
[eni-009] ETSI GR ENI 009, “Experiential Networked Intelligence (ENI); Definition of data processing mechanisms”, v1.2.1, May 2023
[eni-010] ETSI GR ENI 010 (V1.1.1) “Experiential Networked Intelligence (ENI); Evaluation of categories for AI application to Networks”, March 2021.
[eni-012] ETSI GR ENI 012, “Experiential Networked Intelligence (ENI); Reactive In-situ Flow Information Telemetry”, v1.1.1, March 2022
[eni-013] ETSI GR ENI 013, “Experiential Networked Intelligence (ENI); Intent Policy Model Gap Analysis”, v1.1.1, January 2023
[eni-017] ETSI GR ENI 017 (V2.1.1). “Experiential Networked Intelligence (ENI); Overview of Prominent Control Loop Architectures”, August 2021.
[eni-019] ETSI GR ENI 019, “Experiential Networked Intelligence (ENI); Representing, Inferring, and Proving Knowledge in ENI”, v3.1.1, June 2023.
[ENV23] Envoy proxy https://www.envoyproxy.io/docs/envoy/v1.26.2/intro/what_is_envoy
[ESA23] European Space Agency, “ANChOR - Data-driven Network Controller and Orchestrator for Real-time Network Management,” 2023. [Online]. Available: https://artes.esa.int/projects/anchor
[Eta19] L. Etaati, “Machine Learning with Microsoft Technologies: Selecting the Right Architecture and Tools for Your Project,” APress:Aukland, New Zealand, 2019.
[FEP22] Y. Fang, S. Ergüt, and P. Patras, “SDGNet: A handover-aware spatiotemporal graph neural network for mobile traffic forecasting,” IEEE Communications Letters, vol. 26, no. 3, pp. 582-586, 2022.
[FOL+23] T. Faisal, J. Ordoñez, D. Lopez, C. Wang, and M. Dohler "How to Design Autonomous Service Level Agreements for 6G," in IEEE Communications Magazine, vol. 61, no. 3, pp. 80-85, 2023.
[FPS+23] M. Ferriol-Galmés, J. Paillisse, J. Suárez-Varela, K. Rusek, S. Xiao, X. Shi, X. Cheng, P. Beret-Ros, and A. Cabellos-Aparicio, “RouteNet-Fermi: Network Modeling With Graph Neural Networks,” IEEE/ACM Transactions on Networking, 2023.
[FRS+22] M. Ferriol-Galmés, K. Rusek, J. Suárez-Varela, S. Xiao, X. Shi, X. Cheng, B. Wu, P. Barlet-Ros, and A. Cabellos-Aparicio, "RouteNet-Erlang: A Graph Neural Network for Network Performance Evaluation", IEEE INFOCOM, 2022.
[FSP+22] M. Ferriol-Galmés, J. Suárez-Varela, J. Paillissé, X. Shi, S. Xiao, X. Cheng, P. Barlet-Ros, and A. Cabellos-Aparicio, "Building a Digital Twin for network optimization using Graph Neural Networks,” Computer Networks, vol. 217, 2022.
[GA19] D. Gunning, and D. Aha, “DARPA’s explainable artificial intelligence (XAI) program,” AI magazine, vol. 40, no. 2, pp. 44-58, 2019.
[GMM+21] F. Guim, T. Metsch, H. Moustafa, T. Verrall, D. Carrera, N. Cadenelli, J. Chen, D. Doria, C. Ghadie, and R. Prats, "Autonomous Lifecycle Management for Resource-Efficient Workload Orchestration for Green Edge Computing." in IEEE Transactions on Green Communications and Networking, vol. 6, no.1, pp. 571-582, 2021.
[GNM23] gNMI gRPC Network Management Interface, 2023. [online]. Available: https://github.com/openconfig/gnmi
[GRP23] gRPC, 2023. https://grpc.io/
[GSM23] GSMA, “ESG Metric for mobile,” June 2022. [Online]. Available: https://www.gsma.com/betterfuture/wp-content/uploads/2023/02/ESG-Metrics-for-Mobile-February-2023.pdf
[HAS23] HashiCorp, Consul service mesh. https://developer.hashicorp.com/consul/docs/concepts/service-mesh
[HEX21-D12] Hexa-X, “Deliverable D1.2: Expanded 6G vision, use cases and societal values – including aspects of sustainability, security and spectrum,” Apr. 2021.
[HEX223-D11] Hexa-X-II, “Deliverable D1.1 Environmental, social, and economic drivers and goals for 6G”, June 2023.
[HEX223-D21] Hexa-X-II, “Deliverable D2.1: Draft foundation for 6G system design”, June 2023.

[HEX223-D22] Hexa-X-II, “Deliverable D2.2: Foundation of overall 6G system design and preliminary evaluation results”, December 2023.
[HEX223-D32] Hexa-X-II, “Deliverable D3.2: Initial Architectural enablers”, To be published Nov. 2023.
[HEX224-D63] Hexa-X-II, “Deliverable D6.3: Initial Design of 6G Smart Network Management Framework”, to be published in July 2024.
[HEX22-D62] Hexa-X, “Deliverable D6.2: Design of service management and orchestration functionalities”, Apr. 2022.
[HEX23-D14] Hexa-X, “Hexa-X architecture for B5G/6G networks – final release”, July 2023.
[HEX23-D63] Hexa-X, “Deliverable D6.3: Final evaluation of service management and orchestration mechanisms”, Apr. 2023.
[HEX23-D73] Hexa-X, “Special-purpose functionalities: final solutions”, May 2023.
[HOR23] HORSE project website, 2023. [Online]. Available: https://www.horse-6g.eu/
[HSW+21] Y. Han, S. Shen, X. Wang, S. Wang, and V. C. Leung, “Tailored Learning-Based Scheduling for Kubernetes-Oriented Edge-Cloud System,” IEEE INFOCOM 2021 - IEEE Conference on Computer Communications, 1-10, 2021.
[IG1219A] TMForum, “IG1219A AI Closed Loop Automation Management,” v1.3.0, Dec. 2022.
[IG1219F] TMForum, “IG1219F Closed Loop Information Model v1.0.0,” Jan. 2023.
[IG1305] TMForum, “Autonomous Networks – Empowering digital transformation – from strategy to implementation,”, Aug 2022.
[IG1307] TMForum, “DT4DI – Digital Twin for decision intelligence,” Dec 2022.
[IJR+23] M. Iovene, L. Jonsson, D. Roeland et al., « Defining AI native: A key enabler for advanced intelligent telecom networks,” https://www.ericsson.com/49341a/assets/local/reports-papers/white-papers/ai-native.pdf
[IMT+23] W. Issa, N. Moustafa, B. Turnbull, N. Sohrabi, and Z. Tari, “Blockchain-Based Federated Learning for Securing Internet of Things: A Comprehensive Survey,” ACM Comput. Surv. 55, 9, Article 191, September 2023. [Online]. Available: https://doi.org/10.1145/3560816
[INS20] H2020 project INSPIRE-5Gplus https://www.inspire-5gplus.eu/
[IST23] Istio. https://www.solo.io/topics/istio/istio-architecture/
[JKF+19] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, “Explainable reinforcement learning via reward decomposition,” in IJCAI/ECAI Workshop on explainable artificial intelligence, 2019.
[KAR23] karmada, 2023. [Online]: Available: https://karmada.io
[KAT23] Kata Containers, 2023. [Online]: Available: https://www.katacontainers.io
[KEP23] Kubernetes Efficient Power Level Experter (Kepler), 2023. [Online]: Available: https://sustainable-computing.io/
[KOH22] DCAE R11 Kohn M2 Architecture Review [Online], Available at: https://wiki.onap.org/display/DW/DCAE+R11+Kohn+M2+Architecture+Review (Accessed 17 July 2023).
[KSE22] Kubeflow. Kserve [Online]. Available: https://www.kubeflow.org/docs/external-add-ons/kserve/kserve/ (Accessed 21/07/2023)
[KUB23] Kubeflow: The Machine Learning Toolkit for Kubernetes [Online]. Available at: https://www.kubeflow.org/ (Accessed 21/07/2023)
[Lab21] Mark Labbe, “Energy consumption of AI poses environmental problems,” 2021. [Online]. Available: https://www.techtarget.com/searchenterpriseai/feature/Energy-consumption-of-AI-poses-environmental-problems
[LCA+22] I. Leyva-Pupo, C. Cervelló-Pastor, C. Anagnostopoulos, D. P. Pezaros, "Dynamic UPF placement and chaining reconfiguration in 5G networks", Computer Networks, Volume 215, 9 October 2022.
[LFN23] Linux Foundation [Online], Available at: https://www.linuxfoundation.org/ (Accessed 17 July 2023).
[LIN23] Linkerd https://linkerd.io/2.9/reference/architecture/
[LIQ23] Liqo: https://liqo.io
[LL17] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in neural information processing systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017.

[LLS+19] A. Lacoste, A. Luccioni, V. Schmidt, T. Dandres, "Quantifying the Carbon Emissions of Machine Learning", arXiv, 2019. [Online]. Available: https://arxiv.org/abs/1910.09700v2

[LLY+23] J. Li, F. Lin, L. Yang and D. Huang, "AI Service Placement for Multi-Access Edge Intelligence Systems in 6G," in IEEE Transactions on Network Science and Engineering, vol. 10, no. 3, pp. 1405-1416, 1 May-June 2023.

[LON23] DCAE Release Notes. Version: 12.0.0 [Online], Available at: https://docs.onap.org/projects/onap-dcaegen2/en/london/sections/version_12.0.0.html (Accessed 17 July 2023).

[LSZ+19] G. Liu, O. Schulte, W. Zhu, and Q. Li, "Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees," in Machine Learning and Knowledge Discovery in Databases, M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim, Eds., in Lecture Notes in Computer Science, vol. 11052. Cham: Springer International Publishing, 2019, pp. 414–429. doi: 10.1007/978-3-030-10928-8_25.

[MCM+21] F. Malandrino, C. F. Chiasserini, N. Molner and A. de la Oliva, "Network Support for High-Performance Distributed Machine Learning," in IEEE/ACM Transactions on Networking, vol. 31, no. 1, pp. 264-278, Feb. 2023.

[mec-003] ETSI GS MEC 003, "Multi-access Edge Computing (MEC); Framework and Reference Architecture", v3.1.1, March 2022.

[mec-015] ETSI GS MEC 015, "Multi-Access Edge Computing (MEC); Traffic Management APIs", 2020.

[mec-040] ETSI GS MEC 040, "Multi-access Edge Computing (MEC); Federation enablement APIs", v3.1.1, February 2023.

[mec-fed-22] M. Suzuki et al., ETSI White Paper No. #49, "MEC Federation; Deployment Considerations", 1st version, June 2022.

[mec-sec-22] D. Sabella et al., ETSI White Paper No. #46, "MEC Security; Status of standards support and future evolutions", 2nd edition, September 2022

[MGK+22] A. Mesodiakaki, M. Gatzianas, G. Kalfas, C. Vagionas, R. Maximidis and N. Pleros, "ONE: Online Energy-efficient User Association, VNF Placement and Traffic Routing in 6G HetNets," 2022 IEEE Globecom Workshops (GC Wkshps), Rio de Janeiro, Brazil, 2022, pp. 304-309

[MRM20] F. Manessi, A. Rozza, and M. Manzo, "Dynamic graph convolutional networks," Pattern Recognition, vol. 97, pp. 107000, 2020.

[MWW+17] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Briefings in bioinformatics, vol. 19, no. 6, pp. 1236–1246, November 2018.

[NAT23] NATS.io https://docs.nats.io/nats-concepts/what-is-nats

[NET23] IETF Network Modelling (OPSAWG) [Online]. Available: https://datatracker.ietf.org/wg/opsawg/documents/

[NFV21] ETSI, "NFV Release 4 Definition", ETSI, 2021

[NFV23] ETSI NFV web site. 2023. Available at: https://www.etsi.org/technologies/nfv [Accessed 05 April 2023]

[NFV5] NFV Release 5 Description. 2023. Available at: https://docbox.etsi.org/ISG/NFV/Open/Other/ReleaseDocumentation/NFV(22)000172_NFV_Release_5_Description_v0_0_2.pdf [Accessed 05 April 2023]

[NFV5-SBA] First deliverable for NFV Release 5 feature: applying SBA design style to NFV-MANO. 2023. Available at: https://www.etsi.org/newsroom/blogs/entry/first-deliverable-for-nfv-release-5-feature-applying-sba-design-style-to-nfv-mano-1 [Accessed 05 April 2023]

[NFV6] NFV#40, preparing for NFV next decade by planning Release 6. 2023. Available at: https://www.etsi.org/newsroom/blogs/technologies/entry/preparing-for-nfv-next-decade-by-planning-release-6 [Accessed 05 April 2023]

[NG135] GSMA NG.135, "E2E Network Slicing Requirements", version 1.0, July 2022

[NMR23] Network Management Research Group, 2023. [Online]. Available: https://irtf.org/nmrg.html

[OCM23] Open Cluster Management, https://open-cluster-management.io/

[OD22] J Ordonez-Lucena and Felix Dsouza, “Pathways towards Network-as-a-Service: the CAMARA Project”, 2022 Workshop on Network-Application Integration (NAI ’22), August 22, 2022, Amsterdam, Netherlands

[OGW23] GSMA, CAMARA, Linux Foundation Networking and TM Forum, “The Ecosystem for Open Gateway NaaS API Development”, June 2023 [Online]. Link: https://www.gsma.com/futurenetworks/wp-content/uploads/2023/05/The-Ecosystem-for-Open-Gateway-NaaS-API-development.pdf

[ONA23] ONAP web site [Online]. Available at: https://docs.onap.org/en/london/ (Accessed 17 July 2023).

[ONAb23] ONAP architecture overview [Online], Available at: https://docs.onap.org/en/london/ (Accessed 17 July 2023).

[OP23] OpenConfig, https://www.openconfig.net/

[OPE23] Openslice https://openslice.readthedocs.io/en/stable/architecture/architecture/

[OPS23] IETF Operations and Management Area Working Group (OPSAWG) [Online]. Link: https://datatracker.ietf.org/wg/opsawg/documents/

[OPT23] OpenTelemetry, https://opentelemetry.io/, 2023.

[ORA22] Confluence, “G Release,” 2022. [Online]. Available at: https://wiki.o-ran-sc.org/display/REL/G+Release/ (Accessed 21/07/2023)

[ORA23] O-RAN Architecture Overview [Online]. Available at: https://docs.o-ran-sc.org/en/h-release/architecture/architecture.html (Accessed 21/07/2023)

[ORAb23] O-RAN web site [Online]. Available: https://www.o-ran.org/ (Accessed 19 July 2023).

[ORAc23] O-RAN SC H Release Documentation [Online]. Available at: https://docs.o-ran-sc.org/en/h-release/ (Accessed 21/07/2023).

[OSM-12] OSM Release TWELVE: Release Notes. 2023. [Online] Available at: https://osm-download.etsi.org/ftp/osm-12.0-twelve/OSM_Release_TWELVE_Release_Notes.pdf [Accessed 05 April 2023].

[OSM-13] OSM Release THIRTEEN: Release Notes. 2023. [Online] Available at: https://osm-download.etsi.org/ftp/osm-13.0-thirteen/OSM_Release_THIRTEEN_Release_Notes.pdf [Accessed 05 April 2023].

[OSM23] ETSI Open-Source MANO: OSM. 2023. [Online] Available at: https://osm.etsi.org/ [Accessed 05 April 2023].

[OTR20] J. Ordonez-Lucena, C. Tranoris and J. Rodrigues, "Modeling Network Slice as a Service in a Multi-Vendor 5G Experimentation Ecosystem," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, pp. 1-6, 2020.

[Pat22] D. Patterson. Good News About the Carbon Footprint of Machine Learning Training. Feb’22 https://ai.googleblog.com/2022/02/good-news-about-carbon-footprint-of.html

[pdl-001] ETSI GR PDL 001: “Permissioned Distributed Ledger (PDL); Landscape of Standards and Technologies”, September 2021.

[pdl-011] ETSI PDL 011, “Specification of Requirements for Smart Contracts’ architecture and security”, Sept 2022.

[PMS+18] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, ‘‘SoK: Security 2002 and privacy in machine learning,’’ in Proc. IEEE Eur. Symp. Secur. 2003 Privacy, pp. 399–414, Apr. 2018.

[PMW+16] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in 2016 IEEE Symposium on Security and Privacy (SP), pp. 582–597, 2016.

[PRI23] PRIVATEER project, SNS. [Online]. Link: https://www.privateer-project.eu/

[PRO23] Prometheus: Monitoring Platform. Available at: https://prometheus.io/docs/introduction/overview/ [Accessed 05 April 2023]

[PSX+20] W. Peng, J. Shi, Z. Xia, and G. Zhao, “Mix dimension in poincaré geometry for 3d skeleton-based action recognition,” In Proceedings of the 28th ACM International Conference on Multimedia, pp. 1432-1440, October 2020.

[PUL22] A. Paleyes, R. G. Urma, and N. D. Lawrence, “Challenges in Deploying Machine Learning: A Survey of Case Studies,” ACM Comput. Surv., vol. 55, no. 6, Article 114, pp. 1—29, 2022.

[PV20] E. Puiutta and E. M. S. P. Veith, “Explainable Reinforcement Learning: A Survey,” in Machine Learning and Knowledge Extraction, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E.

Weippl, Eds., in Lecture Notes in Computer Science, vol. 12279. Cham: Springer International Publishing, 2020, pp. 77–95. doi: 10.1007/978-3-030-57321-8_5.
[RAB23] RabbitMQ https://www.rabbitmq.com/access-control.html
[RBO21] J. Rico, J. Barateiro, A. Oliveira, "Graph Neural Networks for Traffic Forecasting", arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2104.13096
[RCV22] S Roy, H. Chergui, and C. Verikoukis, "TEFL: Turbo Explainable Federated Learning for 6G Trustworthy Zero-Touch Network Slicing," arXiv preprint, arXiv:2210.10147, 2022.
[RFC6632] IETF RFC 6632: "An Overview of the IETF Network Management Standards" [Online]. Link: https://www.rfc-editor.org/rfc/rfc6632
[RFC7276] IETF RFC 7276: "An Overview of Operations, Administration and Maintenance Tools" [Online]. Link: https://www.rfc-editor.org/rfc/rfc7276
[RFC7951] IETF RFC 7951:" RFC7951 Encoding of Data Modelled with YANG" [Online]. Link: https://datatracker.ietf.org/doc/rfc7951
[RFC8343] IETF RFC 8343: "A YANG Data Model for Interface Management" [Online]. Link: https://datatracker.ietf.org/doc/rfc8343
[RFC8344] IETF RFC 8343: "A YANG Data Model for IP Management" [Online]. Link: https://datatracker.ietf.org/doc/rfc8344
[RFC8348] IETF RFC 8343: "A YANG Data Model for Hardware Management" [Online]. Link: https://datatracker.ietf.org/doc/rfc8348
[RFC8349] IETF RFC 8343: "A YANG Data Model for Routing Management" [Online]. Link: https://datatracker.ietf.org/doc/rfc8349
[RFC8453] IETF RFC 8453, Framework for Abstraction and Control of TE Networks (ACTN), 2018.
[RFC8795] IETF RFC 8795, YANG Data Model for Traffic Engineering (TE) Topologies, 2020.
[RFC8821] IETF RFC 8821, PCE-Based Traffic Engineering (TE) in Native IP Networks, 2021.
[RFC8969] IETF RFC 8969: "A Framework for Automating Service and Network Management with YANG" [Online]. Link: https://www.rfc-editor.org/rfc/rfc8969
[RFC9182] IETF RFC 9182: "A YANG Network Data Model for Layer-3 VPNs" [Online]. Link: https://www.rfc-editor.org/rfc/rfc9182
[RFC9232] IETF RFC 9232: "Network Telemetry framework" [Online]. Link: https://www.rfc-editor.org/rfc/rfc9232
[RFC9291] IETF RFC 9291: "A YANG Network Data Model for Layer-2 VPNs" [Online]. Link: https://www.rfc-editor.org/rfc/rfc9291
[RFC9375] IETF RFC 9375: "A YANG Data Model for Network and VPN Service Performance Monitoring" [Online]. Link: https://www.rfc-editor.org/rfc/rfc9375
[RIG23] https://rigourous.eu/
[RRA+21] M. A. Ridwan, N. A. M. Radzi, F. Abdullah and Y. E. Jalil, "Applications of Machine Learning in Networking: A Survey of Current Issues and Future Challenges," in IEEE Access, vol. 9, pp. 52523-52556, 2021
[RSG+18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11491.
[SAI23] https://www.etsi.org/technologies/securing-artificial-intelligence
[SCA23] https://github.com/hubblo-org/scaphandre
[SGC19] E Strubell, A Ganesh, A McCallum, "Energy and Policy Considerations for Deep Learning in NLP", https://arxiv.org/abs/1906.02243
[SGT+09] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, "The Graph Neural Network Model," in IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61-80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.
[SHR20] T. Subramanya, D. Harutyunyan, R. Riggio, "Machine learning-driven service function chain placement and scaling in MEC-enabled 5G networks", Computer Networks, Volume 166, 15 January 2020
[SKK+22] E. U. Soykan, L. Karaçay, F. Karakoç, and E. Tomur, "A Survey and Guideline on Privacy Enhancing Technologies for Collaborative Machine Learning," IEEE Access, 10, 97495-97519, 2022.

[SMO+20] H. Sami, A. Mourad, H. Otrok and J. Bentahar, "FScaler: Automatic Resource Scaling of Containers in Fog Clusters Using Reinforcement Learning," 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 2020, pp. 1824-1829, 2020 doi: 10.1109/IWCMC48107.2020.9148401.

[SYL22] White Paper Operators Sylva (2022)

[TAP22] TAPI v2.4.0 / v2.4.1 Reference Implementation Agreement - Streaming (TR-548 V2.0), December 2022.

[TAP23] TAPI v2.4.1 Reference Implementation Agreement (TR-547 V2.1), March 2023.

[TIB+20] A. Terra, R. Inam, S. Baskaran, P. Batista, I. Burdick, and E. Fersman, “Explainability Methods for Identifying Root-Cause of SLA Violation Prediction in 5G Network,” in GLOBECOM 2020 - 2020 IEEE Global Communications Conference, Taipei, Taiwan: IEEE, pp. 1–7, Dec. 2020. doi: 10.1109/GLOBECOM42002.2020.9322496.

[TIF22] A. Terra, R. Inam, and E. Fersman, “BEERL: Both Ends Explanations for Reinforcement Learning,” Applied Sciences, vol. 12, no. 21, p. 10947, Oct. 2022, doi: 10.3390/app122110947.

[TIP21] TIP, Open Transport SDN Architecture Whitepaper, 2021.

[TMF23] TM Forum, 2023. [Online]. Available: https://www.tmforum.org/

[TMF-IG1230] IG1230 Autonomous Networks Technical Architecture, 2021. https://www.tmforum.org/resources/how-to-guide/ig1230-autonomous-networks-technical-architecture-v1-0-0/

[TOM+20] O. Tomarchio, D. Calcaterra, G. D. Modica, “Cloud resource orchestration in the multi-cloud landscape: a systematic review of existing frameworks,” J Cloud Comp vol. 9, no. 49, 2020. https://doi.org/10.1186/s13677-020-00194-7

[TRW23] https://www.itu.int/en/ITU-T/focusgroups/ai4ee/Documents/TR-D.WG2_03-Requirements%20on%20EE%20measurement%20models%20and%20the%20role%20of%20AI%20and%20big%20data_Anthopoulos.pdf

[TSK+22] P. Tam, I. Song, S. Kang, S. Ros, and S. Kim, “Graph Neural Networks for Intelligent Modelling in Network Management and Orchestration: A Survey on Communications,” Electronics, vol. 11, no. 20, pp. 3371, 2022. https://doi.org/10.3390/electronics11203371

[TST+22] A. Taneja, N. Saluja, N. Taneja, A. Alqahtani, M. A. Elmagzoub, A. Shaikh, and D. Koundal, “Power Optimization Model for Energy Sustainability in 6G Wireless Networks,” Sustainability. 2022

[TUN23] Tungsten Fabric, https://tungsten.io/

[TZA22] I. Tzanettis, C. M. Androna, A. Zafeiropoulos, E. Fotopoulou, and S. Papavassiliou, “Data Fusion of Observability Signals for Assisting Orchestration of Distributed Applications,” Sensors, vol. 22, no. 5, pp. 2061, 2022. https://doi.org/10.3390/s22052061

[UER+21] M. A. Uusitalo, M. Ericson B. Richerzhagen, E. U. Soykan, P. Rugeland, G. Fettweis, D. Sabella, G. Wikström, M. Boldi, M. H. Hamon, H. D. Schotten, V. Ziegler, M. Latva-aho, P. Serrano, Y. Zou, G. Carrozzo, J. Martrat, G. Stea, P. Demestichas, A. Pärssinen, and T. Svensson, "Hexa-X The European 6G flagship project," 2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Porto, Portugal, 2021, pp. 580-585, doi: 10.1109/EuCNC/6GSummit51104.2021.9482430.

[UQA+18] M. Usama, J. Qadir, and A. Al-Fuqaha, "Adversarial Attacks on Cognitive Self-Organizing Networks: The Challenge and the Way Forward," 2018 IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops), Chicago, IL, USA, 2018, pp. 90-97, doi: 10.1109/LCNW.2018.8628538.

[VCC+17] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” arXiv preprint arXiv:1710.10903, 2017.

[VFM99] M. Van Lent, W. Fisher, and M. Mancuso, “An explainable artificial intelligence system for small-unit tactical behavior,” in Proceedings of the national conference on artifi[1]cial intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907. [3] D.

[VHI+21] K. Vandikas, H. Hallberg, S. Ickin, C. Nyström, E. Sanders, O. Gorbatov, L. Eleftheriadis. “Using AI to Ensure Energy-Efficient Networks”

https://www.ericsson.com/4972d5/assets/local/reports-papers/ericsson-technology-review/docs/2021/ensuring-energy-efficient-networks-with-ai.pdf

[VMC+21] R. Vilalta, R. Muños, R. Casellas, R. Martinez, et al., Teraflow : Secured autonomic traffic management for a tera of sdn flows, 2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit).

[VSD+21] L. Velasco, M. Signorelli, O. Gonzales De Dios, et al., "End-to-End Intent-Based Networking", IEEE Communications Magazine, vol. 59, no. 10, pp. 106-112, 2021. doi: 10.1109/MCOM.101.2100141

[WFC+18] Y. Wang, R. Forbes, C. Cavigioli, H. Wang, A. Gamelas, A. Wade, and S. Liu, "Network management and orchestration using artificial intelligence: Overview of ETSI ENI," IEEE communications standards magazine, 2(4), 58-65, 2018.

[WJH+21] Z. Wu, D. Jiang, C. Y. Hsieh, G. Chen, B. Liao, D. Cao, and T. Hou, "Hyperbolic relational graph convolution networks plus: a simple but highly efficient QSAR-modeling method," Briefings in Bioinformatics, vol. 22, no. 5, 2021.

[WLT+21] H. Wang, D. Lian, H. Tong, Q. Liu, Z. Huang, and E. Chen, "HyperSoRec: Exploiting hyperbolic user and item representations with multiple aspects for social-aware recommendation," ACM Transactions on Information Systems (TOIS), vol. 40, no. 2, pp. 1-28, 2021.

[WWM+20] H. Wang, Y. Wu, G. Min, and W. Miao, "A graph neural network-based digital twin for network slicing management," IEEE Transactions on Industrial Informatics, vol. 18, no.2, pp. 1367-1376, 2020.

[XB20] M. Xu, R. Buyya, "Managing renewable energy and carbon footprint in multi-cloud computing environments," Journal of Parallel and Distributed Computing, vol. 135, 2020

[XLC+22] D. Xu, F. Liu, W. Chen, F. He, X. Tang, Y. Zhang, and B. Wang, "A review of research on multi-cloud management platforms," ISCTT 2022; 7th International Conference on Information Science, Computer Technology and Transportation, Xishuangbanna, China, 2022, pp. 1-16.

[YAX+20] H. Yang, A. Alphones, Z. Xiong, et al., "Artificial-Intelligence-Enabled Intelligent 6G Networks," IEEE Network, vol. 34, no. 6, pp. 272-280, November/December 2020

[zsm-002] ETSI GS ZSM 002, "Zero-touch network and Service Management (ZSM); Reference Architecture". August 2019.

[zsm-008] ETSI GS ZSM 008, "Zero-touch network and Service Management (ZSM); Cross-domain E2E service lifecycle management", v1.1.1, July 2022.

[zsm-009-1] ETSI GS ZSM 009-01, "Zero-touch network and Service Management (ZSM); Closed loop Automation; Part 1: Enablers", v1.1.1. June 2021.

[zsm-009-2] ETSI GS ZSM 009-02, "Zero-touch network and Service Management (ZSM); Closed loop Automation; Part 2: Solutions for automation of E2E service and network management use cases", v1.1.1, June 2022.

[zsm-009-3] ETSI GR ZSM 009-03, "Zero-touch network and Service Management (ZSM); Closed loop Automation; Part 3: Advanced Topics", August 2023.

[zsm-010] ETSI GR ZSM 010, "Zero-touch network and Service Management (ZSM);. General Security Aspects"

[zsm-011] ETSI GR ZSM 011, "Zero-touch network and Service Management (ZSM); Intent-driven autonomous networks; Generic aspects", v1.1.1 February 2023.

[zsm-012] ETSI GS ZSM 012, "Zero-touch network and Service Management (ZSM); Enablers for Artificial Intelligence-based Network and Service Automation", v1.1.1, December 2022.

[zsm-013] ETSI GS ZSM 013, "Zero-touch network and Service Management (ZSM); Automation of CI/CD for ZSM services and managed services". 2022. Available at: https://portal.etsi.org/webapp/WorkProgram/Report_WorkItem.asp?WKI_ID=61979 [Accessed 06 of April 2023]

[zsm-014] ETSI GR ZSM 010, "Zero-Touch network and Service Management (ZSM), General Security Aspects". July 2021.

[zsm-poc2] ETSI ZSM, "Proof-of-Concept 2, Automated network slice scaling in multi-site environments," https://zsmwiki.etsi.org/images/8/84/ZSM_POC_2_Final_Report.pdf

[zsm-poc6] ETSI ZSM “Proof-of-Concept 6, Security SLA assurance in 5G network slices,” https://zsmwiki.etsi.org/images/e/e1/ZSM_POC_6.pdf

[ZYD+23] C. Zhou, H. Yang, X. Duan, D. Lopez, A. Pastor, Q. Wu, M. Boucadair, and C. Jacquenet, "Digital Twin Network: Concepts and Reference Architecture, draft-irtf-nmrg-network-digital-twin-arch-03,” 2023.

[ZYF+22] J. Zhang, P. Yu, L. Feng, W. Li, M. Zhao, X. Yan, and J. Wu, “Fine-Grained Service Offloading in B5G/6G Collaborative Edge Computing Based on Graph Neural Networks,” ICC 2022 - IEEE International Conference on Communications, 5226-5231, 2022.

7 Annex: State of the Art

An exhaustive review of the most relevant State of the Art (SoTA) for Hexa-X-II was presented in [Hexa-X-D61]. That document does not include information since 2021, so this section provides updated information on the current relevant SoTA initiatives on network management & Orchestration (M&O) from 2021 to mid-2023. New relevant initiatives have also been considered. The SoTA overview is structured into the following subsections: i) Standard Development Organizations (SDO) and Open-Source communities, ii) Industry fora, and iii) Research projects.

7.1 Standard Development Organizations and Open-Source Communities

In this section we analyse the work of multiple SDO and Open-Source Software that are related or could impact on the development of the 6G technology.

7.1.1 3GPP

3GPP is a consortium with seven national or regional telecommunication standards organizations as primary members ("organizational partners") and a variety of other organizations as associate members ("market representation partners"). The 3GPP organizes its work into three different streams: Radio Access Networks, Services and Systems Aspects (SA), and Core Network and Terminals. Below the status of the relevant groups for the work in this deliverable are presented: SA2, SA3, SA5 and SA6.

7.1.1.1 SA2/Network data analysis functions and data repositories

The 3GPP SA2 Working Group is responsible for specifying the system architecture, functionalities, and interfaces for the (currently 5G) network architecture. The 3GPP standards define the Network Data Analytics Function (NWDAF) function in the 5G architecture which is in charge of training ML models based on data collected from different functions in the architecture and providing data analytics [23.288]. The 3GPP 5G architecture allows the NWDAF to collect data from and provide analytics to any 5G Core Network Function, and multiple NWDAF instances can be deployed in a layered hierarchy, which is deployment specific.

In the 17th Release of the 3GPP standards, two logical functions were introduced for the NWDAF:

- The Analytics Logical Function (AnLF), which is in charge of performing inference and exposing analytics and predictions.
- The Model Training Logical Function (MTLF), which trains ML models, and provides them to the AnLF.

The current work of the SA2 Working Group in the Release 18 of 3GPP [23.700-80,23.700-81] is exploring NWDAF analytics exposure to the User Equipment (UE), authorized 3rd party AFs, or to NWDAFs in different PLMNs in case of roaming. However, user data privacy issues are being raised, and more work on security aspects is required.

Further functionalities for the NWDAF are also discussed, such as assisting application traffic detection in the UPF, and assisting the PCF in generating UE Route Selection Policy (URSP) rules. New Application Data Transfer (ADT) policies are also introduced to differentiate and apply specific Quality of Service (QoS) rules to the Application AI/ML traffic transport.

7.1.1.2 SA5/Management system and network slice management

SA5 is the 3GPP working group responsible for:

- Management and Orchestration, which covers aspects such as operation, assurance, fulfillment and automation, including management interaction with entities external to the network operator (e.g., verticals). This activity is conducted by SA5 OAM sub-group.

- Charging, which covers aspects such as Quota Management and Charging Data Records (CDRs) generation, related to end-user and service-provider. This activity is conducted by SA5 CH sub-group.

As for Hexa-X-II, the line of work that is most interesting is SA5 OAM, which at the time of writing, is closing Rel-18 work, and starts drafting the agenda and time plan for Rel-19 work. As for the Rel-18 work, SA5 OAM structures the topics into three families: i) **intelligence and automation**, aiming at progressing towards a data-driven, zero-touch network and service operation; ii) **management architecture and mechanisms**, capturing all the system architecture aspects and capabilities for the management of 5G network resources and associated services; iii) **support of new services**, including all the value-added capabilities that allow defining beyond-connectivity services, accompanied with new revenue streams for network operators. As for the Rel-19 plan, 3GPP SA5 has started working out the timing in conjunction with the rest of 3GPP groups. The set of topics to be developed in Rel-19 is still under discussion, and the final drop is expected to be agreed by the end of this year.

7.1.1.3 SA3/Security and privacy

SA3 working group is responsible for the security and privacy aspects of the architectures designed by other working groups such as SA2 and SA6. The working group is also responsible for the availability of cryptographic algorithms, so specifies the crypto profiles to be used in 3GPP system. Some of the items that are currently studied and worked could be listed as follows:

- Security of network automation based on network data analytics function: Some of the main key issues studied are about authentication, authorization and secure transfer of data and ML models between network functions; authentication and authorization aspects in support of federated learning.
- Security of edge computing support over 3GPP system: SA3 is working on security and privacy aspects of architectures defined by both SA2 and SA6. In the SA2 defined architecture, security of edge application server discovery messages is focused. For the SA6 defined edge application layer enabler architecture, SA3 works on the solutions to protect the communication between newly introduced entities and on the authentication and authorization aspects.
- Security and privacy aspects of the enhancement in Common API Framework for 3GPP northbound APIs (CAPIF): SA6 is enhancing the 3GPP defined common Application Programming Interface (API) framework to make the framework subscriber-aware, meaning that if the API invocation is related to accessing UE data, which needs authorization from the user/subscriber of the UE, CAPIF core function or the authorization function needs to ensure that the user authorizes the API invocation before it issues an access token to the API invoker. Since the authentication and authorization are in the remit of SA3, SA3 is working on possible solutions to specify the mechanism required to have a subscriber-aware common API framework.

In addition to the listed topics above, SA3 working group currently have more than 20 topics including maintenance and enhancements of the features introduced in previous releases, and studies and works for the new features introduced in the current release.

7.1.1.4 SA6/Application Enablement and Critical Communication Applications

The SA6 Working Group of 3GPP specializes in the application layer specifications and provides the architecture and procedures for mission critical services, service frameworks, and vertical applications.

The SA6 WG is currently working on multiple items of interest from Hexa-X-II perspective, among which:

- *EdgeApp:* To support Edge applications in 5G systems, 3GPP standards define an architecture and a set of procedures for enabling UEs to access Edge Application Servers (EAS) that are managed by Edge Enabler Servers (EES) in Edge Data Networks (EDN) [23.558]. To ensure service continuity, the Application Context Relocation (ACR) procedure allows user context relocation between two EASs, this procedure can be triggered in case of detected or predicted UE mobility. The target EAS for ACR is selected through an EAS Discovery mechanism using detected or predicted UE position, Application Client characteristics, and specific application criteria for selecting the required target EAS such as EAS type, availability schedule, service area and required features. If no EAS matching

the criteria is discovered, the EES may trigger a new EAS instantiation by sending a request to the EAS management system.

- *Application Data Analytics Enablement (ADAE):* In Release 18 of the 3GPP specification, the Application Data Analytics Enablement (ADAE) [23.436] functional architecture has been introduced to provide a unified exposure interface for the data analytics services of the 3GPP domain to verticals. Data analytics collection procedures are also designed to allow the ADAE layer component to collect and correlate analytics data from the 5G Core, Operations, Administration and Maintenance (OAM), the Vertical Application Layer, and optionally EDNs, and expose them to the requesting client. The ADAE currently supports features such as (slice-specific) application performance and usage pattern analytics, service API analytics, and edge load analytics.
- *CAPIF:* To optimize support for northbound applications, the 3GPP standards have introduced a CAPIF [23.222] which provides mechanisms for the operation and discovery of service APIs from trusted 3rd party API providers including security-related mechanisms, charging, OAM, monitoring and logging, and policy configuration.

7.1.2 ETSI

The European Telecommunications Standards Institute (ETSI) is an independent, not-for-profit, standardization organization in the field of information and communications technologies. Industry Specification Groups (ISG) operate alongside traditional standards-making committees in a specific technology area. In this section, the latest work of several ISG is presented.

7.1.2.1 Zero touch network and Service Management (ZSM)

ETSI Zero touch network and Service Management (ZSM) is a ISG within ETSI, formed in December 2017, that focuses on developing an advanced framework for network and service management, using the latest advancements in automation and Artificial Intelligence (AI). A number of Proof-of-Concept implementations have been also implemented and reported to demonstrate the feasibility of ZSM concepts applied to mobile networks, applied to e.g., automated management and scaling of network slices, automated provisioning of Industry 4.0 services in public-private 5G network environments, SLA assurance and security management.

ETSI ZSM specifically targets the management of end-to-end services across multiple domains. In this context, [zsm-008] analyses the applicability of ZSM concepts to the management system defined by 3GPP and proposes a mapping between ZSM management services to their counterparts in 3GPP Core and RAN domains, as well as multi-technology transport domains. One of the key features of ETSI ZSM is its specifications regarding Close-Loop Automation solutions [zsm-009-1] [zsm-009-2], which enable networks to self-optimize and self-heal in response to changes in the network environment or service requirements. This is achieved through a combination of monitoring, AI-based analytics, policy-based decision making, and automated control actions, which together form a closed feedback loop that continuously evaluates and adjusts network and service parameters. An additional specification [zsm-009-3] has been recently released to analyse advanced topics related to closed loop automation, like cognitive closed loops, coordination of interdependent closed loops, dynamic composition of closed loops and intent-based closed loops, and to propose potential solutions for each of them. Some of these topics are considered in the context of enablers 10 and 11 (Sections 3.10 and 3.11). As such, and due to the early stage of the research, good opportunities are expected to impact the future versions of ETSI ZSM 009 specifications ([zsm-009-1] [zsm-009-2] [zsm-009-3]), e.g., with contributions on the coordination of cross-layer closed loops operating at the network and service layers, or on the orchestrated provisioning of closed loop functions as part of the network management procedures.

Another important standard developed by the ETSI ZSM ISG, is their Intent-Driven approach to network and service management [zsm-011], which allows operators to define their service and performance objectives in terms of high-level business goals, rather than low-level network configurations. This simplifies the process of service provisioning and enables operators to focus on delivering value to their customers, rather than on technical details. The latest version of the specification introduces interesting concepts related to conflict management, between different intents or in association with coexisting imperative operations, as well as the support of optional interfaces and operations to ensure trust in intent-driven autonomy, with notifications, testing and verification of intent outcomes. Finally, it is worth to note that the work on a new ETSI ZSM

Specification (the future ETSI GS ZSM-016) entirely dedicated to Intent-Driven Closed Loops has just started in the ISG, confirming the strong interest in this area.

Moreover, this ETSI ISG also incorporates Continuous Integration/Continuous Deployment (CI/CD) automation [zsm-013], which streamlines the process of deploying new services and updates and ensures consistency and reliability across different environments. This study has a great focus on version control and model's release compatibility between vendors and operating teams.

Additionally, ETSI ZSM places a strong emphasis on security management [zsm-010], providing a comprehensive set of tools and mechanisms for detecting, preventing, and mitigating security threats aiming at achieving zero-touch security controls at the operator M&O context. An additional specification on security aspects is currently under preparation (the future ETSI GS ZSM 014), with focus on trustworthiness, intra-domain and inter-domain access control, AI/ML models' robustness. AI presents a key enabler for zero-touch automation as demonstrated in [zsm-012], which outlines AI-based capabilities supporting management and orchestration activities automation. These capabilities are concentrated in areas like action, interoperation, governance, execution environment, and data (including data gathering and analytics).

7.1.2.2 Network Functions Virtualization

The ETSI ISG Network Functions Virtualization (NFV) [NFV23] was formed in 2012 with the aim of defining a standard framework for virtualizing Network Functions (NFs), allowing for more flexible, efficient and cost-effective network infrastructure. Since the latest release in October 2019, ETSI NFV Release 4, this ETSI ISG has not published any further specifications regarding newer releases except for a first ETSI NFV Release 5 whitepaper description [NFV5] and some short news regarding the integration of the SBA design style into NFV-MANO [NFV5-SBA] and ETSI NFV Release 6 potential standardization paths [NFV6].

7.1.2.3 Experiential Networked Intelligence

The Experiential Networked Intelligence (ENI) ISG is currently in the process of defining a comprehensive framework that involves the application of AI techniques and context-aware policies in a cognitive network management system. This framework is aimed at automating decisions regarding network orchestration, with the objective of increasing the level of automation in procedures for service provisioning, lifecycle management, operation, resource orchestration, and optimization. ENI complements the ETSI ZSM architecture by enabling the application of AI techniques to network management [eni-010]. The goal is to increase the level of automation in network operations by continuously collecting data from the network and applying ML algorithms to analyse this data. This would allow the network to adapt dynamically to changing conditions, and to provide better quality of service to end-users. In this direction, the ENI ISG has released a set of new specifications in 2022 and 2023, related to reactive in-situ flow information telemetry [eni-012], the definition of data processing mechanisms [eni-009], and information models, data models and ontologies to represent, infer and prove knowledge in ENI framework [eni-019]. All of them can be considered as enablers for closed loop network control and management, supported by ML and reasoning techniques.

ETSI ENI ISG strongly focus on intent management (ENI InTent Aware Network Autonomicity [eni-008]) proposes extensions to the ENI architecture for InTent Aware Network Autonomicity (ITANA), with mechanisms for translation and validation of intent policies, continuous assurance and conflict detection. ENI Intent Policy Model Gap Analysis [eni-013] provides a gap analysis of various intent policy models and a survey of the standardization work in that area, covering 3GPP SA5, ETSI NFV and ZSM, TM Forum ANP and IRTF NMRG. An additional specification is currently in progress (ETSI GR ENI 015) on processing and management of intent policies, addressing topics of conflict detection and resolution between different intent policies and the adoption of Knowledge Graph for managing intent policies and their lifecycle.

7.1.2.4 Multi-access Edge Computing

The ETSI Multi-access Edge Computing (MEC) ISG is defining a framework that allows operators to open their RAN edge domains towards third-parties, like application and content providers, in a secure and authorized manner to facilitate the delivery of novel services to mobile subscribers and vertical industries.

ETSI MEC ISG is currently at the stage 3 of the specification activities, with the consolidation of the MEC framework architecture and the finalization of REpresentational State Transfer (REST) APIs for a variety of MEC applications and services. In parallel, some additional research items are currently addressed. Particular focus is on MEC security, enhanced lifecycle management for cloud and NFV services to be applied in MEC environments, the support of mobile components, as well as management of consumer-owned cloud resources.

In the **security** area, ETSI MEC has released a recent white paper [mec-sec-22] analysing security use cases and requirements for MEC, related to infrastructure security, data protection and security, user security, as well as network and application security layers. Security in the area of MEC **federation** is specifically addressed, following the latest advancements of the MEC Architecture in ETSI MEC GS 003 [mec-003] towards inter-MEC communications and sharing of information and resources among different operators and edge computing service providers. MEC federation is indeed a key research topic. The white paper on deployment considerations for MEC federation [mec-fed-22] analyses the related business cases, deployment options, and technological implications, e.g., in terms of inter-MEC connectivity and multi-domain orchestration, while a dedicated set of MEC APIs has been recently released in [mec-040] for MEC federation enablement.

7.1.2.5 Permissioned Distributed Ledger

The ETSI ISG on Permissioned Distributed Ledger (PDL) analyses and provides the foundations for the operation of permissioned distributed ledgers, with the ultimate purpose of creating an open ecosystem of industrial solutions to be deployed by different sectors, fostering the application of these technologies, and therefore contributing to consolidate the trust and dependability on information technologies supported by global, open telecommunications networks. The group puts its focus on addressing infrastructure and operational aspects that had not been covered by previous or parallel standardization activities.

The ISG PDL started from already available experiences in the field of permissioned distributed ledgers, seeking for: i) the definition of open and well-known operational mechanisms to validate participant nodes; ii) support the automation of the lifecycles of the ledger and individual nodes; iii) publish and execute operations regarding the recorded transactions through smart contracts; iv) improve security of ledgers during both their design and operation, and v) establish trusted links among different ledgers using these mechanisms. After an initial phase where the goals of PDL activities were delimited and the general DLT landscape analyzed in pdl-001 [pdl-001], different documents have been produced, of three different natures: informative (studies and recommendations for further work), normative (specifications) and demonstrative (in the form of proof-of concept reports and interoperability assessment events).

One of the most important documents is the PDL reference architecture [pdl-011], which provides the foundation framework to elaborate on capabilities, features and PDL application services. It consists of three layers:

- **DLT Layer**: This layer includes various DLT networks (e.g., an implementation of a specific DLT type) and potentially the abstraction of DLT networks.
- **PDL Platform Service Layer** it provides useful services for applications using PDL technology. As a result, an application could leverage services from the PDL Platform Service Layer rather than embed such services within the application itself.
- **PDL Application Abstraction Layer**: it conveys the set of applications that leverage PDL services as provided by the Platform Service Layer described above in order to interact with different DLT networks. For next steps, ISG PDL plans to explore new application environments, especially those enabled by the emergence of next-generation networking infrastructures, such as those related to resource trading at all levels, from compute nodes to spectrum, as well as new industrial scenarios.

Hexa-X-II work can leverage the work that PDL community is conducting on smart contracts, as well as other operational aspects such as reputation management, identity and trust management (including eIDAS qualification), all of them relevant in the scope of:

- Trusted 3rd party management, impacting Enabler 4.
- Resource orchestration mechanisms in the IoT-edge-cloud continuum, when this continuum spans across different administrative domains. This has an impact on Enablers 5 and 6.

7.1.2.6 Securing Artificial Intelligence

The ETSI ISG on Securing Artificial Intelligence (SAI) [SAI23] is a technology standardization group focusing on securing AI. The primary responsibility of ETSI ISG SAI is to develop technical specifications that mitigate against threats arising from the deployment of AI –and threats to AI systems –from both other AIs and from conventional sources. As input to the ISG SAI work, the group collected security concerns arising from AI in order to build the foundation of a longer-term response to the threats to AI and promote the development of normative technical specifications to ensure that AI systems are secure. Stakeholders impacted by the activity of the group include end users, manufacturers, operators and governments.

Specifically, the intent of the ISG SAI is to address three aspects of AI in the standards domain:

1. Securing AI from attack (e.g., where AI is a component in the system that needs defending).
2. Mitigating against AI (e.g., where AI is the 'problem' (or used to improve and enhance other more conventional attack vectors)).
3. Using AI to enhance security measures against attack from other things (e.g., AI is part of the 'solution' (or used to improve and enhance more conventional countermeasures)).

The group has published some reports such as "Problem statement", "AI Threat Ontology", "The role of hardware in security of AI", among others.

7.1.3 Internet Engineering Task Force and Internet Research Task Force

The Internet Engineering Task Force (IETF) is a standards organization for the Internet and is responsible for the technical standards that make up the Internet protocol suite. The Internet Research Task Force (IRTF) is an organization that focuses on longer-term research issues related to the Internet. Both organizations are related as the Internet Research Task Force (IRTF) focuses on longer term research issues related to the Internet while the parallel organization, the Internet Engineering Task Force (IETF), focuses on the shorter term issues of engineering and standards making.

7.1.3.1 IETF Operations and Management Area Working Group

The Operations and Management Area Working Group (OPSAWG) is a forum for developing work items within IETF, producing normative documents focused on operational and management topics. The focus of the work will be on the topics that govern the behavior of small, highly focused projects that either do not merit a Working (WG) of their own, or that belong to WGs that have already concluded.

Focusing on Hexa-X-II goals, the following activities are considered:

- State-of-the-art solutions for network management, aimed to reviewing and consolidating these solutions into stand-alone RFCs. Examples of these RFCs are [RFC6632] ("An Overview of the IETF Network Management Standards"), [RFC7276] ("An Overview of Operations, Administration and Maintenance Tools") and [RFC8969] ("A Framework for Automating Service and Network Management with YANG").
- Model-based service specification, focused on the development of models for data modelling language YANG to capture the semantics of L2/L3 Virtual Private Network (VPN) connectivity services in vendor-agnostic way, facilitating their provisioning and configuration on different carrier networks, with no dependencies on underlying vendor solutions. Decoupling services from technology allows for replicability, alleviating scalability burdens in the management plane. As of today, OPSAWG defines YANG models for the provision of Layer-2/Layer-3 VPNs [RFC9291] [RFC9182], and their operation in the context of monitoring [RFC9232] and performance assurance [RFC9375].

Hexa-X-II may incorporate, among other inputs, outcomes produced by the OPSAWG as a baseline solution for the network programmability and monitoring features which will designed and developed in enabler 1, all of them focused on advanced SDN networks.

7.1.3.2 IETF Network Modeling

The Network Modeling (NET) is a WG which addresses general topics related to the use of the YANG modelling language [RFC7950] and YANG models for management activities, including interface management language [RFC8343], IP management [RFC8344], hardware management [RFC8348], and routing management [RFC8349]. The widespread adoption of YANG in the network domain is explained by the rich set of capabilities this modeling language bring, including: i) **hierarchical configuration data models**, which enables YANG to model the hierarchical organization of data as a tree, while providing a clear and concise description of individual nodes, as well as the interaction among them; ii) **data modularity**, which enables YANG to structure data models into modules and submodules, for the purposes of easy composability; iii) **extensibility**, which mechanisms that allow YANG to augment the model-submodel hierarchy existing on data models, with one module adding data nodes to be hierarchy defined in another modules; and iv) **reusable types and groupings** (structured types), enabling YANG to tailor the modes for particular needs, with mechanisms like range or pattern restrictions.

Hexa-X-II might explore and select specific the YANG specifications produced by the NET WG to develop network configuration and monitoring solutions for those managed resources interpreting YANG language, these might include customer equipment (CPE), transport nodes and connectivity services, as well as 3GPP network functions (they both support YANG and YAML, as represented in [NRM]).

7.1.3.3 IETF Deterministic Networking

The Deterministic Networking (DetNet) Working Group is focused on creating deterministic data paths for Layer 2 and Layer 3 networks that can provide bounds on latency, loss, and packet delay variation, as well as high reliability. Some of the challenges related to DetNets are documented in [RFC 8857]. Deterministic networking requires a high degree of coordination between network devices to ensure that data is delivered with low latency and jitter. This can be difficult to achieve in large networks with many devices. Another challenge is that deterministic networking requires specialized hardware and software. This can make it more expensive to implement than traditional networking technologies. The Working Group is responsible for the overall DetNet architecture and specifications that encompass the data plane, OAM, time synchronization, management, control, and security aspects required to enable a multi-hop path with controlled latency, low packet loss, low packet delay variation, and high reliability.

Deterministic networking can be achieved using existing technologies such as TSN, MPLS and Segment Routing. These technologies provide the necessary mechanisms for traffic engineering and path computation that are required for deterministic networking. MPLS provides a way to label packets so that they can be forwarded along a specific path through the network. This allows for traffic engineering and path computation to be performed in a centralized manner. Segment Routing is another technology that can be used to achieve deterministic networking. It allows for the creation of explicit paths through the network by encoding the path information in the packet header. By using these technologies, it is possible to achieve deterministic networking without requiring significant changes to the existing network infrastructure. Now that the baseline RAW work (Section 7.1.3.5) has been finalized, it is time for the IETF DetNet to discuss which items to focus on, being many potential candidates, such as: specification of wireless specific protocol extensions for data and/or control plane, integration of multiple wireless and wired domains and administrative domains, mobility support, integration with edge deployments, OAM extensions, etc.

7.1.3.4 IETF Traffic Engineering Architecture and Signaling

The Traffic Engineering Architecture and Signaling (TEAS) Working Group is currently responsible for defining traffic engineering architecture for IP, MPLS, and GMPLS networks. In addition, the group identifies the required control-protocol functions such as routing and path computation element functions. Moreover, the

TEAS group standardizes RSVP-TE signaling protocol mechanisms that are not related to a specific switching technology.

The term Traffic Engineering (TE) refers to techniques that enable operators to control how specific traffic flows are treated within their networks [RFC8795]. TE is usually applied to packet networks via MPLS TE tunnels and LSPs. However, other mechanisms such as forwarding rules, similar to policy-based routing, may also be provided. GMPLS generalized the MPLS-TE control plane to additionally support non-packet technologies. RSVP-TE is the signaling protocol used for both MPLS-TE and GMPLS. The TEAS WG supports centralized and logically centralized control models, such as Abstraction and Control of Traffic Engineered Networks (ACTN) and stateful-PCE [RFC8453].

7.1.3.5 IETF Reliable and Available Wireless

The Reliable and Available Wireless (RAW) WG was created in 2020 to look into wireless specific aspects of deterministic networking. It was chartered as a standalone WG to allow for faster progress, instead of adding its goals to the DetNet WG, though both groups are expected to work very tightly coordinated. Wireless operates on a shared medium, and transmissions cannot be fully deterministic due to uncontrolled interferences, including self-induced multipath fading. RAW (Reliable and Available Wireless) is an effort to provide deterministic networking on a path that include a wireless interface. RAW provides for high reliability and availability for IP connectivity over a wireless medium. The wireless medium presents significant challenges to achieve deterministic properties such as low packet error rate, bounded consecutive losses, and bounded latency. RAW extends the DetNet Working Group concepts to provide for high reliability and availability for an IP network utilizing scheduled wireless segments and other media, e.g., frequency/time-sharing physical media resources with stochastic traffic: IEEE Std. 802.15.4 timeslotted channel hopping (TSCH), 3GPP 5G ultra-reliable low latency communications (URLLC), IEEE 802.11ax/be, and L-band Digital Aeronautical Communications System (LDACS), etc. Similar to DetNet, RAW technologies aim at staying abstract to the radio layers underneath, addressing the Layer 3 aspects in support of applications requiring high reliability and availability.

RAW separates the path computation time scale, at which a complex path is recomputed, from the path selection time scale, at which the forwarding decision is taken for one or a few packets. RAW operates at the path selection time scale. The RAW problem is to decide, amongst the redundant solutions that are proposed by the Patch Computation Element (PCE), which one will be used for each packet to provide a Reliable and Available service while minimizing the waste of constrained resources. To that effect, RAW defines the Path Selection Engine (PSE) that is the counterpart of the PCE to perform rapid local adjustments of the forwarding tables within the diversity that the PCE has selected for the Track. The PSE enables to exploit the richer forwarding capabilities with Packet (hybrid) ARQ, Replication, Elimination and Ordering (PAREO), and scheduled transmissions at a faster time scale.

The RAW WG was announced to be dissolved, moving its chartered items back to the DetNet WG.

7.1.3.6 IETF Distributed Mobility Management

The Distributed Mobility Management (DMM) WG has been running for many years now. Originally chartered to work on solutions that explore and standardize how to distribute the traditionally centralized mobility management approaches (Mobile IPv6, Proxy Mobile IPv6, etc) – having an impact on what 3GPP has finally standardized in terms of session and service continuity modes – it is now responsible of the maintenance of IP mobility protocols in general in the IETF.

In addition to the maintenance work, DMM has also been analysing more “disruptive” approaches to mobility and connectivity architectures, for example evaluating the impact of slicing in the transport networks, or which protocol alternatives could be used instead of the traditional GTP protocol.

Give this nature of “exploratory” WG, DMM might be a good option to disseminate and discuss new mobility approaches, that might for example integrate deterministic networking requirements (such as [BM23]). It is also worth mentioning that DMM is a WG that could be chartered down in the future, but it is not foreseen that this can happen during the Hexa-X-II timeframe, and therefore it is considered a good target WG for having impact.

7.1.3.7 IETF MAC Address Device Identification for Network and Application Services

The MAC Address Device Identification for Network and Application Services (MADINAS) WG is a recently chartered group at the IETF aimed at exploring the impact of Randomized and Changing MAC addresses (RCM) on networking, transport and applications. RCM has been massively adopted by mobile and fixed operating systems and this has had an impact on networking protocols, as many use cases and applications make an implicit assumption that a device is represented by a unique and permanent layer 2 address.

The group is currently chartered to document the current RCM state of affairs by: (*i*) identifying relevant network and application services scenarios and examining the effect of RCM schemes on them; (*ii*) analyzing various existing identifiers (i.e., beyond the MAC address) that can be used by the network to provide seamless services, and, (*iii*) identifying scenarios where device identity is not required. The group will also generate a Best Current Practices (BCP) document recommending means to reduce the impact of RCM on the documented use cases while ensuring that the privacy achieved with RCM is not compromised. For scenarios where device identity stability is desirable, the BCP document will recommend existing protocols that can be used to protect the request and exchange of identifiers between the client and the service provider.

Given the impact of privacy in current and future network architectures, MADINAS is relevant for the Hexa-X-II work, especially when considering in conjunction other networking and orchestration solutions that might be in place.

7.1.3.8 IRTF Compute In the Network Research Group

The Compute In the Network Research Group (COINRG) [COI23] is an Internet Research Task Force (IRTF) research group that aims to explore and promote the use of in-network computing for improving network and application performance and user experience. The group focuses on investigating the benefits of the emerging disruption to the Internet architecture brought about by Compute In the Network, which involves using network devices for computation, storage, and management purposes. COINRG's mission involves investigating the use of programmable network devices, languages, and abstractions to implement network functions and improve Internet performance. The group also aims to identify potential benefits of in-network functionality such as compute, cache, manage, and control, among others. The group aims to investigate novel architectures, data-plane abstractions, and new network protocol designs to efficiently federate decentralized computing resources across the infrastructure, regardless of where in the network the compute is placed.

7.1.3.9 IRTF Network Management Research Group

The Network Management Research Group (NMRG) [NMR23] serves as a platform for researchers to explore novel technologies for managing the Internet. Its primary goal is to devise solutions for issues that have not yet been fully comprehended enough for practical implementation within the IETF. The NMRG places emphasis on management services that work in conjunction with the current Internet management framework. This involves communication services between management systems, which could belong to separate management domains, as well as management services aimed at customers.

The ultimate goal of self-driving/-managing networks is fully autonomous network operations. However, there are intermediate levels where human users remain "in the loop" and are progressively assisted and replaced by more intelligent mechanisms. Interfaces between humans and a self-driving system are essential for bidirectional communication. On one hand, users should be able to express guidance and needs without having to handle the full complexity of the underlying infrastructures. On the other hand, users should understand the decisions made, the reasons why, be informed about the future actions the system will initiate, and be provided with recommendations. Intent-Based Networking (IBN) provides high-level, user-friendly abstractions to describe business and operational goals, alleviating the need for the user to know and derive technical details on how to achieve those goals. IBN is an essential component of self-driving networks but requires the introduction of intelligent mechanisms to process intents with minimal human involvement. Advances in AI can provide intelligent mechanisms to process intents. Different forms of AI have been used for decades in network management. The combined progress in data, computing power, AI algorithms, and flexible

capabilities of networks in recent years make it highly relevant to re-examine the coupling between AI and network management in depth.

7.1.4 O-RAN Service Management and Orchestration

The O-RAN Alliance [ORAb23] global community, founded in February of 2018, is formed by MNOs, vendors, and research institutions working on the Radio Access Network (RAN) industry in order to enhance RAN functionalities by making it more intelligent, open, virtualized, disaggregated, and extensible. As already depicted in [Hexa-D61], the main goal of O-RAN Alliance is to define specifications and references for implementation and development about Open RAN.

O-RAN includes the so-called O-RAN Service Management and Orchestration (SMO) framework, which belongs to the highest layer of the O-RAN architecture, shown in Figure 7-1, covering the functionality of orchestrating, managing, and automating the O-RAN elements.

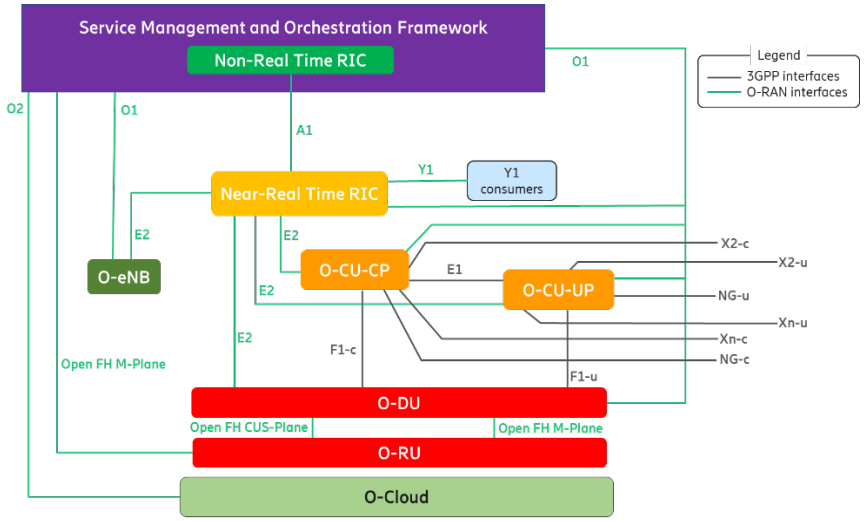


Figure 7-1: Logical Architecture of O-RAN [ORA23]

Since [HEX-D6.1], certain improvements have been added to the SMO, the most relevant being the introduction of an innovative framework known as the AIML Framework (AIMLFW, in the [ORA22] release). This framework operates as a standalone installation, existing outside the current deployment infrastructure. AIMLFW is based on Kubeflow [KUB23], intended for the training phase of the AI/ML models, and Kserve [KSE22], for the inference stages. Among other improvements in the last release [ORAc23], it is worth mentioning the AIMLFW module, also related to the SMO, allowing to deploy and manage AI/ML models in Near-RT RAN Intelligent Controller (RIC) and Non-RT RIC (Non-RT RIC is part of the SMO dealing with management of RAN functions in a non-real-time manner), giving to these components AI/ML capabilities while ensuring their Life Cycle Management (LCM).

7.1.5 Open Networking Foundation

The Open Networking Foundation (ONF) is an operator-driven, community-led non-profit consortium fostering innovation in software-defined programmable networks.

7.1.5.1 Transport API

Transport API (TAPI), a RESTCONF YANG interface, is designed to facilitate communication between SDN controllers and orchestrators. It offers a neutral model that covers photonic, optical transport network (OTN), and Ethernet networks, with forwarding-technology-layer specific augments. TAPI supports a range of functions, including Topology, Connectivity, OAM, Path Computation, Equipment, Notification, Streaming, and Fault Management, and enables a TAPI client to stay in sync with the network properties controlled by the TAPI provider via ongoing state updates.

TAPI v2.4.0 includes two Reference Implementation Agreements, which specify the application of the TAPI models for various use cases [TAP22]. The latest version builds on previous releases by enhancing existing

capabilities and introducing new features. These include improvements to photonic impairment models, refinements to network layer modelling, extensive OAM enhancements, consolidation of alarm and performance monitoring (PM) structures, support for physical route describing equipment, and enhancements to Network Edge Points and Service Interface Points (NEP/SIP) relationships.

TAPI's combination of features and enhancements makes it a strong choice for integrating control over photonic, OTN, and Ethernet networks. TAPI v2.4.0 is a significant step forward into the path of openness and disaggregation of Optical Transport Networks. data retrieval for Optical Tributary Signal (OTSi) path planning and validation.

7.1.6 Open-Source Software

In this section, multiple open-source software projects are referenced. They are related to management and orchestration mechanisms to be considered for future 6G networks.

7.1.6.1 ETSI OpenSource MANO

Open-Source MANO (OSM) is the ETSI open-source stack implementing the ETSI NFV specifications. It is an example of a community-led effort that produces a MANO stack of production quality that satisfies the needs of operators for use in commercial NFV deployments [OSM23]. The objective of OSM is to enable automated management and orchestration of Network Services (NSs) that comprise multiple Virtual Network Functions (VNFs), across several administrative domains. The project aims to deliver a modular, scalable, and cloud-native MANO solution, supporting the whole lifecycle of VNFs, from onboarding and service design to monitoring and optimization. OSM enables the definition of Network Slice templates that can be customized for different tenants each with its own policies and security requirements. This feature allows operators to offer NFV services to different customers, while ensuring isolation and security between them. Furthermore, it is capable of interacting with several Virtual Infrastructure Managers (VIMs) and to manage different types of NFs i.e., VNFs, Containerized Network Functions (CNF), Kubernetes Network Functions (KNF), Physical Network Functions (PNF), and Hybrid Network Functions (HNF).

Since 2022, ETSI OSM has launched Release TWELVE [OSM-12], which is the second Long Term Support (LTS) release of ETSI OSM and one of its most prolific releases and also, launched Release THIRTEEN [OSM-13] with a new scalable architecture for massive closed loop operations. The ability to recover NFs from infrastructure failures is a notable addition in the OSM TWELVE LTS release. This functionality is applicable across all cloud types supported by OSM, aligning with the platform's multi-cloud strategy. Aside from enhancing Day-2 NS operations (i.e., focused around maintaining, monitoring, and optimizing the system), this release also enables updating a running NF Instance (NFI) to a newer package version with minimal downtime. On the other hand, OSM Release THIRTEEN integrates a new scalable architecture for Service Assurance and Closed-Control loop (CCL) operations. This architecture utilizes a cloud-native version of Apache Airflow [airflow] and Prometheus [PRO23] to cope with the demands of even the most challenging service assurance scenarios, including auto-healing and auto-scaling across large clouds and multiple edge sites. In addition, new workflows have been added to obtain the state of NFs, NSs and VIMs. Further capabilities are expected to be added in upcoming releases [OSM23].

7.1.6.2 ETSI TeraFlowSDN

TeraFlowSDN is an SDN controller developed by ETSI Open-Source Group TeraFlowSDN (OSG TFS) that utilizes a micro-services architecture, providing improved resource allocation and faster development cycles [VMC+21]. Notably, TeraFlowSDN supports IP over optical transport networks and offers OpenConfig support for the IP layer, while interfacing with the optical layer through the ONF Transport API. This feature makes it a unique solution for the transport network orchestration, with no other open-source alternative currently available.

TeraFlowSDN supports a diverse range of use cases established by various ETSI standardization groups and strives for interoperability with ETSI Open-Source MANO (OSM). It seamlessly integrates with existing frameworks such as NFV and MEC and provides a platform for standardization groups and research initiatives to experiment with features such as flow aggregation, management (service layer), network equipment integration (infrastructure layer), AI/ML-based security, and forensic evidence for multi-tenancy.

The development group of TeraFlowSDN closely collaborates with various standardization bodies and initiatives, such as IETF, ETSI ZSM, ETSI NFV, ETSI MEC, ETSI Millimetre Wave Transmission (mWT), and ETSI SAI. It is worth noting that ETSI TeraFlowSDN emerged from the European Community financed H2020 TeraFlow project, and the group is committed to ensuring compliance with European cybersecurity requirements.

In February 2023, TeraFlowSDN Release 2 was launched, offering expanded and validated support for transport network slicing across multiple domains. This release features complete SDN orchestration for L2/L3VPN provisioning, microwave networks, Point-to-Multipoint integration of XR optical transceivers, and interaction with optical SDN controllers via the Open Networking Foundation (ONF) Transport API (TAPI).

The ETSI TeraFlowSDN community has made a commitment to implement the Telecom Infra Project (TIP) Mandatory Use Case Requirements for SDN for Transport (MUST) in their innovative cloud-native SDN Controller. This move positions TeraFlowSDN as a reference implementation in the TIP Open Optical & Packet Transport group (TIP OOPT), accelerating the adoption of SDN standards for IP/MPLS, Optical, and Microwave transport technologies. The objective of MUST is to promote the development of standards-based, interoperable solutions that are more accessible to broader communities.

By using TeraFlowSDN as a reference implementation, the networking community can benefit from an open, standards-based solution that simplifies the development, testing, and deployment of new functionality.

7.1.6.3 5G Monitoring Platform

The 5G Monitoring Platform is a tool designed and originally developed in the context of ESA Artes ANChOR project [ESA23] as part of an orchestration and control platform in charge of implementing an AI-based data-driven closed loop for the automatic management of network slices. The platform consists of a set of open-source software for monitoring and data collection, properly managed by custom software module called Config Manager, as shown in Figure 7-2.

A set of Telegraf instances collects the data from different and heterogenous data-sources and performs the data adaptation by embedding the values into specific messages. Such messages are then published both to the Kafka bus, for a real-time data exposure, and to InfluxDB in form of time series. Both forms can be consumed by an AI for i) training models and ii) reacting to specific events. Prometheus is exploited for data aggregation and manipulation and to set alarms through the Alarm Manager. Such alarms are usually triggered upon exceeding a certain threshold and are published in Kafka to guarantee a prompt reaction.

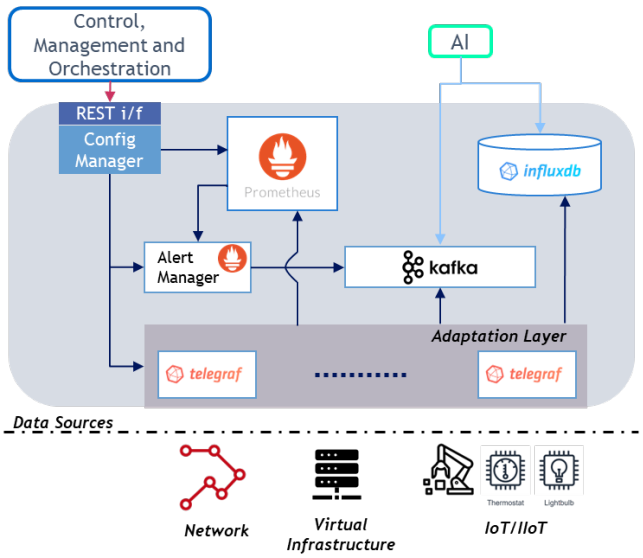


Figure 7-2: 5G Monitoring Platform software architecture

The different elements of the platform are dynamically configured by the Config Manager, which executes requests coming from Management, Orchestration and Control platforms, both at orchestration time and at

runtime, as the elements to be monitored may vary during the service/slice lifetime. Such configurations include the type of data sources, the number of Telegraf instances, alarms, and topics for Kafka bus.

7.1.6.4 OpenTelemetry

OpenTelemetry [OPT23] is a vendor-neutral observability framework aimed at standardizing the collection, export, and processing of telemetry data from distributed systems. It provides a set of APIs, libraries, and agents for capturing telemetry data across various programming languages and platforms. OpenTelemetry enables the collection of metrics, traces, and logs, facilitating comprehensive observability and enabling operators to gain insights into the behavior and performance of complex systems.

OpenTelemetry makes it easy to collect, analyze, and correlate telemetry data from different parts of an application or system. Here are some of the benefits of using OpenTelemetry: a) Vendor neutrality; as it can be used with a variety of different collectors and processors. This makes it easy to choose the right collector or processor for specific needs; b) Scalability; as the OpenTelemetry Collector can be deployed on-premises or in the cloud, and can scale to meet the needs of even the most demanding applications; and c) Flexibility; it can be used to collect a variety of telemetry data, and it can be exported to a variety of destinations. This makes it easy to customize OpenTelemetry to meet the specific needs of your application or system.

7.1.6.5 Prometheus

Prometheus [PRO23] is an open-source monitoring system that specializes in time-series data collection and alerting. It offers a flexible query language, powerful visualizations, and seamless integration with other monitoring tools, making it a popular choice for monitoring and metrics collection in modern infrastructure environments.

7.2 Industry fora

7.2.1 TM FORUM

TM Forum [TMF23] is an international organization that brings together global companies (digital service providers, technology suppliers, consultancies and system integrators) that concentrates on transforming digital industry. By using collaborative programs and industrial communities, they facilitate the creation of rapid prototypes, spanning from digital business models to open APIs and related procedures. TM Forum offers industry best practices and standards to hasten adoption. Under the umbrella of TM forum, there are multiple projects that are interested in various industry relevant topics covering digital transformation, zero-touch management, and Digital Twin (DT) and AI based decision making for the usage of CSPs.

7.2.1.1 Autonomous Networks

The Autonomous Networks (AN) Project [IG1305] aims to define fully automated zero wait, zero touch, zero trouble innovative network/ICT services for vertical industries' users and consumers, supporting self-configuration, self-healing, self-optimizing and self-evolving telecom network infrastructures for telecom internal users: planning, service/marketing, operations and management. Autonomous Networks incorporate a simplified network architecture, autonomous domains, and automated intelligent business/network operations for the closed loop control of digital business, offering the best-possible user experience, full lifecycle operations automation/autonomy and maximum resource utilization. Hence, the four fundamental concepts introduced in IG1230 [TMF-IG1230] to support the AN Objectives are:

- Autonomous Domains.
- Objects (or Managed Entities) and their taxonomy.
- Business Services that are exposed by Autonomous Domains are based on the notion of Intent-driven Service definitions.
- Support of (closed) Control Loops within Autonomous Domains and across Autonomous Domains.

The four closed loops identified to fulfil the full lifecycle of the inter-layer interaction are:

- User closed loop – the interaction across three layers and three closed loops to support fulfilment of the user's service.
- Business closed loop – the interaction between business and service operations may trigger related service and resource closed loops in its fulfilment.
- Service closed loop – the interaction between service and network resource operations may trigger related resource closed loops in its fulfilment.
- Resource closed loop – the interaction of network resource operations is in the granularity of autonomous domains.

The user closed loop is the main thread to streamline the business/service/resource closed loops, while each of the business/service/ resource closed loops addresses the interaction between adjacent layers. The interaction between adjacent layers is simple, business driven and technology/implementation independent, i.e., communicating and fulfilling the intents (business/ service/resource) rather than technology-prone commands based on the intent mechanisms and interfaces. The different intents are used for the interactions of different layers, i.e., business intent, service intent and resource intent.

7.2.1.2 DT4DI – Digital Twin for decision intelligence

This collaboration project [IG1307] aims to set up a foundational ontology, including domain concepts, definitions, and baselines, and to design and engineer knowledge-driven decision intelligence solutions powered by Digital Twins (DT) and AI, supporting vital CSPs' business processes and tactical and strategic decision-making.

AI and Digital Twins boost each other to support Decision Intelligence to elaborate decisions and provide the most optimal scenarios to humans.

- AI enables the exploration of Digital Twin models with powerful discovery, analysis, prediction, correlation, learning, and reasoning capabilities. AI programs are also fundamental to DT systems to represent the operations of the real world for complex systems that cannot be modelled using rule-based equations.
- Digital Twins provide structured, comprehensive, and dynamically updated data models simulating the real world (potentially every real entity) that facilitate the development, deployment, and operations of AI solutions. Highly reliable DT virtual models, massive twin data, and real-time two-way dynamic interaction enable diverse and accurate AI models. DT can produce simulated data in virtual environments and go through infinite repetitions and scenarios. The simulated data and virtual environments can be used to effectively and efficiently perform AI model training.

7.2.2 GSMA

The GSM Association is a non-profit industry organization that represents the interests of mobile network operators worldwide.

7.2.2.1 Operator Platform Group

The Operator Platform Group (OPG) is an Industry Specification Issuing Group that defines the architecture, the requirements, and the APIs for an Operator Platform (OP). An OP facilitates access to the Edge Cloud and other capabilities of an operator or federation of operators and their partners. The platform exposes these capabilities through APIs to non-telco developers who could use them for example to support improving or even enabling the delivery of their application's service to its users.

On the other hand, the Operator Platform API Group (OPAG) is a subgroup of the OPG focussing on identifying the APIs that could be used for the realisation of the different interfaces of the Operator Platform, a federated platform concept that facilitates access to the Edge Cloud and other capabilities of an operator or federation of operators and their partners. Where available OPAG may refer to existing APIs from SDOs or they can define APIs themselves where such existing APIs are not considered suitable.

OPAG can also start Sandboxes where members agreeing to participate in those according to OPAG's terms of reference can work on contributions for Linux Foundation's CAMARA project (in Section 7.2.4.4).

The high-level architecture of the OP for edge computing consists of the following elements:

- Operator Platform. Single point of entrance for application providers for delivering edge computing capabilities at the location of the users. OP interconnects edge computing capabilities across different footprints that aggregate IT capabilities at the edge infrastructure equipment and the edge data center facilities that are deployed..
- Application provider. Uses northbound APIs to request edge capabilities through OP, including onboarding, instantiation, resources and application provider.
- Edge Computing Platform. Connects to the OP by SBI for resources management.
- Cloud Management Platform. Access point which combines a set of features of different cloud environments and resources, ensuring proper capacity handling. This element is associated to cloud edge resources and interacts with the edge computing platform.
- End user. Final service user, who requests app usage by User Network Interface APIs and make use of resources allocated on a proper edge instance, which selection is based on user location or network requirements/status.

7.2.2.2 Open Gateway

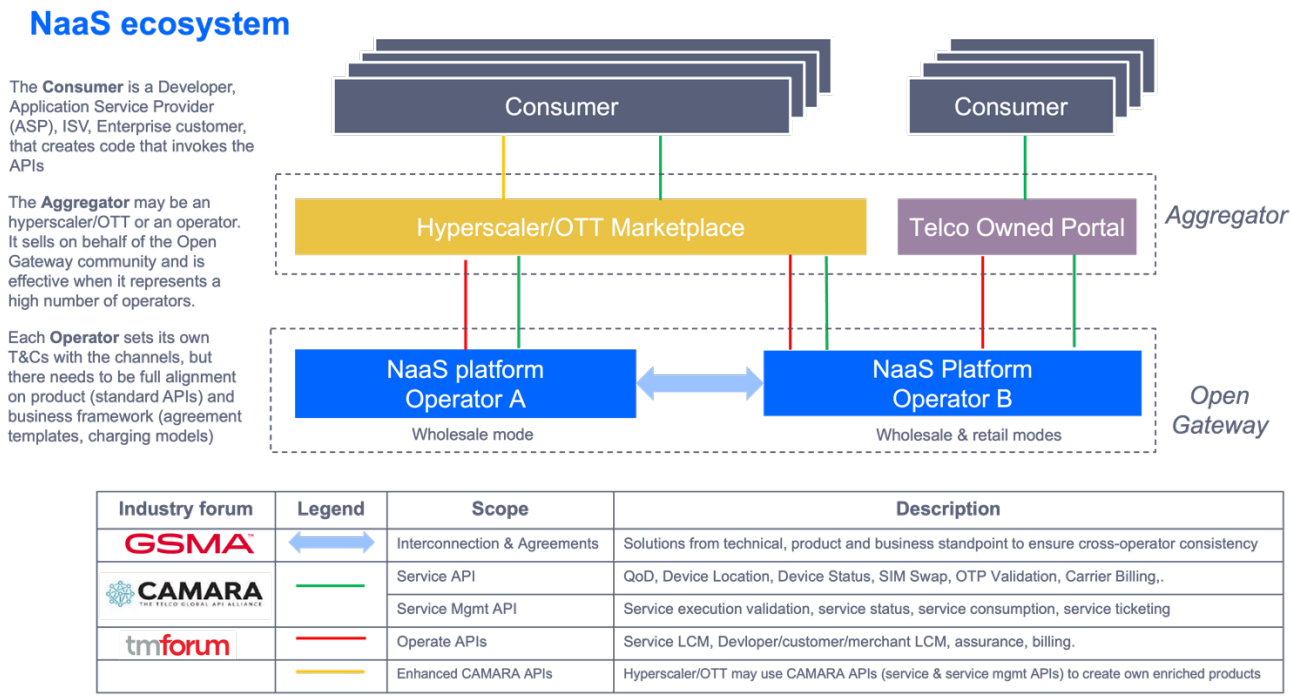
Open Gateway is a GSMA initiative designed to allow operators to expose and monetize telco capabilities to third party service providers in a programmatic manner through Application Programming Interfaces (APIs). This approach, referred to as Network as a Service (NaaS), helps developers enhance and deploy services quicker across operator networks via single points of access to the world's largest connectivity platform.

Open Gateway was launched in the MWC Barcelona 23, with a twofold mission: i) to provide a governance framework for NaaS, covering technical and business aspects; ii) to get operator commitment to launch universal NaaS API services in 2023.

Open Gateway initiative recognizes that the NaaS concept builds on the work developed by three organizations:

- Linux Foundation's CAMARA: it represents the "exposure" doctrine, i.e., how capabilities are exposed for external consumption through 3rd party facing APIs. CAMARA defines these APIs and is responsible for their hosting and release management. 3rd party facing APIs are user-friendly (semantics tailored to service and business needs of 3rd parties) and open (following Apache2.0 license). Further details on CAMARA are captured in section 2.4.4.
- GSMA: it represents i) the "technical" doctrine, by specifying how 3rd party facing APIs are to be supported by underlying telco capabilities, including network and cloud capabilities; and ii) the "business" doctrine, with the definition of agreement templates for federation between the operator networks and for relationship with 3rd parties, ensuring a consistent yet fair commercial framework for exposing services. GSMA conducts the technical workstream through OPG/OPAG (Operator Platform Group / Operator Platform API Group) and the business workstream through WAS (Wholesale Agreement Services) group.
- TM Forum: it represents the "operational" doctrine, i.e., how 3rd party facing APIs are to be operated and managed, to make a commercial product out of them. Representing the IT side of operators, that includes Operation, Administration and Maintenance (OAM) functionality provided by OSS, BSS, and on-line charging systems, now structured following a new digital architecture referred to as Open Digital Architecture (ODA).

Figure 7-3 illustrates the contributions of these organizations into the Open Gateway actor-role model.



Industry forum	Legend	Scope	Description
GSMA		Interconnection & Agreements	Solutions from technical, product and business standpoint to ensure cross-operator consistency
CAMARA		Service API	QoD, Device Location, Device Status, SIM Swap, OTP Validation, Carrier Billing,.
		Service Mgmt API	Service execution validation, service status, service consumption, service ticketing
tmforum		Operate APIs	Service LCM, Devloper/customer/merchant LCM, assurance, billing.
		Enhanced CAMARA APIs	Hyperscaler/OTT may use CAMARA APIs (service & service mgmt APIs) to create own enriched products

Figure 7-3: Open Gateway ecosystem

On the one hand, the scope of **GSMA** is circumscribed to the telco domain. GSMA prescribes the capabilities that all operators must make available for 3rd parties, to ensure global reach and scale. These must-be capabilities are referred to as Open Gateway services. The GSMA is also responsible for i) the prioritization and roadmap management of Open Gateway services, according to market needs and commercial readiness of underlying technologies; and ii) architecting the platform that individual operators will use to realise and expose Open Gateway services.

On the other hand, the scope of **CAMARA** and **TM Forum** is on the APIs that allows programmatic access to Open Gateway services. These APIs can be clustered into three groups:

- Service APIs: they allow invoking Open Gateway services. Examples of these APIs are Quality on Demand API, Device Location API, Device Status API, SIM Swap API, OTP validation API or Carrier Billing API. As noted, these APIs provide purpose-specific application-tailored functionality.
- Service Management APIs: they allow running certain management actions from within Open Gateway services, like ordering the activation/de-activation of certain functionality for that service, monitoring, eligibility check or consumption check.
- Operate APIs: they provide all the transversal (non-service specific) functionality that is required to make a commercial product out of the Open Gateway services, making them operable and monetizable. Examples of OAM functionality provided by the Operate APIs include registration/onboarding of 3rd parties, service fulfilment (e.g., provisioning, activation, and modification), service assurance (e.g., incident management, service supervision and performance) and billing.

With regards to API owners, Service and Service management APIs are defined, developed, tested, and maintained by CAMARA, while TM Forum is responsible for Operate APIs.

With regards to targeted consumers, it is worth noting that CAMARA APIs are used by 3rd parties in their applications. These APIs are delivered though aggregators (wholesale model), or by the telco itself using its portal (retail model). The aggregators may develop and expose additional "Enriched APIs", by adapting or combining CAMARA APIs. The Operate APIs (red line) are however not available for 3rd parties; they are used for integration with aggregators and portals instead.

Further details on Open Gateway can be found in the White Paper published in [OGW23].

7.2.3 OpenConfig

OpenConfig [OP23] is an open-source project that provides a common, vendor-independent software layer for managing network devices. The project operates with contributions from network operators, equipment vendors, and the wider community. OpenConfig is led by an Operator Working Group consisting of network operators from various segments of the industry.

One of the core features of OpenConfig is its consistent and coherent data models designed by users for vendor-neutral management in a wide variety of networking use cases. OpenConfig's initial focus is on compiling a consistent set of vendor-neutral data models written in YANG. These models are based on actual operational needs from use cases and requirements from multiple network operators. The models can be developed directly by OpenConfig or compiled from third-party modules that conform to the OpenConfig requirements. The project is also interacting with standards bodies and network equipment manufacturers with the aim of making these models the basis of widely-adopted standardized interfaces.

Another key feature of OpenConfig is streaming telemetry, a subscription-based model for efficiently and accurately monitoring network devices based on OpenConfig models. Streaming telemetry is intended to replace SNMP, a protocol widely used for monitoring network devices, but which has limitations in terms of scalability, accuracy, and efficiency. OpenConfig's streaming telemetry offers a more efficient and accurate way of monitoring network devices, enabling operators to quickly identify and resolve issues.

OpenConfig also provides device management and control protocols based on gRPC [GRP23], a modern, secure RPC framework built for distributed services. gNMI [GNM23] is a network management interface that defines a gRPC-based protocol for the modification and retrieval of configuration from a target device, as well as the control and generation of telemetry streams from a target device to a data collection system. The intention is that a single gRPC service definition can cover both configuration and telemetry, allowing a single implementation on the target device, as well as a single NMS element to interact with the device via telemetry and configuration RPCs. This simplifies the management and control of network devices, making it easier for network operators to manage their networks.

Finally, OpenConfig provides vendor-independent automation to simplify and accelerate compliance testing of OpenConfig implementations. This allows network operators to quickly and easily test the compatibility of their network devices with OpenConfig, ensuring that they meet the required standards and are ready to be deployed in production networks.

7.2.4 Linux Foundation

The Linux Foundation (LF) is a non-profit technology consortium that hosts and promotes the collaborative development of open-source software projects. Several projects are related to Networking and in the next subsection, the most significant ones are presented.

7.2.4.1 Open Network Automation Platform

ONAP (Open Network Automation Platform) is an open-source platform that is part of Linux Foundation [LFN23] projects. ONAP aims to provide a framework for real-time, policy-driven orchestration and automation of network services based on virtualized functions [ONA23]. Since the first version released in November of 2017 until the last one in July of 2023 several enhancements have been added to the platform such as physical functions, containerized functions, network slicing, improvements according to networking standards, etc.

The main blocks are:

- Design-Time that is a workspace that offers a set of tools, techniques, and repositories for defining resources, services, and products.
- Run-Time that is a framework that runs the rules and policies and other models previously created by the design and creation environment.
- Manage ONAP, which offers management capabilities for the ONAP itself.

Additionally, some subcomponents conform the different aforenamed main blocks.

Regarding the Design-Time framework, the most important subcomponent regarding M&O is the so-called Service Design and Creation (SDC) framework [ONAb23]. It provides tools, techniques and repositories that allow to, not only define, but also simulate and certify system assets and the related processes and policies. Since release [KOH22], launched in December of 2022, SDC supports the onboarding of Application Service Description (ASD) packages, which are deployment descriptors for cloud native applications and functions.

Concerning the Run-Time framework the most important modules for M&O are the so-called Service Orchestrator (SO), Controllers, Data Collection, Analytics and Events (DCAE), and Active and Available Inventory (A&AI). DCAE adds to the ONAP architecture the possibility to collect events and host analytic applications. As already depicted in [Hexa-D61], DCAE is also targeted to collect and process data coming from the management systems and network. Additionally, it proposes a framework to develop analytic applications, and a Hadoop cluster for processing and storing data. It also includes a microservices catalogue to address these functionalities. The main feature here in order to address the smart network management for M&O is the addition of AI/ML support through the Acumos adapter, as depicted in [DCA20] already mentioned in [Hexa-D61]. However, it has been deprecated in the release [KOH22]. Nevertheless, in the last release [LON23] a new feature called "AI/ML mS for Intent-based network (IBN) based closed loop in E2E Network Slicing" is presented and shown through a proof of concept (PoC). It aims to predict the real time configurations of cells and slices. More information about the PoC can be found at the repository where the source code is located [DCA23]. It is also important to mention that the SO allows to perform basic orchestration operations such as creation, modification, or removal of network services and resources as VNFs, PNFs and CNFs, enabling the management from an E2E perspective.

The only component that can be found at Manage ONAP is the subcomponent called ONAP Operations Manager (OOM). It encompasses the lifecycle management of the whole ONAP platform in order to cover orchestration functionalities like deployment, configuration, monitoring, restart, clustering and scaling, upgrade and deletion of the ONAP components.

Due to the different ONAP orchestration capabilities, some of which based or including AI/ML, ONAP could address some of the goals of the enablers mentioned in Section 4, adding M&O functionalities.

7.2.4.2 EMCO

The Edge Multi-Cluster Orchestrator [EMC23] is an open-source project committed to create a universal control plane to securely connect and deploy Kubernetes workloads (e.g., a simple application or a complicated application that is composed of multiple simple applications, or it can be a network function in the form of container/VM) across public and private clouds and edge locations (e.g., IoT edge site cluster, a 5G cell cluster, etc.), enabling, at the same time, inter-application communication. The EMCO is a geo-distributed intent-based framework that works as a universal control-plane and application orchestrator for Kubernetes, capable of deploying cloud native applications to multiple Kubernetes clusters, spanning enterprise data centers, different cloud providers, and multiple edge locations. EMCO handles also the automation of the infrastructure needed by the deployed workloads: from the networking setup to service mesh to application security. The architecture of EMCO makes it modular and highly scalable, giving it the flexibility to be applicable to any environment with multiple clusters requiring end-to-end inter-application communication. Operating at a higher level than Kubernetes, EMCO performs decision making processes to select on which clusters a workload should run, then interacts with the Kubernetes API Server to perform the workload deployment. EMCO supports multiple placement constraints, some of which are inherited from Kubernetes itself: Affinity and Anti-Affinity patterns, Platform capabilities, and Latency/Cost.

The communication between deployed applications and with an external service within or across clusters (with or without mutual TLS) are enabled by the configuration of service mesh and security policies (e.g., NAT, Firewall). EMCO also provides application lifecycle management capable of application upgrades and comprehensive status monitoring.

7.2.4.3 Tungsten Fabric

Tungsten Fabric [TUN23] is an open-source project committed to fostering rapid innovation in software-defined networking aiming to provide a single networking and security tool to achieve: a) connecting multiple

orchestration stacks, b) choosing an SDN plug-in, c) networking and security across legacy, virtualized and containerized applications, and d) multitask and across-stack policy control, visibility and analytics.

Tungsten Fabric provides a scalable virtual networking platform working with a variety of virtual machines and container orchestrators and can integrate with physical networking and compute infrastructure using networking standards like BGP EVPN control plane and VXLAN overlays to connect workloads in different orchestrator realms. Tungsten Fabric implements the virtual networking in cloud environments leveraging OpenStack and Kubernetes orchestrators and using overlay networks between vRouters that run on each host. Tungsten Fabric provides a variety of enhanced features over the native networking implementation of orchestrators such as highly scalable multi-tenant networking, multi-tenant IP address management, DHCP, ARP proxies, etc.

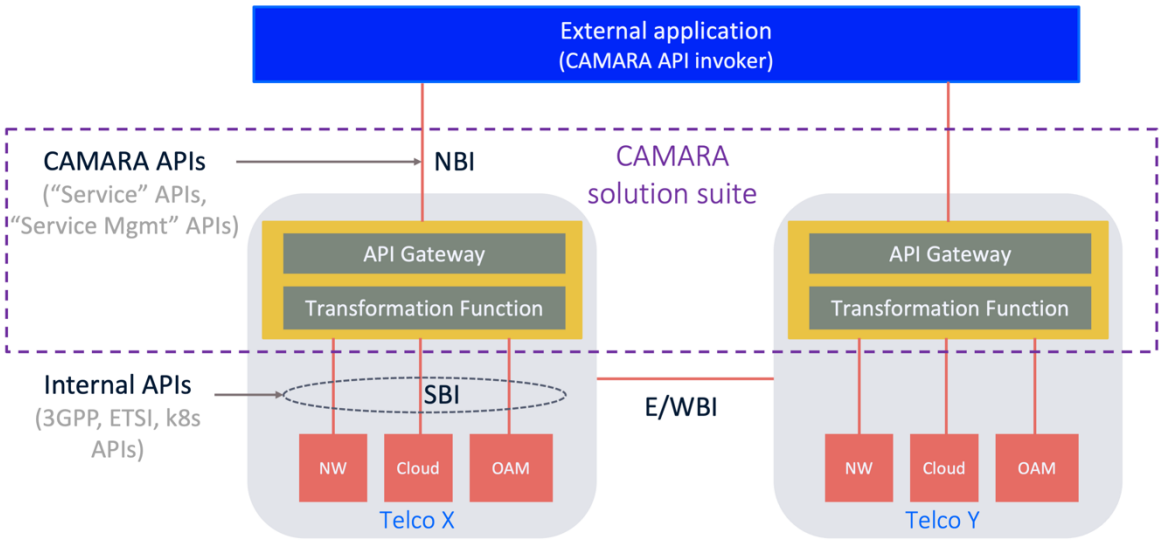
At the base of the Tungsten Fabric work there are the Tungsten Fabric Controller and the Tungsten Fabric vRouter modules. The Tungsten Fabric Controller ensures that when a Virtual Machine or a container is created, it is provided with network connectivity according to network and security policies in the controller or orchestrator; it also includes a set of software components that maintains a model of networks and network policies. The Tungsten Fabric vRouter module is installed in each host that runs workloads (i.e., Kubernetes, OpenStack, etc.) performing packet forwarding and enforcing network and security policies. The Tungsten Fabric controller includes software plugins to interact with the compatible orchestrators; those plugins implement the networking service of the target orchestrators: the Tungsten Fabric plugin for OpenStack implements the Neutron API, while the Kubernetes plugin implements the K8s API Server interface to listen to network-related event through the watch capabilities powered by Kubernetes API themselves. The Tungsten Fabric vRouter, running on each compute node host, replaces the Linux bridge and IP tables or the Open vSwitch networking; the controller configures the vRouter to implement the networking and the security policies that should be enforced. Communication between the controller and vRouters is performed through XMPP messaging protocol and the controller itself is responsible for installing the set of routes in each VRF of each vRouter that implements network policies.

7.2.4.4 CAMARA

CAMARA project [CAM23b] responsible for the definition, implementation, testing and maintenance of APIs that 3rd parties will use to gain access to GSMA Open Gateway services.

Figure 7-4 illustrates the reference architectural framework of CAMARA. As seen, it builds on the following components:

- **Internal APIs**. These are the APIs which are implemented in telco assets, including network resources (core, access, transport functions), cloud resources (virtualized and cloud-native workload hosting infrastructures) and OAM resources (OSS and inventories). These APIs are typically defined in standard bodies or industry fora, and tied to the underlying technology. Examples of these APIs include the ones defined by 3GPP, ETSI, IETF or Kubernetes, among others.
- **CAMARA APIs**. These are the “Service” and “Service Management”, i.e., the APIs made available for consumption to external/3rd party application. CAMARA APIs design guidelines are reported in [CAM23]. They include the following two principles: openness (API source code is Apache 2.0) and intent-based (service API definition follows a dev-friendly style, hiding the telco complexity inherent to network APIs).
- **Transformation function**. It keeps the information on correspondences between CAMARA APIs and internal APIs and executes workflows to enforce these mappings. This component can be deployed as a microservice provisioned with a workflow engine.
- **API Gateway**. It provides all the capabilities that are needed to policy the interaction between the operator and the external applications, in relation to CAMARA API invocation. These capabilities include service API publication & discovery, access control (authentication & authorization of applications), auditing, accounting, and logging.



NOTE: E/WBI stands for East/Westbound Interface, and calls for federation capabilities across telco operators. This is transparent to 3rd parties, and therefore out of scope of CAMARA project

Figure 7-4: CAMARA Reference Framework

As of today, CAMARA is developing APIs for a number of Open Gateway services, including Quality on Demand, edge cloud, device location, device status, carrier billing or SIM Swap. The complete list of APIs and their status is captured in [CAM23]. A paper published in 2022 showed some results done for Quality on Demand API [OD22].

7.2.4.5 SYLVA

Project Sylva [SYL22] is an initiative launched by Linux Foundation Europe to create an open-source and cloud-native framework for telcos that meets the federated and regulatory challenges of the EU. It aims to reduce stack complexity across edge, cloud and application layers and helps reduce transformation costs in a rapidly evolving sector. The Open-Source releases resulting from this project will incorporate the capabilities that are required for a CaaS (Container as a Service) to address specific use cases identified by European Telcos (such as 5G, O-RAN and Edge) and will be the basis for a common infrastructure among European operators allowing the federation and integration of edge applications.

The Sylva project has two main objectives. Firstly, to release a cloud software framework that will identify and prioritize telco and edge requirements, develop solutions to specific technical challenges in the infrastructure layer of the telco ecosystem, and integrate these solutions with existing open-source components. This cloud software framework, while not "production ready," aims to be "production grade" so that it can be utilized by third parties, such as operators, network function vendors, and cloud providers, to create commercial products. Secondly, the project aims to develop a Reference Implementation of this cloud software framework and establish an Integration and Validation program to validate commercial network functions against the framework, validate implementations based on the released framework and its components, and accelerate the commercial adoption of network functions and their compliance with this cloud software framework.

This initiative encompasses five technical pillars. Firstly, network performance, which involves implementing the performance requirements for 5G Core and Open Ran CNF. Secondly, distributed cloud, which aims to address use cases such as distributed UPF in 5G Core, CDN or O-RAN. Sylva will provide an architecture that can manage cloud infrastructures from central locations to far edge sites, utilizing a declarative approach, Kubernetes Clusters Life Cycle Management, and hybrid deployment and Bare Metal Automation. Thirdly, security is a key pillar of the initiative. Fourthly, energy efficiency will be prioritized, integrating measures for consumption and optimization mechanisms, with energy control integrated as a core aspect. Lastly, open-source and standardized API will be incorporated as a fundamental element of the initiative. The cloud native

infrastructure stack to be released by Sylva can be used to implement a validation platform aiming to verify that CNFs can be deployed on it using making the most of its capabilities.

7.2.4.6 Kubernetes

Kubernetes (K8s) is an open-source container orchestration system, designed and implemented by Google but currently hosted in Cloud Native Computing Foundation (CNCF). Due to its prevalent usage, it is considered a de-facto standard for managing container-based workflows. This has also led to efforts to use the technology in the orchestration of Containerized (or Cloud-native) Network Functions, and it was also identified in the Hexa-X project [Hexa-X-D6.2] as a cloud-native enabler and a possible part of the Management Function block, which is a specialised computing block for executing Management Functions. Although K8s was not designed to manage CNFs, there have been many extensions and variants that make it capable to do so.

7.2.4.7 Network Service Mesh

Network Service Mesh (NSM) is a CNCF [NSM23] project which takes the K8s service mesh concepts and brings them down to layer 3 of the networking stack. It enables connecting container workloads within and across clusters with L3 connectivity, allowing them to operate on this layer and above (e.g., network functions). The main concept facilitating this is a Network Service, which is “a set of Connectivity, Security and Observability features applied to traffic”. Effectively, it creates connections called vWires between containers and service endpoints where those packets are processed. It also allows more complex and selective compositions of services, as well as topology awareness when selecting endpoints. In spite of the fact that it is closely related to K8s and container-based workloads, it also allows for other types (VNF or PNF) to connect to the service.

NSM is not a management tool but can facilitate the management of workloads that operate on lower levels of the networking stack since their connectivity to the other workloads is handled elsewhere. The only thing needed by the management process is to mark them as requiring the Network Service. In the case of an extremely distributed user plane, this would largely simplify any management logic. In the last year NSM has seen many feature updates related to hardware acceleration (SR-IOV), packet forwarding (VPP and OVS), connecting external clients (VMs, Bare Metal), MTU handling and resilience. Its current roadmap contains adding support for eXpress Data Path and SR-IOV improvements related to VLAN tagging.

7.3 Related EC Research projects

In this section we provide description of Management and Orchestration solutions in SoTA in multiple research projects. They have been grouped according to the EC call to which they belong: 5GPPP ICT-52 (Smart Connectivity beyond 5G), SNS Stream B-01 (6G System Architecture) and Stream B-04 (Secure Service development and Smart Security).

7.3.1 5GPPP ICT-52 projects: Smart Connectivity Beyond 5G

5G PPP Smart Connectivity beyond 5G related projects cover the long-term transformation of networks into a distributed smart connectivity platform with high integration with (edge) computing and storage resources. ICT-52 projects started in January’21. Some of them have already finished in June’23 and others will finish in December’23 at the latest.

7.3.1.1 Hexa-X

The Hexa-X project [UER+21] has been an ambitious European initiative aimed at developing a sixth generation (6G) wireless communication platform in the scope of the 5GPPP. Hexa-X-II project is the continuation of Hexa-X project, which will adopt its key findings and evolve towards 6G materialization. Hexa-X project's vision was to connect human, physical, and digital worlds with a fabric of key enablers that will drive growth, global sustainability, trustworthiness, and digital inclusion. To achieve this vision, Hexa-X defined six main objectives that focus on developing a consolidated set of requirements for the future 6G platform in terms of Key Performance Indicators (KPIs) and Key Value Indicators (KVIs), systemizing and

integrating the 6G platform into a blueprint considering key values such as sustainability, inclusion, and trustworthiness, enabling breakthrough technologies and interfaces for connectivity services as well as novel digital services building on new network capabilities of sensing, compute, and AI. Additionally, Hexa-X aimed to ensure that the 6G system is realizable, implementable, and manageable in a resource-efficient manner while also contributing to a holistic European 6G view by impacting standardization activities. At the forefront of radio innovation, Hexa-X explored the use of terahertz frequencies for radio access, a promising solution that aims at delivering much higher data rates than current systems.

In addition, one of the critical innovations in the Hexa-X project was related to network management and orchestration (M&O). Specifically, the project analysed several advanced M&O mechanisms that enable agile and flexible network operations (i.e., integration of the compute-continuum on M&O operations, optimized placement, zero-touch automation, AI-driven orchestration, automation in multi-stakeholder scenarios, etc.), supporting a diverse set of use cases and requirements. The Hexa-X project defined a M&O architecture [HEX22-D62] that helped to cope with the aforementioned mechanisms through the integration of the following innovations:

- **Clear Separation of M&O and Managed Contexts:** The M&O architecture includes a clear separation of concerns between Managed Resources and Managing resources. Managed resources describe objects that can be operated and controlled, the relationships between these objects are represented in a class UML diagram called the Information Model (IM). Managed Objects (MOs) are instances of managed resources. In contrast, M&O resources provide the management tools (such as provisioning, monitoring, etc.) needed to act on Managed Objects.
- **Design Layer:** A new layer, the Design Layer, has been added to represent M&O-related operations involving third-party software providers. It exemplifies the implementation of cloud-native concepts in terms of bringing together development and operational teams through the use of DevOps approaches, facilitating how services are delivered and updated with a very high degree of automation (e.g., CI/CD pipelines).
- **Compute-Continuum:** Hyperscalers, private networks, and the extreme-edge domain have been included in the *Infrastructure Layer* to reflect the so-called compute-continuum.
- **Closed-Control Loops:** The architecture includes new CCL such as the *DevOps Control Loop* (i.e., CI/CD continuous operations between the MNO scope and the Design Layer) and the *Infrastructure Control Loop* (i.e., automates the infrastructure discovery, monitoring and management processes).
- **Network Functions Abstraction Sets:** Functions are associated with different groups at the *Network Layer* to ease function clustering and management. These sets include Radio Access Functions, Core Network Functions, Monitoring Functions, Security Functions, M&O Functions, AI/ML Functions, and third-Party Functions.
- **AI-based Orchestration:** The architecture includes an AI/ML Functions block, including AI/ML functions that can work in close connection with the Monitoring and the M&O Functions block. It also includes AI/ML-based collaborative components that can be distributed across all layers in the network.
- **API Layer:** A cross-layer Application Programming Interface (API) Management Exposure block has been integrated in order to aid the communications across layers and domains and enhance the capability exposure at all levels.

This architecture continues to represent a paradigm shift in the design of the telco stack based on switching from conventional network/service management systems (hard to evolve, with siloed managers coupled with point-to-point protocol interfaces) to a cloud-native management system (built out of modular composable management services that are offered for consumption using HTTP-based RESTful APIs). This service-based management architecture (SBMA) architecture provides for the creation of a group of management services, each of which represents a certain management capability (such as provisioning, performance assurance, or trace control) that enables the manipulation of a specific resource (e.g., network slice, CN function, etc.). Management functions, which may be mapped to vendor solutions, create and consume management services.

Finally, some relevant features of this M&O architecture were demonstrated and validated in [HEX23-D63] within two PoCs (section 4): i) Handling unexpected situations in industrial contexts, and, ii) Data-driven device-edge-cloud continuum management.

7.3.1.2 TeraFlow

TeraFlow is delivering a new generation open-source cloud-native SDN controller to provide secured and smart connectivity services to B5G/6G networks. This new SDN controller has been established as an open-source group (OSG) at ETSI as TeraFlowSDN (TFS). TeraFlowSDN controller is able to integrate with current NFV and MEC frameworks as well as to provide revolutionary features for flow aggregation, management (service layer), network equipment integration (infrastructure layer), and AI/ML-based security and forensic evidence for multi-tenancy. The project proposes an integrated solution for tackling various challenges of B5G networks to support service providers and telecommunication operators in their journey towards future networks. Section 7.1.6.2 has presented TeraFlowSDN as an state-of-the-art SDN controller.

7.3.1.3 Daemon

This project aims at developing AI solutions that are tailor made to fit networking problems. The Daemon project is designing an end-to-end Network Intelligence (NI) architecture that will fully integrate Beyond 5G functionalities. Innovations derived from this project will enable the efficient usage and optimal performance of radio and computational resources while minimizing the energy required to run the mobile network. These practical advances in NI will also enhance the reliability of 5G beyond the current SoTA.

The project will provide objective proofs of the benefits of a deep, structured and pragmatic integration of NI into network infrastructures. These proofs will be based on the achievement of a comprehensive range of Key Performance Indicators (KPIs) with regard to performance, reliability and sustainability of the proposed solutions in realistic settings including experimentation on testbeds and real-world network measurement data.

7.3.1.4 DEDICAT 6G

The major objective of DEDICAT 6G project is to create a smart connectivity platform (DEDICAT 6G platform) using artificial intelligence and blockchain techniques, that supports human-centric applications securely while being dependable, adaptive, ultra-fast and green. This platform will be remotely located in the cloud to support the whole system, monitor, and retrieve the necessary metrics required.

DEDICAT 6G focuses on techniques for combining current communication infrastructure with novel distribution of intelligence (data, computation, and storage) at the edge to allow flexible, and energy efficient real-time experience. The project also aims to design and build mechanisms for dynamic coverage extension by utilising new terminals and mobile client nodes, such as smart connected cars, robots, and drones. Additionally, DEDICAT 6G addresses security, privacy, and trust assurance, particularly for mobile edge services and enablers for novel human-digital system interactions.

The goal is to (i) use resources more effectively; (ii) lower latency, response time, and energy consumption; (iii) cut down operational and capital costs; and (iv) strengthen security, privacy, and trust. DEDICAT 6G focuses on four use cases, smart warehousing, enhance experiences, public safety and smart highway and the developed solutions will be tested through simulations and demonstrations in laboratory settings, and larger field trials using various assets and testing facilities.

All the nodes from the far edge to the core cloud are assumed part of the Network Function Virtualization Infrastructure (NFVI), simplifying the creation of network slices and instantiation of network services. According to ETSI NFV [NFV21] architecture, the NFV Orchestrator (NFVO), which is located at the control plane part of the network, orchestrates the NFVI. It is responsible for performing NFV Management and Network Orchestration (MANO) functions in a 5G-based environment. According to the results from the Network Operation Decision Making (NODM) Functional Component (FC), developed within this project, and actions taken by the Orchestration FC, also developed within this project, the DEDICAT 6G platform helps the NFVO when it comes to the instantiation of network services and network slices [DED-D2.4].

7.3.1.5 Marsal

MARSAL is a research project that focuses on the development and evaluation of a complete framework for the management and orchestration of network resources in 5G and beyond intelligent networks. The project aims to utilize a converged optical-wireless network infrastructure in the access and fronthaul/midhaul segments to achieve this goal. One of the key objectives of the MARSAL project is to develop novel cell-free based solutions that allow for a significant scaling up of the wireless access points in a cost-effective manner.

To accomplish this, MARSAL is exploiting the distributed cell-free concept and serial fronthaul approach to contribute innovative functionalities to the O-RAN project [ORA23]. Additionally, MARSAL aims to increase the flexibility of optical access architectures for Beyond-5G Cell Site connectivity via different levels of fixed-mobile convergence in the fronthaul/midhaul segments.

In terms of network and service management, MARSAL seeks to provide a comprehensive framework for the management of the entire set of communication and computational network resources. This is achieved by utilizing novel ML-based algorithms of both edge and midhaul data centers and incorporating the Virtual Elastic DataCenters/Infrastructures paradigm.

In the network security domain, MARSAL's goal is to introduce mechanisms that provide privacy and security to application workload and data. This is achieved by developing AI and Blockchain technologies to guarantee a secured multi-tenant slicing environment while allowing applications and users to maintain control over their data when relying on the deployed shared infrastructures.

7.3.1.6 AI@Edge

AI@EDGE [AIA23] project is developing a connect-compute fabric, specifically leveraging the serverless paradigm, for creating and managing resilient, elastic, and secure end-to-end slices. Such slices are expected to be capable of supporting a diverse range of AI-enabled applications. Privacy-preserving machine learning and trusted networking techniques are being used to ensure each stakeholder can use the platform without disclosing sensitive information.

The AI@EDGE architecture is divided into 2 main layers: i) Network and Service Automation Platform (NSAP) and ii) Connect-Compute Platform (CCP). The 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 CCP brings distributed computation over the cloud, far edge and near edge. The architecture is aligned with ETSI/MEC reference architecture and implement some components of ETSI/MEC (e.g., MEC Application Orchestrator (MEAO), MEC Platform Management (MEP(m)), Virtual Infrastructure Manager (VIM), NFVO) and O-RAN components (e.g., Near-Real-Time RIC (Near-RT RIC)) [AIA-D2.2]. Closed loops are an important enabler for automation in the AI@EDGE system architecture. There are three types of closed loops considered in AI@edge 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.
- NSAP Closed loop 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 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.

7.3.2 SNS StreamB-01: 6G System Architecture

SNS StreamB-01 System architecture projects support the vision of a massively digitised economy and society calling for intelligent connectivity and service provision across a huge number of heterogeneous domains, resources, and with an unlimited number of application requirements. These projects have started in January 2023, so there are no mature results to be considered yet.

7.3.2.1 DETERMINISTIC6G

The main objective of DETERMINISTIC6G is to establish the architecture of 6G in order to enable deterministic communication. This includes creating features that facilitate the progression of the current IEEE TSN standards, modelling wireless systems for 6G, developing edge computing solutions to support deterministic communication for edge-based applications, and designing relevant security features. Also, DETERMINISTIC6G project aims to create an end-to-end system architecture for deterministic communication that seamlessly integrates and interworks with TSN and DetNet standards. In addition, an advanced data-driven solution will be developed to enhance latency performance characterization and self-optimization for 6G network architecture.

The system architecture design for DETERMINISTIC6G will offer new capabilities that go beyond the existing Service Based Architecture (SBA) and will cater to several use cases for 6G. We can categorize the vision of DETERMINISTIC6G into three parts: a) Architectural aspects for enhanced E2E deterministic communication; b) Edge computing; and c) Improvements in the implementation for data-driven architecture.

7.3.2.2 PREDICT-6G

PREDICT-6G's mission is set towards the development of an end-to-end 6G (E2E) solution including architecture and protocols that can guarantee seamless provisioning of services for vertical use cases requiring extremely tight timing and reliability constraints. To succeed, the solution will target determinism network infrastructures at large, including wired and wireless segments and their interconnections. PREDICT-6G will develop a novel Multi-technology Multi-domain Data-Plane (MDP) overhauling the reliability and time sensitiveness design features existing in current wired and wireless standards.

PREDICT-6G aims to create a secure, modular, interoperable, and extensible deterministic network and management framework that automates the definition, provisioning, monitoring, fulfilment, and life-cycle management of end-to-end (E2E) deterministic services over multiple network domains, hiding the complexity of continuously balancing and re-configuring the constituent domain specific enablers to maintain a consistent E2E determinism. PREDICT-6G builds on top of three pillars:

- To extend the reliability and time sensitiveness features of IEEE 802.11 (targeting WiFi7/8) and 3GPP (targeting contributions to R19/20) networks, including APIs for the monitoring and control of such capabilities, enabling predictability. This pillar also considers the provision of a certain level of determinism (for specific applications) involving links or networks without native support (at layer-2) of deterministic features or not capable of guaranteeing a certain level of time sensitiveness, reliability, or predictability.
- To develop a Multi-technology multi-domain Data-Plane (MDP) jointly with an AI-driven multi-stakeholder inter-domain Control-Plane (AICP). This will enable the creation of E2E deterministic paths, by leveraging on IETF Deterministic Networking (DetNet) and Reliable and Available Wireless (RAW) mechanisms.
- To enhance the predictability of the network through intelligence, enabling the forecasting of the occupancy of network resources and the effect of accepting a new flow into the network. This feature will be enabled through AI and network digital twining approaches.

7.3.2.3 ADROIT-6G

ADROIT6G aims at building a 6G mobile network architecture that applies AI/ML-powered optimizations to reach high performance and automation, based on zero-touch paradigms. ADROIT6G also plans to migrate to a fully cloud-native network software, which can be implemented across heterogeneous edge-cloud platforms, including Non-Terrestrial Networks, with built-in security in the network user plane.

From a practical perspective, the envisioned ADROIT6G architecture evolves the SBA/SBMA of 5G networks towards a fully distributed and dynamic approach, with functional elements automatically deployed on-demand as cloud-native virtual functions across multi-stakeholder extreme-edge, edge and cloud domains. The goal is to build loosely coupled architecture elements and components integrated by means of a common, secure and scalable inter-domain communication framework to regulate the exposure of the capabilities offered

and consumed by each domain, enabling dynamic publish/subscribe of APIs and resources, and integrating access control and authorization mechanisms.

ADROIT6G architecture is built around three cooperative frameworks, dedicated to (i) management and orchestration, (ii) network control, and (iii) AI management, which operate over a programmable inter-computing and multi-domain infrastructure. These frameworks adopt a cooperative approach for distributed ML techniques and crowd-sourcing AI. Monitoring, analysis, decision and execution actions are implemented by distributed functions that operate at different layers (physical infrastructures, network slices, and services), realizing multiple real-time and medium-/long-term close loops.

The AI-driven Management and Orchestration framework implements the logic for end-to-end automated management of the inter-computing, multi-domain and multi-technology infrastructure. It provides features for the composition of multi-tenant services and slices across extreme-edge, edge, transport, and core domains, integrating public and non-public networks. The Belief-Desire-Intention (BDI) and AI-driven Unified and Open Control framework implements real-time and near-real-time closed control loops in support of full network automation and short-term resource allocation, based on current network condition and context-aware predictions. The AI/ML framework for CrowdSourcing AI assists the other two frameworks in their decisions, offering mechanisms for energy-efficient provisioning of AI agents, training, discovery and selection of AI models, as well as data management and distribution mechanisms.

7.3.2.4 DESIRE6G

DESIRE6G will design and develop a novel zero-touch control, management, and orchestration platform, with native integration of AI, to support eXtreme URLLC application requirements over a performant, measurable and programmable data plane. DESIRE6G aims to revolutionize mobile networks by using intent-based control and end-to-end orchestration to achieve near real-time autonomic networking, as well as a cloud-native unified programmable data plane layer that supports multi-tenancy. The control and orchestration plane of DESIRE6G employs a hybrid platform, with a centralized E2E Service Management and Orchestration Layer (SMO) and distributed intelligence moved to the network nodes.

This approach aims to support dynamic mobile traffic, reduce operational costs using AI/ML, simplify network operations, and leverage disaggregated networking technologies. The E2E SMO layer oversees orchestration and lifecycle management, generating guidelines for the agents to operate autonomously. The IBN approach starts from the Service Design Tools and defines policy rules to guide service behavior, such as quality of service, security, and performance. To secure agent interactions, DESIRE6G leverages Distributed Ledger Technology (DLT), and to support the execution of services across multiple domains, it uses blockchain-based federation at the SMO level.

At the data plane, DESIRE6G uses a cloud-native and microservice-based architecture toward RAN-Core convergence for 6G. A Serverless approach is adopted for mobile network and application-specific function deployment to take advantage of programming simplicity and automatic scaling and migration capabilities. DESIRE6G proposes a unified programmable data plane layer supporting multi-tenancy, providing a uniform approach for delivering slice-specific SW stacks utilizing multi-HW accelerators. Additionally, DESIRE6G enables seamless offloading of computations/functions to the network, introducing the "In network computing as a service" paradigm.

DESIRE6G's data, control, management, and orchestration plane are supported by a pervasive monitoring system that employs in- and out-of-band E2E network telemetry extending from the network to the user equipment or IoT terminal, ensuring the trustworthiness of the exchanged monitoring data.

7.3.3 SNS StreamB-04: Secure Service development and Smart Security

Target outcomes from this SNS stream qualify the needed level of reliability, trust and resilience that applies to a critical infrastructure like 6G based on a globally connected continuum of heterogeneous environments supported by the convergence of networks and IT systems to enable new future digital services.

7.3.3.1 CONFIDENTIAL6G

The ambition of CONFIDENTIAL6G [CON23] is to research post-quantum cryptography (PQC) enablers to develop tools, libraries, SDKs and other artifacts needed for confidential 6G technologies. These enablers will be articulated into three domains:

- **Confidential computing**. This will include lattice-based cryptography for Fully Homomorphic Encryption (FHE), Secure Multi-Party Computation (SMPC), Trusted Execution Environment (TEE) attestation handling and collaborative AI/ML.
- **Confidential networking**. This will include PQC TLS and other protocols, Zero-knowledge Proofs (ZKPs), confidential Smart Contracts (SC), Decentralized Identifiers (DIDs) and Anonymous Credentials (AC).
- **Confidential Edge and IoT.** This will include embedded FHE, PQC for constrained devices, large-scale networks of connected devices involved in federated learning, and cryptography to necessary support this.

Among these enablers, the confidential Edge and IoT is the most remarkable for Hexa-X-II scope, considering their applicability for a secure orchestration of extreme-edge devices in the IoT-edge-cloud continuum.

7.3.3.2 RIGOUROUS

RIGOUROUS [RIG23] will introduce a new holistic Smart service framework leveraging new ML and AI mechanisms, which can react dynamically to the everchanging threat surface on all orchestration layers and network functions.

The RIGOUROUS smart service framework is capable of ensuring a secure, trusted and privacy-preserving environment for supporting the next generation of trustworthy continuum computing 6G services along the full device-edge-cloud-continuum on heterogeneous multi-domain networks. This includes establishing compliance with the design of software, protocols and procedures, as well as AI-governed mechanisms to cope with the security-related requirements in the full DevOps lifecycle, from the service onboarding up to the day-2 operations. Main objectives are:

- Holistic Smart Service framework for securing the IoT-Edge-Cloud continuum lifecycle management.
- Human-Centric DevSecOps.
- Model-based and AI-driven Automated Security Orchestration, Trust Management, and deployment.
- Advanced AI-driven Anomaly Detection, Decision and Mitigation Strategies.

7.3.3.3 HORSE

HORSE [HOR23] will demonstrate how applications can leverage the ongoing evolution of 6G capabilities, as well as deal with the technology solutions, and system evaluation not yet foreseen, towards an omnipresent, smart and secure network service provisioning in the future network-of-networks landscape. HORSE project will address the challenge towards 6G infrastructure operation for smart connectivity and service management, and beyond, showing its effectiveness at the intersection of 6G connectivity, computing infrastructure management and security.

7.3.3.4 PRIVATEER

PRIVATEER [PRI23] aims to provide a privacy-centric security framework specifically designed for future 6G networks. More specifically, PRIVATEER aims to tackle four distinct privacy challenges which are closely linked to the top-tier security enablers that have been introduced in the 5G Advanced landscape. PRIVATEER framework will be evaluated through representative application scenarios, with focus on two vertical domains: i) Intelligent Transportation Systems, and ii) Smart Cities.

PRIVATEER's capabilities which are most relevant for Hexa-X-II are the privacy slicing-aware orchestration (in scope of M&O Enablers 3 and 4) and proof-of-transit (in scope of security enablers described in [HEX223-D21]). The first one will be used to provide trustworthy management of 3rd parties, keeping their credentials and data of their user securely stored, both needed for multi-tenancy support in resource sharing environments.