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 D3.2
Initial Architectural enablers

Co-funded by the European Union

6GSNS

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:	31/10/2023	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:	D3.2
Document Title:	Deliverable D3.2 Initial Architectural enablers
Editor(s):	Mårten Ericson (EAB), Merve Saimler (EBY), Ozgur Akgul (NFI), Panagiotis Botsinis (APP), Sokratis Barmpounakis (WIN), Milan Groshev (UC3)
Authors:	Ozgur Akgul (NFI), Michael De Angelis (NXW), Sokratis Barmpounakis (WIN), Jaap van de Beek (LTU), Giacomo Bernini (NXW), Panagiotis Botsinis (APP), Pere Garau Burguera (AAU), Panagiotis Charatsaris (ICC), Siddharth Das (TUD), Panagiotis Demestichas (WIN), Maria Diamanti (ICC), Toni Dimitrovski (TNO), Sameh Eldessoki (APP), Mårten Ericson (EAB), Afsaneh Gharouni (NGE), Milan Groshev (UC3), Alperen Gundogan (APP), Hasanin Harkous (NGE), Hamed Hellaoui (NFI), Selim Ickin (EAB), Paola Iovanna (EAB), Grigorios Kakkavas (ICC), Bahare M. Khorsandi (NGE), Slawomir Kukliński (OPL), Gerald Kunzmann (NGE), Hannes Larsson (EAB), Marvin Manalastas (NFI), Antonio de la Oliva (UC3), Ece Ozturk (NGE), Torgny Palenius (SON), Symeon Papavasliou (ICC), Ignacio Labrador Pavón (ASA), Merve Saimler (EBY), Erin Seder (NXW), Vivek Sharma (SON), Mohammad Soliman (NGE), Heiko Straulino (NGE), Olav Tirkkonen (AAU), Nassima Toumi (TNO), Vasilis Tsekenis (WIN), Stefan Wänstedt (EAB), Zi Ye (LTU), Milan Zivkovic (APP)
Contractual Date of Delivery:	31/10/2023
Dissemination level:	PU[1]
Status:	Final
Version:	1.0
File Name:	Hexa-X-II_D3.2_v1.0

Revision History

Revision	Date	Issued by	Description
0.1	2023-02-01	Hexa-X-II WP3	Template for Deliverables/IRs
0.2	2023-06-21	Hexa-X-II WP3	Sent to cross-WP review
0.3	2023-09-01	Hexa-X-II WP3	Sent to external review and PMT
0.4	2023-09-29	Hexa-X-II WP3	Sent to GA for approval
1.0	2023-10-31	Hexa-X-II WP3	Final version, submitted to EC

[1] PU = Public

Abstract

This is the first public deliverable from WP3, called D3.2 “Initial Architectural enablers” from the Hexa-X-II project. Hexa-X-II is a flagship initiative bringing together key stakeholders in Europe for 6G research, continuing the Hexa-X flagship project work.

In this report results from work in WP3 are presented, which deals with the 6G architecture. The overarching objective of WP3 is to develop a 6G architecture framework and innovative enablers for beyond communications and data driven architecture to power new services, modular cloud-native network for improved signalling and new access and flexible topologies for improved reliability.

The main areas of the 6G architecture are the data-driven architecture, modular network, new access and flexible topologies, beyond communication and finally, the cloud transformation.

Keywords

6G architecture, data-driven architecture, AIaaS, modular networks, JCAS, flexible topologies, cloud transformation

Executive Summary

This is the first public deliverable from WP3, called D3.2 “Initial Architectural enablers” from the Hexa-X-II project. Hexa-X-II is a flagship initiative bringing together key stakeholders in Europe for 6G research, continuing the Hexa-X flagship project work.

In this report results from work in WP3 are presented, which deals with the 6G architecture design. The overarching objective of WP3 is to develop a 6G architecture framework and innovative enablers for beyond communications and data driven architecture to power new services, modular cloud-native network for improved signalling and new access and flexible topologies for improved reliability.

The main areas of the 6G architecture are the data-driven architecture, modular network, new access and flexible topologies, beyond communication and finally, the cloud transformation.

In the realm of 6G data-driven architecture, a set of AI enablers assumes a pivotal role in unlocking the power of Artificial Intelligence (AI). These enablers, comprising architectural means and protocols, Machine Learning Operations (MLOps), Data Operations (DataOps), AI as a Service (AIaaS), and Intent-based management, collectively form a robust framework for seamlessly integrating AI into the fabric of 6G networks.

For enabling flexibility without increasing complexity, 6G needs an easily deployable architecture of modules that can grow and adapt on the current needs. Network modularity targets to decompose the 6GS into orthogonal building blocks (i.e., network functions, services and interfaces) with the right level of granularity. Modularisation of the network functions needs to be performed with an E2E vision, considering not only the network function granularity but also the necessary interfaces and deployment options to incorporate existing and new use cases such as NTN, programmability and Everything as a Service (XaaS).

New access and flexible topologies consist of the “network of networks” enabler, which deals how to integrate subnetworks and Non-Terrestrial Networks (NTN). To support new accesses, new 6G multi-connectivity innovations are proposed, both for the terrestrial network but also between the Terrestrial Network and NTN.

The beyond conventional connectivity is expanding the network’s scope by processing data, generating insights, and delivering added value from societal, innovation, and business perspectives. Examples of new services comprise sensing, enhanced localization and tracking, compute-as-a-service, and AI-as-a-Service.

Cloud computing became the de-facto standard for managing web-based and web-scale applications. While this architectural paradigm is suitable for a big subset of multimedia human-scale applications, it shows its limitations when it comes down to supporting the upcoming latency sensitive 6G use cases. The cloud-based architectures have some limitations when it comes down to latency, throughput, connectivity and security and interoperability and therefore there is a need for what we call the cloud transformation, i.e., adapting the cloud for the 6G requirements.

Table of Contents

List of Tables

List of Figures

Acronyms and abbreviations

Term	Description
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
5GC	5G Core
5G-PPP	5G Infrastructure Public Private Partnership
6G	Sixth Generation
6GS	6G System
AI	Artificial Intelligence
AP	Access Point
AIaaS	Artificial Intelligence as a Service
AMF	Access and Mobility Management Function
AUSF	Authentication Service Function
API	Application Programming Interface
AR	Augmented Reality
ARQ	Automatic Repeat reQuest
B2B	Business-to-business
B2C	Business-to-consumer
BCS	Beyond Communication Services
BLER	Block Error Rate
BS	Base Station
BSS	Business Support Systems
CA	Carrier Aggregation
CaaS	Compute-as-a-Service
CAPIF	Common API Framework
CC	Compute Continuum
CD	Continuous Deployment
CI/CD	Continuous Integration/Delivery
CM	Continuous Monitoring
cMEC	Constrained Multi-access Edge Computing

CN	Core Network
CNF	Core Network Function
COTS	Commercial-Off-The-Shelf
CP	Control Plane
CPSSO	CPS Status Observer
CRUD	Create, Read, Update, Delete
CSI	Channel State Information
CSMF	Communication Service Management Function
CSP	Communication Service Provider
D2D	Device to Device
DC	Dual-Connectivity
DCF	Data Centre Features
D-MIMO	Distributed MIMO
DN	Data Network
DSP	Digital Service Provider
DSS	Dynamic Spectrum Sharing
DT	Digital Twin
E2E	End-to-End
EDCA	Evolved Data Collection Architecture
eMBB	enhanced Mobile Broadband
EN-DC	E-UTRA – NR Dual Connectivity
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
EVM	Error Vector Magnitude
FA	Federation Agent
FaaS	Function as a Service
FDD	Frequency Division Duplex
FL	Federated Learning
FR	Frequency Range
GCL	Global Connectivity Layer
GHz	Gigahertz

GSMA	Groupe Special Mobile Association
GTP	GPRS Tunnelling Protocol
GW	Gateway
HARQ	Hybrid ARQ
HGNF	Highly granular network function
HFL	Hierarchical Federated Learning
HW	Hardware
IBN	Intent-Based Networking
IBS	Intent-Based Systems
IETF	Internet Engineering Task Force
IID	Independent Identically Distributed
ISAC	Integrated Sensing and Communication
ISL	Inter-Satellite Link
IoT	Internet of Things
IP	Infrastructure Provider
IT	Information Technology
JCAS	Joint Communication and Sensing
KPI	Key Performance Indicators
KVI	Key Value Indicators
LEO	Low Earth Orbit
LoS	Line of Sight
LTE	Long Term Evolution
LTM	Layer 1-Layer 2 Triggered Mobility
MA	Management Agent
MAC	Medium Access Control
MANO	Management and Orchestration
MARL	Multi-Agent Reinforcement Learning
MaaS	Mobility-as-a-Service
MC	Multi-Connectivity
MEC	Multi-access Edge Computing
MgtN	Management Node

MIMO	Multiple-Input and Multiple-Output
ML	Machine Learning
mMTC	massive Machine Type Communication
MN	Master Node
MNO	Mobile Network Operator
MR	Mixed Reality
MRSS	Multi-RAT Spectrum Sharing
MT	Multi Technology
M&O	Management and Orchestration
NaaS	Network as a Service
NF	Network Function
NFV	Network Function Virtualization
NFVI	Network Function Virtualization Infrastructure
NR	New Radio
NRAR	Network Reliability, Availability and Resiliency
NRF	Network function Repository Function
NSA	Non-standalone
NTN	Non-Terrestrial Network
NW	Network
NWDAF	Network Data Analytics Function
O-RAN	Open Radio Access Network
OS	Operating System
OSS	Operations Support System
PCell	Primary Cell
PDCP	Packet Data Convergence Protocol
PCA	Principal Component analysis
PCT	Procedure Completion Time
PCF	Policy Control Function
PDU	Packet Data Unit
PHY	Physical
PNF	Physical Network Function

PoC	Proof of Concept
PPDR	Public Protection and Disaster Relief
PSCell	Primary Secondary Cell
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RCS	Radar Cross-Section
REC-EXEC	REsource orchestrator for Continuum across EXtreme-edge, Edge, Cloud
RF	Radio Frequency
RIS	Reconfigurable Intelligent Surfaces
RL	Resource Layer
RLC	Radio Link Control
RLF	Radio Link Failure
RNF	Radio network functions
RRC	Radio Resource Control
RRM	Radio Resource Management
RU	Radio unit
SA	Standalone
SaaS	Sensing-as-a-Service
SBA	Service-Based Architecture
SCCP	Service-Centric Control Plane
SCG	Secondary Cell Group
SDN	Software-Defined Networking
SDO	Standards Development Organizations
SDP	Semidefinite program
SLA	Service-Level Agreement
SLAM	Simultaneous Localization and Mapping
SMF	Session Management Function
SN	Secondary Node
SNF	Shared network functions

SNS	Smart Network and Services
SNS JU	Smart Network and Joint Undertaking
SP	Service Provider
SoC	System on Chip
SoTA	State-of-The-Art
SVM	Support Vector Machines
TCP	Transmission Control Protocol
TDD	Time Division Duplex
THz	Terahertz
TN	Terrestrial Network
UAV	Unmanned Aerial Vehicle
UDM	Unified Data Management
UDR	Unified Data Repository
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low-Latency Communication
V2X	Vehicle-to-everything
VNF	Virtual Network Function
VNO	Virtual Network Operator
VR	Virtual Reality
WG	Working Group
WP	Work Package
XaaS	Everything as a Service
XR	Extended Reality
ZSM	Zero-touch network and Service Management

1 Introduction

The Hexa-X-II project is a flagship initiative bringing together key stakeholders in Europe for 6G research, continuing the Hexa-X flagship project work. Hexa-X-II includes the key industry players in telecom as well as new value chain for future connectivity solutions and major research institutes.

The Hexa-X-II project comprises several work packages that study different areas. In this report results from work in WP3 are presented, which deals with the 6G architecture.

The overarching objective of WP3 is to develop a 6G architecture framework and innovative enablers for beyond communications and data driven architecture to power new services, modular cloud-native network for improved signalling and new access and flexible topologies for improved reliability.

This is the first public deliverable from WP3, called D3.2 “Initial Architectural Enablers”.

1.1 Objective

The main objective of this document is to make an initial description of the different studies in WP3. Based on the different studies, initial descriptions of the enablers are made. The description includes the reason and motivation why this enabler is important for the 6G architecture. The long-term objectives of WP3 are found in Table 1-1. The three WP3 objectives includes the 6G architecture for AI and beyond communications, how to combine the cloud technology for a modular, scalable and extendable architecture and an architecture for flexible topologies.

Table 1-1 WP3 Objectives

Objective	Objective description	Chapter
WPO3.1	Develop and analyse a 6G architecture framework and new innovative enablers for the beyond communications and data driven architecture, identify requirements a data-driven architecture will have on protocols, interfaces, data, and network nodes.	Chapter 4 and 7
WPO3.2	Define and analyse solutions that combine cloud technology flexibility with distributed processing nodes into self-contained modules with minimum dependency that can be used to extend and scale the network deployments in stepwise manner	Chapter 5 and 8
WPO3.3	Develop and analyse new access for flexible topologies and local communications, including different types of multi-connectivity, node roles and node coordination, as well as design control and management solutions for programmable and context-aware transport	Chapter 6

1.2 Structure and 6G E2E architecture

This document is structured as follows: Chapter 2 gives a brief overview of previous and current initiatives of designing 6G. Chapter 3 sets the scene by describing the use cases applicable to the enablers in WP3. Chapter 4 describes the AI enablers for a data driven architecture. Chapter 5 describes the network modularisation, i.e., how to build an architecture of modules that can grow, and change based on current needs. Chapter 6 describes new 6G access for flexible topologies and local communications. Chapter 7 describes the “beyond communications”, the new services in 6G not based on MBB communication, such as sensing and computing. Chapter 8 describes the virtualization and cloud transformation. Chapter 9 is the conclusions and Chapter 10 contains the references.

Another way to see the structure of this document is to refer to Figure 1-1, which shows the 6G E2E system blueprint from [HEX2-D21]. Figure 1-1 depicts an end-to-end architecture “blueprint” and consists of four

different layers, namely Application, Network-centric application, Network functions, and Infrastructure and compute layers. There is also something called “Pervasive functionalities”. These functionalities can reside in any of the four layers. These functionalities include DataOps (involving data selection and collection, potentially sensitive and voluminous), exposure of AI services, intent-based management, and AI Framework (MLOps) and is addressed by Chapter 4. Chapter 5 addresses the network function layer, and how the RAN and CN NFs can communicate with each other in an efficient manner. Chapter 6 also resides in the Network function layer, but deals more with the subnetwork, UEs and the architectural aspects (i.e., protocols, interfaces etc) of the relation between RAN and UEs. Chapter 7 belongs to the “Beyond-communication functions” box (e.g., sensing and compute offloading), but also works with how to expose the service. Finally, Chapter 8 handles the cloud continuum and also the compute and devices in the Infrastructure and compute layer.

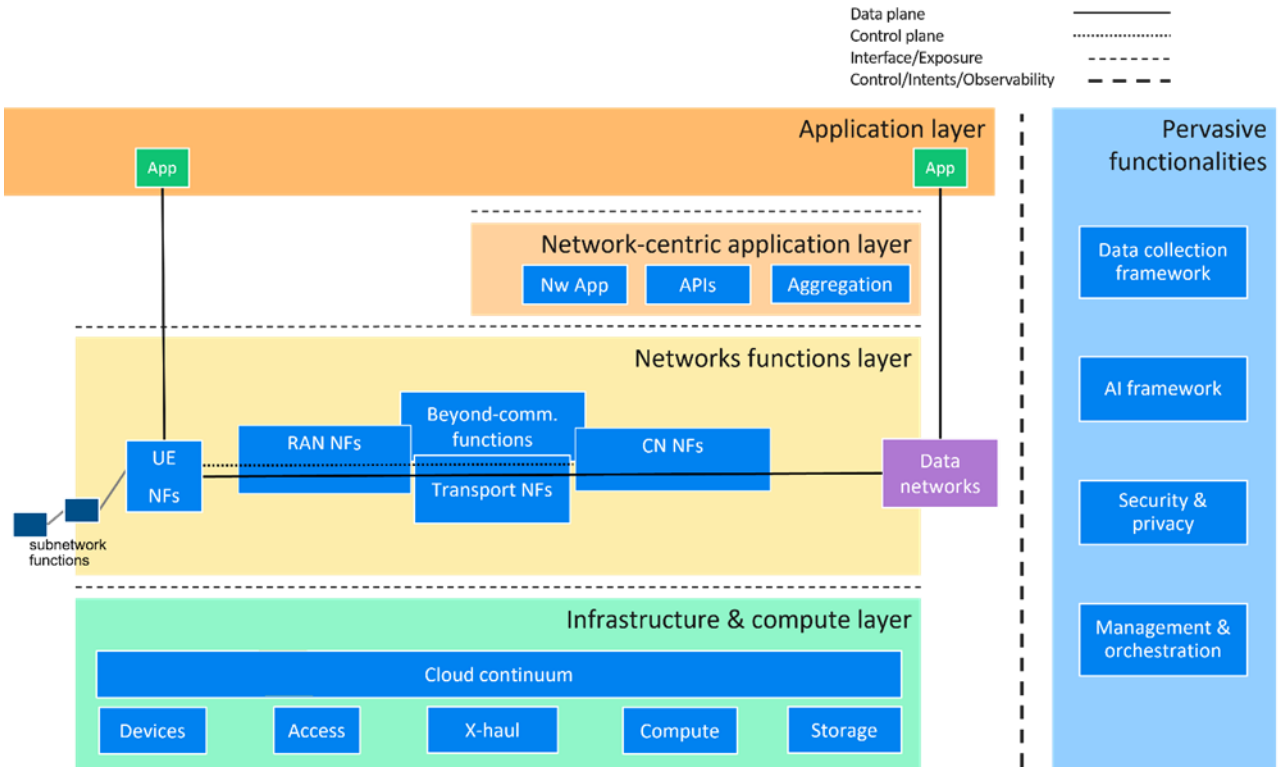

Figure 1-1 Initial 6G E2E system blueprint [HEX2-D21].

2 Outlook and previous work

Every day the users demand more from their cellular connections, where "more" can be new services, higher bit rates, better coverage or all. Therefore, cellular systems need to be continuously improved, resulting in new releases of the cellular standards and new improved hardware, both in user equipment and in networks. When there appears to be limits to how much more the current generation of cellular systems can be improved and when advances in technology provides opportunity, work with a subsequent generation commences.

Therefore, different groups, initiatives and projects have started working on shaping the 6th generation of cellular networks. The abovementioned improvements comprise all aspects of the system, not only the services to users. Consequently, changes in society can create requirements that were not anticipated during the design of previous generations of cellular systems. Examples of such changes in society are that today most users want to access contacts with various institutions digitally, AI has evolved from a research project to actual usable applications, and sustainability becomes increasingly important. All of the above affects the work on designing the 6th generation of cellular networks. In the following some major initiatives are described.

According to the 5G Infrastructure Association (5GIA) [5GIA21] a direct integration of many resources, such as networking, computation, and sensing, is expected in 6G. As a result, the scope of the 6G Architecture is expanded beyond RAN and CN to include terminals and data centres to assure complete, end-to-end resource awareness.

In the report [M.2516-0] the ITU-R FTTR presents technologies for a possible 6G particularly targeted at enhancing the radio interface. Technologies include Sub-THz (above 100 GHz) frequencies, extreme MIMO, multiple physical dimension transmission including reconfigurable intelligent surfaces (RIS), advanced modulation, coding and multiple access schemes, co-frequency co-time full duplex communications, as well as ambient backscatter communication.

In Asia, there are several initiatives looking at various aspects of 6G. The IMT2030 (6G) Promotion Group [6GPG] was established in China to advance 6G research and creates a global forum for perspective exchange. Another initiative, the MSIT 6G Research Program in South Korea [MSIT22] is developing a smart strategy to be the first country to deploy 6G networks. The Japanese government launched the Japan 6G/B5G promotion plan [B5G6G] to encourage research and development of 6G wireless communications services. To investigate technological advancements and potential commercial applications, the Department of Telecommunications (DoT) of the Ministry of Communication established a task force under the Technology Innovation Group (TIG) on 6G technology. The task forces will facilitate pre-standardization, manufacturing, research and development, and a market readiness framework.

Currently India is deploying 5G at a very high pace, working very hard to provide affordable services to most. The introduction of 5G is a major change to the Indian market since many users are still on 2G. Spectrum was issued as recently as August 2022. Recently, an Indian 6G initiative was unveiled called Bharat 6G Initiative [Bha6G], which from a distance resembles the European projects and brings together a diverse consortium of stakeholders, public and private companies, academia, research institutions and organisations providing standardization. The primary objective is to define the requirements on 6G from the Indian society. Requirements are also in this case both technical and societal. According to [Bha6G] this initiative positions India as one of the key players in the cellular landscape as well as the country's commitment to drive socio-economic progress.

In North America, the Next G Alliance initiative [NGA] targets advancing the leadership of North American Wireless technology and the leadership of the private sector in this domain. Moreover, the Resilient and Intelligent NextG Systems (RINGS) program [RINGS] aims to advance research in areas that could have important impact on emerging Next Generation (NextG) wireless and mobile communication, networking, sensing, and computing systems, as well as services at global scale [NextG]. The program specifically focuses on significantly improving the resiliency of such networked systems, besides other performance metrics.

In Europe, Hexa-X [HEXA] is a flagship project for 6G vision and intelligent fabric of technology enablers connecting human, physical, and digital worlds. Different architectural enablers for building intelligent, flexible, and efficient networks were identified within the scope of this project. Other European initiatives are, the Finnish 6G Flagship, a research initiative centred on "6G Enabled Wireless Smart Society and Ecosystem"

[6GFlagship], and 6G-ANNA [6G-ANNA], a German-funded project looking at holistic approaches for 6G mobile networks.

3 Use cases

The use cases described below represents the 6G use cases that are applicable to the WP3 studies and enablers. They are here included to give an introduction on how the different enables can be used in a 6G system.

3.1 Immersive telepresence for enhanced interactions

This use case family consists of the mechanisms supporting being present and interactive anytime anywhere, using all human senses if so desired [HEX-D13]. Immersive telepresence refers to an expanded version of the virtual world where users and devices can interact with each other in a more complete and seamless manner. Mixed reality and holographic telepresence may offer an interaction between mutually remote people reminiscent to being physically present in the same place [HEX-D13]. In the context of 6G mobile networks, the increased capabilities of the network can be leveraged to create a more realistic and engaging virtual environment that allows merging different worlds i.e., physical, digital, and human. It aims at creating seamless integration between digital and physical systems, requiring real-time communication between devices, sensors and machines. Examples of use cases include fully merged cyber-physical worlds [HEX2-D11], merged reality game/work, immersive education, as well as immersive sport events and entertainment.

The network requirements of this use case family include low latency, high data rates and high reliability [26.928, 26.998]. 6G also needs to support fundamental features such as mobile media support among multiple users, user data management and data security [22.856].

In the immersive telepresence use cases, users can move in and between the worlds and can interact with them in a very dynamic manner. The need to download new 3D aspects prior to the user interaction poses not only very strict Key Performance Indicators (KPI) but also the need for a highly flexible network support. Moreover, the location-based services need to identify the location of the UE and expose this information to the authorized 3rd parties. The wider usage of bandwidth-hungry and latency-sensitive applications (e.g., VR 360 video streaming) forces 6G to provide runtime scalability and ultra-high flexibility to dynamically meet the Quality of Service (QoS) / Quality of Experience (QoE) requirements [TCZ+22]. Therefore, the orchestration and the allocation of the edge resources (e.g., storage and computation) become a critical aspect to maximize the user QoE [LBZ+21].

Subnetworks, flexible topologies, multi-connectivity and E2E context-awareness management may also enable the immersive telepresence use case family. Subnetworks can provide high-throughput and low-latency communication as well as enable local information exchange between the nodes of the subnetwork, which could be useful in such use cases. Flexible topologies are particularly important when user mobility is taken into consideration. Multi-connectivity may be used for data offloading and higher resilience since it enables network connection via multiple paths and/or radio access technologies. E2E context awareness management may provide dynamical adaptation of the network to the user's sensory context.

Immersive telepresence promises to create immersive experiences not only for the consumers but also in the enterprise and the industrial applications. Vertical domain driven use cases such as immersive education, mine inspection, remote rendering and disaster handing can utilize the network beyond communication services such as Joint Communication and Sensing (JCAS) enhanced localization and tracking, Compute as a Service (CaaS), Artificial Intelligence as a Service (AIaaS), digital twining, etc.

The capability of using computing resources in the locality of the user may also be required for enabling the immersive telepresence use cases. For example, computing appliances used for immersive telepresence may belong to the user or the service provider and be in the office or the user's home. Such appliances may be integrated in the computing continuum to host functions used by the network to provide such services.

3.2 From robots to cobots

The "Robots to Cobots" use case encompasses the transition from traditional robots to collaborative robots (cobots) in diverse industries and applications [HEX2-D11]. It involves the integration of cobots to work alongside human operators, fostering cooperation and synergy between humans and machines. Unlike traditional robots, cobots are designed to collaborate with human workers in shared workspaces, rather than performing autonomous tasks independently. Cobots have the capability to support human workers by

undertaking physically demanding or hazardous responsibilities, thereby increasing productivity, and enhancing worker safety. Equipped with advanced sensors and algorithms, cobots can detect the presence of humans and respond appropriately, ensuring safe interaction. By collaborating with cobots, human workers can shift their focus to more complex and cognitive tasks, while leaving repetitive or strenuous activities to the robots. This symbiotic approach not only improves operational efficiency but also enhances job satisfaction and facilitates skill development for human workers. The "Robots to Cobots" use case exemplifies the transformative potential of human-robot collaboration in creating a harmonious work environment where both humans and robots thrive together.

Within the "Robots to Cobots" use case, the adoption of cobots in various industries and applications is achieved through advanced techniques and infrastructure. Multi-domain training using a Split Neural Network architecture is employed, along with multi-task training to reduce computation and enhance model reusability. Additionally, use case-specific APIs are identified to enable cobots to run their AI services effectively. Furthermore, automated deployment of AI cloud-native functions and the implementation of MLOps pipelines for lifecycle management and continuous monitoring of AI/ML models used by cobots are essential components. These advancements aim to enhance the capabilities and efficiency of cobots, enabling them to perform a wide range of tasks effectively and contribute to a harmonious work environment where humans and robots collaborate seamlessly.

In the realm of 6G systems, the transition from conventional command and control robots to cobots holds significant importance. Cobots within this new technological ecosystem can establish relationships with other cobots and humans to successfully accomplish complex tasks. With their multi-dimensional ambient sensing, computing, model building, and communication capabilities, cobots execute projects cohesively. One exemplary application is real-time cooperative safety protection, where cobots collaborate with other cobots, security staff, and remote security controllers to provide security within a specific geographical area [22.916]. The symbiotic relationship between cobots and humans enhances operational sustainability. An example of this collaboration is the cooperative gathering of measuring data, where a group of cobots collaborates in data collection to save energy, enhance outcome quality, or achieve both objectives. Cobots offer efficient resource usage, high flexibility, and situation-aware communication, enabling them to support group tasks effectively. They excel in individualized on-demand tasks, enabling lot size one production and effectively utilizing novel production methods such as additive manufacturing. Meeting the requirements of 6G systems necessitates flexibility in network topologies and resource allocation, functional support for extreme requirements, and closed-loop control of network functionality. To address dynamic transmission requirements, transmission opportunities can be allocated based on cobots' intents, such as fusion levels and traffic. Edge computing reduces computational load, while efficient data transmission ensures seamless collaboration among cobots and synchronization with collaborating groups [22.916].

For successful synergies among cobots, achieving a high level of clock synchronization accuracy is crucial. Time synchronization ensures safe functioning and enables time-sensitive collaborative interactions between humans and robots when they share a space to achieve a common goal. Although the Precision Time Protocol (PTP) protocol offers nanosecond-level accuracy, it may still experience clock deviations and jitters under varying network conditions]. To overcome these issues, quantum time synchronization can stabilize the oscillation frequency of qubits, achieving femtosecond-level accuracy through simulations [NPS+23]. Quantum synchronization involves integrating a quantum physical layer into the existing classical architecture, not only resolving synchronization challenges but also providing secure communication paradigms.

In the context of the "Robots to Cobots" use case, the integration of a network of networks and trustworthy, flexible topology management plays a significant role. This integration involves leveraging subnetworks and device-to-device (D2D) connectivity to facilitate communication and coordination between cobots and other networked devices. The mobility of cobots necessitates dynamic adaptation to changing contexts, which is facilitated by multi-connectivity and end-to-end (E2E) context awareness management. These advancements enhance the resilience and reliability of the network, ensuring seamless communication and collaboration between cobots and other networked components. This utilization of a network of networks and flexible topology management supports the dynamic nature of cobots, enabling efficient and context-aware communication within the collaborative robotics ecosystem.

Moreover, the network's role extends beyond communication services within the "Robots to Cobots" use case. It provides additional functionalities such as sensing, compute-as-a-service, indoor localization, mapping, and massive twinning. These services empower cobots to perceive and interact with their environment, including other robots, humans, and objects, in real-time. The network enables cobots to navigate with high accuracy, safety, and efficiency while showcasing adaptive behaviour. To support cobot autonomous/manual control, AI/ML-driven decision-making, and collaborative and coordinated operations, both computing and communication capabilities are required. The advanced environment perception facilitated by the network enhances the capabilities of cobots, enabling them to perform tasks effectively and interact seamlessly in dynamic work environments.
Furthermore, cloud or edge robotics is leveraged in the "Robots to Cobots" use case to deploy processing and control algorithms for robot fleets. Depending on the desired latency, computing devices can be located either in the cloud or at the edge of the network. This approach greatly benefits cobots as it simplifies the coordination of control instances. Managing a single program in the cloud to control the entire fleet is easier than coordinating multiple distributed instances. While the control may be logically centralized, it can run in a distributed manner across the computing continuum. This approach enables efficient and coordinated control of cobots, enhancing their capabilities and enabling seamless collaboration in various work environments.

3.3 Sustainable development

Sustainable development in 6G has two major meanings, firstly it means that the 6G system should be developed to be sustainable concerning, e.g., choices of material, methods of deployments, power consumptions, etc. Secondly, the 6G system provides characteristics that enable sustainable services to society, e.g., coverage everywhere or inclusiveness, trustworthiness, etc. Therefore, the 6G system will be very well suited to provide solutions contributing to meet the United Nation Sustainable Development Goals (UN SDG) or helping verticals to reduce their environmental impact [HEX-D12]. Among use cases, this use case family addresses the need for inclusion by delivering key digital services, such as providing health services to remote or isolated areas and monitoring large areas in nature to help protection of the environment. Requirements of this use case family include extreme and sustainable performance, global service coverage and trustworthiness.

To provide trustworthiness in 6G systems, preservation of user privacy is important. Examples of activities needed to support privacy are privacy-aware data classification, investigations of privacy-preserving data collection and learning methods. Further, coupled to learning and expected coverage, there is a need for distributed machine learning model training that harnesses the data and computing resources of geographically dispersed end-user devices. Finally, to orchestrate all the above model-driven, privacy-aware, and distributed intelligence solutions need to be developed, that can manage the system. The orchestrated system may also include a large number of heterogeneous 6G network slice instances, all addressing the specific requirements of the use case.

To meet requirements on coverage, existing technology needs to be revisited to find how coverage can be provided everywhere at a reasonable cost. One important concept is the network of networks (e.g., NTN and TN integration, subnetworks), which really provides global service coverage. Having a network that may adapt dynamically to the application’s context is a piece of technology needed to achieve coverage. The adaptability is, among others, made possible by trustworthy flexible topologies and E2E context awareness management.

New services put new requirements on cellular systems in general. Using 6G as a bearer for sustainable services or services that support sustainability in societies may drive development of services that do not exist yet.

3.4 Massive Twinning

Massive twinning is a virtualized model and live representation of a physical asset or a digital representation of the item function and activity. The requirements to accomplish this, -especially in large scales, in terms of devices and models- will test the limits of current technology, as massive amounts of information will need to be transferred with minimal latency. Massive twinning, or the application of the fundamental concept of Digital Twins (DT) in a variety of use cases, aims to expand and improve the production and manufacturing. It includes managing our environment, transportation, logistics, entertainment, social interactions, digital health, defence,

and public safety. Examples of use cases include digital twins for manufacturing, immersive smart cities, and digital twins for sustainable food production.

Besides the latency requirements, high reliability, availability, safety, maintainability, integrity, and data rates will be required. As part of the computation requirements, e.g., for AI/ML-related processing workloads related to the digital representation of the physical assets, their functions and activities, high interpretability levels are required. Finally, as part of localization and sensing requirements, high service availability, high safety, maintainability, integrity, and high location accuracy.

Massive Twinning may be enabled by MLOps pipelines to collect the data for generating and updating the Digital Twin models through continuous monitoring. Enhancing the existing infrastructure for data exchange and enabling AI solutions to gather the necessary data for predicting and designing massive twinning models, will facilitate the efficient flow of data required to train and optimize the AI models, leading to more accurate predictions and improved design outcomes. Context-aware connectivity, where information about the use case environment may be gathered at any given time and the system behaviour may be adapted accordingly. Finally, services beyond communications, such as sensing and localisation for leveraging real-time data from various sources, including radio, towards highly accurate representations of physical objects or systems will be an important driver for this use case. IoT data will be also leveraged (such as environmental data from sensors, or data from wearables) in order to augment the digital representation of the environment and physical assets.

4 AI enablers for data-driven architecture

In the realm of 6G data-driven architecture, a comprehensive set of AI enablers assumes a pivotal role in unlocking the transformative power of AI. These enablers, comprising architectural means and protocols, Machine Learning Operations (MLOps), Data Operations (DataOps), AI as a Service (AIaaS), and Intent-based management, collectively form a robust framework for seamlessly integrating AI into the fabric of 6G networks. Architectural means and protocols provide the foundation for efficient data flow and communication, ensuring the interoperability and scalability of AI systems. Within this context, MLOps assume paramount importance it enables the efficient deployment and management of machine learning models. MLOps ensures that these models can seamlessly adapt to changing network conditions, process voluminous data, and deliver real-time results- (inference results), thereby enabling advanced applications such as autonomous vehicles and augmented reality. Similarly, DataOps focuses on designing and maintaining a distributed data architecture, encompassing crucial disciplines such as data collection, transformation, and quality control. In the realm of 6G, DataOps ensures the reliability, accuracy, and availability of collected data, thereby promoting trustworthiness and effective data management for various applications. Furthermore, AIaaS provides a comprehensive framework for AI functionalities, allowing for the seamless exposure of AI capabilities from 6G networks of AI models within 6G networks. AIaaS empowers applications with AI-driven decision-making and automation, facilitating closed-loop network and service automation. Lastly, Intent-based management, in conjunction with zero-touch networks, addresses the inherent complexity resulting from the adoption of AI and the diverse use cases within 6G networks. Intent-based management is a powerful AI enabler that transforms the way complex systems, such as 6G networks, are configured, managed, and optimized. At its core, intent-based management leverages the capabilities of AI to bridge the gap between high-level objectives and technical implementation, leading to more efficient, adaptive, and scalable management processes. By effectively leveraging the capabilities of MLOps, DataOps, AIaaS, and Intent-based management, 6G networks can fully harness the potential of AI, ensuring efficient AI/ML model deployment, AI exposure capabilities, reliable data processing, and autonomous network management. Ultimately, this comprehensive approach delivers transformative capabilities to a wide range of applications and services within the 6G ecosystem.

4.1 Data-driven architectural means and Protocols

Architectural means and protocols play a pivotal role in facilitating the integration of AI technologies within the framework of 6G networks. As the landscape of 6G continues to evolve, 6G promises to bring forth unprecedented capabilities, including ultra-high-speed connectivity, predictable/bounded low latency low latency, and extensive device connectivity. To fully capitalize on the potential of AI within this context, it becomes imperative to establish a robust architectural framework and implement appropriate protocols. The integration of AI into 6G networks aims to empower intelligent decision-making, adaptive resource allocation, efficient network management and efficient network operation due to the advancement in the physical layer by AI-driven air interface design. By harnessing AI algorithms and techniques, 6G networks can optimize overall network performance, enhance user experiences, and enable a wide array of innovative applications. The motivation behind incorporating AI into 6G networks stems from the need to effectively address the challenges posed by the dynamic and complex nature of the network environment, while maximizing the utilization of available resources and delivering intelligent services tailored to diverse use cases. The architectural means and protocols devised for AI in 6G aspire to integrate AI technologies, empowering the network to adapt, learn, and evolve, thereby catering seamlessly and efficiently to the diverse demands of the future landscape.

The studies in this section focus on communication protocols, including discovery procedures and signalling, as well as control architecture, as outlined in Section 4.1.1. Additionally, Section 4.1.2 delves into the MLOps control loop, while the Section 4.1.3 explores the AI-driven coordinator/recommender. Lastly, Section 4.1.2 addresses the repository for modelling and data dependency relationships.

4.1.1 Architectural support for cooperative learning

Cooperative learning is an ML-based cooperative intelligence technique, where network nodes collaboratively share data and/or models towards achieving a common task, by taking advantage of each other's knowledge

and experience. The learning task requires participation of multiple nodes at the same time and needs multiple successive steps. Depending on the sensitivity of the local data and models, the latter will be either shared partially or totally or owned locally.
Currently, cooperative learning is operable on the application level, and thus network architecture, protocols and procedures are not tailored to its specifics, such as its stringent requirements on privacy/security and data accuracy. Therefore, it is necessary to introduce the corresponding architectural changes and modifications of network protocols and procedures.
On-device/UE machine learning training and inference could be partially offloaded in a collaborative fashion to network or other UEs, while preserving user privacy requirements.
Moreover, the guarantees on accuracy and power consumption should be achieved by cautiously introducing and investigating the UE role and functionality in cooperative learning.

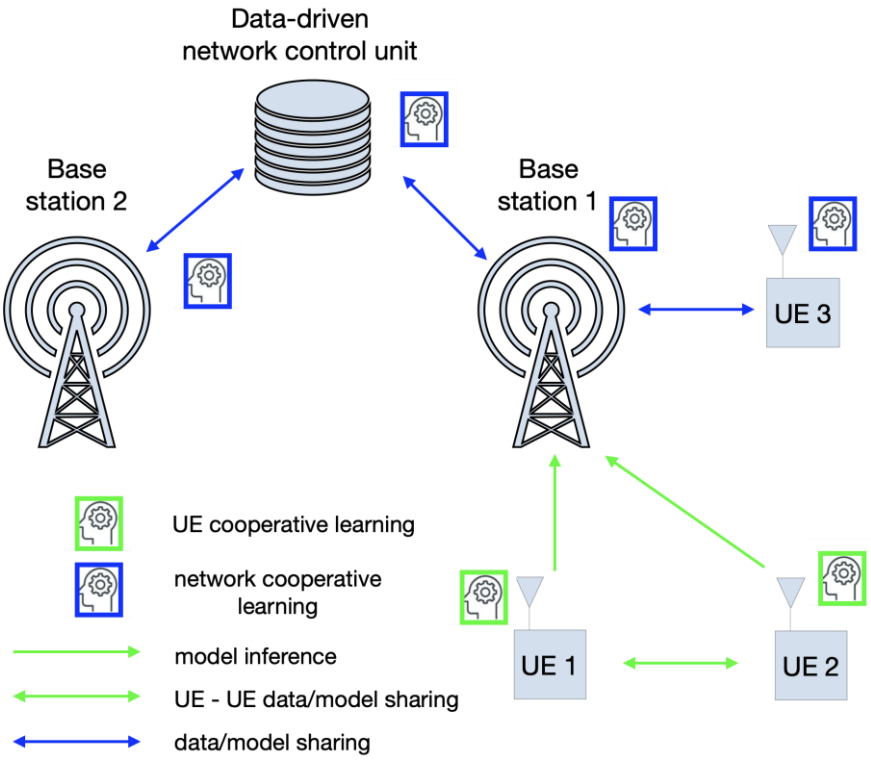

Figure 4-1 Cooperative Learning in 6G.

Based on the privacy-aware data classification and trust levels between UEs and network, data and/or model sharing can be done locally or cooperatively. The general model for cooperative learning is shown in Figure 4-1. UE1 and UE2 have low trust level with Base station 1 and share only the inference of locally trained model. Base station 1 includes this inference in cooperative learning. Moreover, UE3 has high trust level with Base station 1 and participate in cooperative learning. Base station 2 might participate in cooperative learning as well.
The discovery procedures and signalling for capability exchange among cooperative cellular nodes will be studied. Moreover, the control architecture of cooperative cellular nodes will be proposed. To meet the individual needs and relevant KPIs and KVIs, the traditional communication mechanisms should be modified and/or enhanced.

4.1.2 AI -Native Architecture

The standardized 3GPP network architecture is becoming more data driven with a focus on extending and improving the analytics provided by the NWDAF (Network Data Analytics Function) [23.288]. Particularly, a mechanism for ML model performance monitoring and improvement has been included to trigger a model update process when detecting model degradation. The ML model storage has also been enhanced to include metadata on the data that has been used to train each ML model. Hexa-X D5.2 [HEX-D52] also extends the

analytics framework with APIs for AIaaS functions and MLOps for model deployment (see Section 4.2). However, further study is required in order to integrate MLOps into the network architecture and functionalities. To automate ML model training and deployment, automated closed control loops and ML sandboxes have been proposed in the literature to train and update the models [SIM+22][ITU3172]. However, the dependencies and interplay between multiple ML models are not taken into consideration in the automated ML orchestration loops.

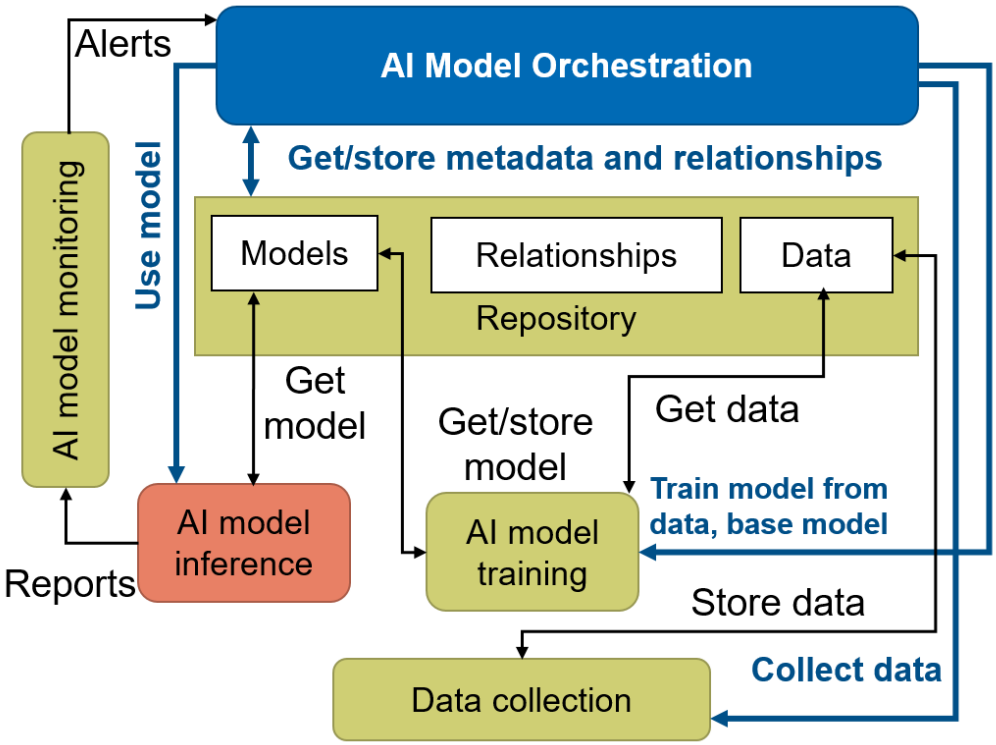

Figure 4-2 AI-Native architecture for efficient ML model orchestration.

The proposed solution will include an enhanced MLOps control loop for ML model orchestration that employs multiple AIaaS functions as illustrated in Figure 4-2. Through ML model performance monitoring, it determines model degradation from inputs of inference reports and identifies the cause and scope of the degradation such as changes in the network or service behaviour. The appropriate mitigation action is determined accordingly, such as re-training a new model with new data from new sources if necessary or optimizing the model training process by re-using existing models for Transfer Learning, or Split Learning to extend the ML model with new data while preserving previously learned knowledge. This optimization can also be used for training the models needed when deploying new services.

Furthermore, based on the cause and scope of the degradation (e.g., mobility of users), the possible impact on other inference instances and other models can be determined and proactive ML model update or replacement can be performed for the identified instances. The proposed solution will introduce a new repository function which stores dependencies and relationships between different models, and between models and data sources. Using those dependencies, the MLOps orchestration module can determine the models that have direct or indirect dependencies with the degraded model and perform the appropriate mitigation actions. If the ML model degradation has been caused by a change in the environment, it is updated (retrained, replaced by another model, or others) to reflect those changes, and the related ML models are also updated.

The use case is AI/ML model orchestration during service lifecycle. The aim is to achieve proactive and efficient ML model update to maintain performance.

4.1.3 AI-driven coordination of multiple control loops

The network or service optimisation based on Control Loops (CLs) typically uses different control-loops to optimise different KPIs. In such a case, improving one system KPI may cause a degradation of other KPIs to an unknown extent. It may also lead to a ping-pong effect or chaotic behaviour of the managed systems. This is typically caused by changing several CLs the same system parameters or impacting the ecosystem (indirect impact). The problem has already been discovered in the case of SON [32.500], and some RAN-specific mechanisms have been proposed. Still, no generic solution has been found as the complexity is very high due to the complexity of CL-managed solution. The FP7 Socrates project: Self-Optimisation and Self-Configuration in Wireless Networks (2008-2011), has tried to solve the coordination problem in SON using the RAN-specific SOCRATES Coordination Framework [SAE+11].

The problem can be partially solved by assigning to CLs different priorities or decoupling them by various time scales if their activity is semi-periodic. Another technique used is scalarisation, i.e., using a scalar value to evaluate the system state obtained by summating weighted objectives (KPIs). The weights can be elements of the management policy and modified over time.

In general, multi-objective optimisation based on multiple single-objective optimisations can be achieved by (1) cooperation or (2) coordination of multiple CLs.

The cooperation means that before the decision by any CL is taken, the decision's impact not on a single but on multiple KPIs is taken into account (evaluated). In the case of coordination, an additional entity (coordinator) decides to accept or reject the initiated CL reconfiguration (a post-factum operation). The coordination in case of rejection of the proposed reconfiguration generates a problem in the case of using CLs based on online learning, as the feedback information about the denial of the reconfiguration has to be provided to a control loop to invalidate the eventual learning step. For the coordination problem the Multi-Agent Reinforcement Learning (MARL) approach [GD22] can be used. In MARL, each agent has its rewards, and MARL allows both competition and cooperation of agents. It has to be noticed that learning in multi-agent systems suffers from the fact that both the state and the action space scale exponentially with the number of agents. Fortunately, despite this loss of theoretical guarantees, Q-learning with multiple agents often converge to optimal policies [FV07] because the agents do not necessarily need to converge to an optimal Q-value, and if all agents are playing optimally, they must settle to a Nash equilibrium, which tends to be self-reinforcing.

Figure 4-3 Illustration of the proposed impact of values of KPIs on acceptance or rejection of optimisation of reconfiguration. The values on the figure are exemplary only.

An essential factor that can be considered in multi-CL-based optimisation is the analysis of the difference between the actual values of KPIs and the target KPIs. In such a case the Kaldor-Hick's improvement criterion instead of Pareto improvement can be used [Pos07]. In practice, it means that the CL-based decisions can be accepted even if they lead to the degradation of some KPIs as long as their value is higher than the predefined target. Such an approach means, therefore, a higher percentage of accepted reconfigurations than the Pareto optimal based. To that end, having in mind the uncertainty of the reconfiguration impact on KPIs, it is proposed to avoid reconfigurations if some KPIs are only slightly higher than the threshold, to prevent their eventual degradations below the acceptable level. When one or more KPIs are below the threshold, however, a reconfiguration of any of the functions must be accepted, even if it will degrade some KPIs above the threshold. The idea is illustrated in Figure 4-3.

Yet another technique regarding the analysis of the status of the network based on KPIs is to use the Fuzzy Set Theory (FST) that can be nicely combined with the Q-learning (FQL) [Ber94].

The above-described variants of the network/service KPI evaluations, as well as Cooperative RL and MARL, will be in the future, compared via simulations. The simulation will take into account a realistic networking environment. The main purpose of the evaluation will concern the efficiency of the coordination in terms of the assumed goals (system KPIs). The experiments will also show the dependencies between the functions. Another factor that will be taken into the account will be the convergence of MARL.

4.2 MLOps

MLOps represents a set of tools designed to manage the entire machine learning development lifecycle, encompassing data preparation, model training/retraining, deployment, and monitoring, while taking privacy concerns into account. In Figure 4-4, it is shown that datasets are inherently distributed, and ML models are trained where the data is collected, primarily due to privacy considerations and the substantial volume of raw data at the edge. This necessitates the distribution of atomic AI functions across the telecommunications system and their collaboration. Within MLOps, these distributed models need to be efficiently managed while maintaining their effectiveness and minimizing overhead, with the additional goal of reusing them for similar tasks. To achieve these objectives, algorithms and technologies are employed to reduce communication, computation, storage, and energy costs during data collection, model training, and inference. The core algorithms within MLOps enable collaborative and decentralized model training and inference, incorporating mechanisms such as model transfer, model parameter exchange, and model federation.

One of the primary drivers behind the importance of MLOps in the realm of 6G lies in the demand for robust and dependable machine learning models. In the expansive 6G landscape, where billions of interconnected devices are expected to operate, machine learning models must adapt to dynamic network conditions, manage large-scale data processing, and deliver real-time accurate results. MLOps addresses these challenges by providing automated procedures for model training, version control, and deployment. By integrating MLOps practices, organizations can ensure that their machine learning models remain up-to-date, continuously monitored, and progressively improved to meet the evolving demands of the 6G environment.

In the dynamic landscape of 6G, ML models take on a multifaceted role, contributing significantly to both end-user experiences and network performance optimization. These models are pivotal for enabling transformative end-user applications. Ideally, they collaborate seamlessly with applications, UE, and even other ML models to deliver innovative and highly personalized experiences. For instance, in autonomous vehicles, ML models are not only responsible for enhancing navigation and safety but also work in tandem with applications to provide real-time information about traffic, road conditions, and even entertainment preferences. Similarly, in augmented reality (AR) applications, ML models seamlessly integrate with UEs to enable immersive experiences by understanding user gestures, surroundings, and preferences.

Moreover, beyond their role in end-user applications, ML models are indispensable for optimizing the underlying network infrastructure. They are specifically designed to analyze network traffic, predict congestion, and allocate resources dynamically. In an ideal scenario, they collaborate synergistically with radio and network components to ensure ultra-high-speed connectivity, minimal latency, and efficient resource utilization. For instance, ML models can collaborate effectively with base stations to optimize signal strength and bandwidth allocation in real-time, ensuring uninterrupted and seamless connectivity for UEs.

Collaborative training is a key aspect in 6G, where multiple ML models, potentially distributed across the network, work together to improve their capabilities collectively. Through sharing insights and data, these models learn from each other and adapt to evolving network conditions and user demands. This collaborative approach not only enhances the overall intelligence of the network but also ensures that UEs and applications benefit from the collective knowledge and expertise of these models.

At the UE and application level, the impact of ML models is substantial. These models optimize device performance, extending battery life, enhancing processing speed, and enabling advanced features for UEs. Simultaneously, applications leverage the capabilities of ML models to deliver context-aware and personalized

experiences to users. For example, a healthcare app could effectively utilize ML models to monitor vital signs and provide real-time health recommendations, thus enhancing user well-being.

The interaction between ML models and radio and network components is seamless and dynamic in 6G. ML models adaptively manage radio frequencies, spectrum allocation, and network routing, ensuring that network adjustments align with the specific requirements and preferences of UEs and applications. This collaborative and interconnected ecosystem underscores the pivotal role of ML models in creating a symbiotic relationship between end-user applications and network optimization, ultimately providing seamless, personalized, and highly efficient experiences for users and applications while optimizing the network's overall performance.

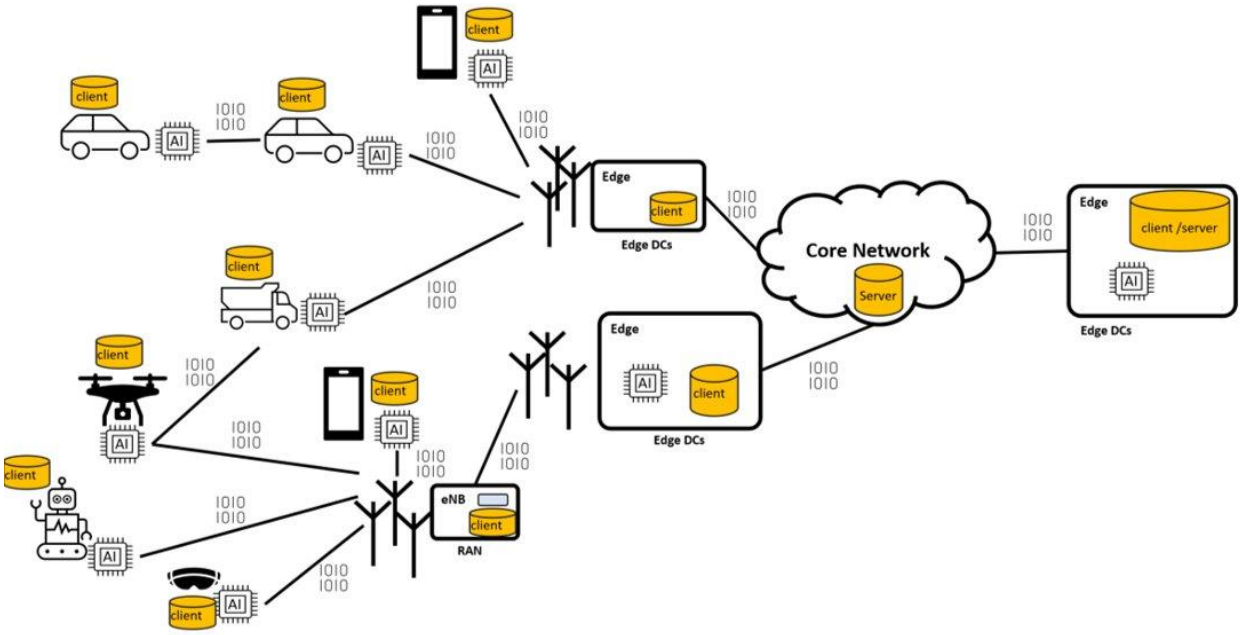

Figure 4-4 MLOps functionalities [Eri22].

Within the context of MLOps, the proposed studies focus on various aspects of model evaluation and orchestration mechanisms for efficient distributed AI model lifecycle management (LCM). The MLOps enabler serves as the foundation for all proposed studies. The studies address topics such as transfer learning and split learning in Section 4.2.1, federated learning and hierarchical federated learning (Sections 4.2.3, 4.2.4 and 4.2.5), and distributed AI for coordinating functions (Section 4.2.1). Additionally, there is a focus on distributed model and feature selection, privacy-preserving collection and learning methods (Sections 4.2.2), as well as distributed, and federated intelligence in the 6G network slicing framework to address data privacy, security, and efficiency concerns (Section 4.2.6).

4.2.1 Distributed Model Training and Inference

Next-generation mobile networks will consist of a massive number of decentralized and intelligent network devices and elements [ILR+22]. Towards the goal of achieving fully automated zero-touch networks, new methods for training machine learning models need to be developed to accommodate these complex and diverse ecosystems. For example, these decentralized network elements may be producing substantial observation data to be used in estimation, prediction of faults and/or poor performance in advance and assist actuation of appropriate actions pro-actively. The existence of high volume and potentially privacy- and/or business-sensitive decentralized datasets motivate moving from centralized learning methodologies to distributed learning ones.

High communication and computation overhead as well as sensitive information leakage may occur when moving large amounts of potentially private and business sensitive datasets from where they are collected to a central location for training in conventional centralized setting. Distributed learning techniques help overcome this challenge of centralized data collection, however DI (distributed intelligence) itself has a set of challenges:

- the decentralized datasets being heterogeneous with respect to attributes and distributions may lead to misleading correlations, slower training, and higher communication overhead;
- in the case of supervised learning, the data labels may not be accessible from one decentralized node to another (or may not be available due to deactivated measurement point);
- training distributed models jointly may introduce significant signalling overhead and network footprint as there may be many iterations;
- a decentralized node where training is supposed to take place might not host this easily due to limitations in the hardware hence might necessitate offloading some of the layers of the neural network to other available nodes;
- the models trained for multiple tasks might be similar (e.g., using similar input datasets), but are trained and stored separately consuming extra computation and storage resources as well as yielding long training time;
- there may exist data-drift between training and test sets in timeseries datasets, hence a model trained during training may under-perform after deployment.

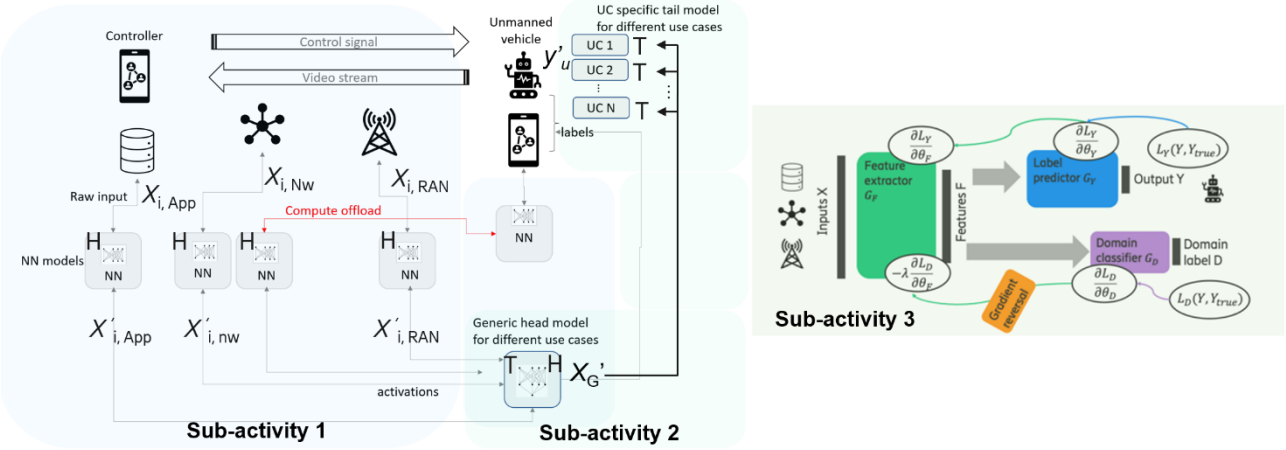

Figure 4-5 Split learning as an enabler for cross-layer ML model training and generalization to multiple tasks and domains.

The challenges listed above are studied within three sub-activities as follows. Sub-activity 1 enables cross-layer training, Sub-activity 2 and 3 enables generalization to multiple tasks (use cases) and domains.

Sub-activity 1) Multi-modal (cross-layer) training and inference involves training and inference of a split neural network, where the input model layers and parameters are split over multiple different physical nodes. The goal is to estimate QoS using multiple cross-layer input observations. Suppose that there are clients i: {1..N}, and X_i is a data source (e.g., data received from MAC, RRC, network, system load, user), and there are N different data sources of different attributes such as SINR, received power, uplink and downlink throughput observed at the core network. Y is the observed target output QoS variable such as throughput or latency observed at the application layer. The goal is to obtain a joint model with H and T, where H is head and T is a tail ML node, where each H transforms $H_i(X_i)$: X_i' for all input i, then T concatenates and encodes all X' to a common representation X_G', such that $T(X'_1, X'_2, \dots, X'_I)$: X_G'.

Sub-activity 2) Multi-task training and inference is beneficial for the cases when the dataset attributes are the same but the final target variables (i.e., tasks) are different. Suppose u represent use case (UC) identifiers, X_G': common input (e.g., all input representations), y: target use case variables (e.g., throughput, latency). X_G' is common to all use cases, y is different in all use cases. The goal is to split the model to generic and customized partitions, such that $H(X_G')$: y'_u, minimize jointly the estimation error, i.e., $y_u - y_u'$, for all u.

Sub-activity 3) Multi-domain training and inference: Unsupervised domain adaptation: Many methods for domain adaptation are in fact split neural networks. These methods can be used where data collection is not possible in target environment when there is lack of training data due to reasons including limited data collection capabilities, or low storage. For example, we have an QoE model for predicting a QoS variable with data as described in sub activity 1. This model is trained and deployed, but during time the usage pattern from users changes over time, and therefore the model does not perform well anymore. In this case it would

be beneficial to update the model without collecting new QoS labels from users. This can be done by utilizing methods for Unsupervised Domain Adaptation, where the goal is to make sure that the feature extractor maps the input to a subspace where the representations of the labelled source samples and unlabelled target samples are indistinguishable. This is done by adding a domain classifier and a gradient reversal layer in order to learn a domain invariant mapping in an adversarial manner.

In Figure 45, an experiment setting is presented for remote controlled and collaborating robots. A set of robots are capturing video frames in real time via a camera attached to them and sending them over a communication carrier link provided via a gNB to a cloud server (uplink). These robots then receive the control signals back from the cloud server for accomplishing different operations (downlink). With an internal simulation tool, the dataset is simultaneously collected from various measurement points from different communication network layers (from radio link to the application layer).

A cross-layer joint ML training is performed for two tasks of estimating playout bitrate and delay observed at the application layer. In addition, this is achieved by generalization techniques such that the portion of the model trained for delay can also be substantially reused and personalized for other tasks and use cases simultaneously. The use case is performed via multiple different simulation configurations and different data availability scenarios, and domain adaptation technique is used to enable model transfer between a source domain and a target domain. The role of split neural networks with respect to privacy aspects was previously studied in the area of health care [PVC+19] and also in estimation of QoE (Quality of Experience) [IFV21] in telecommunication domains. Moreover, its role in sustainable distributed model training and inference are as follows. i) *cross-layer learning*: training a large neural network where the split portions of it are located in decentralized fashion in multiple entities, i.e., multiple input nodes (layers, network functions) can be enabled by split neural network training in a multi-head topology. The input attributes from decentralized entities do not necessarily have to be the same, in fact can be complementary, hence enabling multi-modality as well; ii) *multi-task learning*: split neural networks with a multi-tail topology helps to train a joint global model that is generalized to multiple tasks, i.e., multiple output nodes. Generalization helps to reduce storage and memory requirements in the case when similar tasks are being trained using similar (if not the same) ML input features; iii) multi-domain learning: domain adaptation techniques such as Domain Adversarial Neural Networks [GUA+16] are known to have split neural network architectures, where these techniques help to transfer a model that is pre-trained on a good quality dataset (e.g., without missing values) into domains where the same quality is not present (e.g., with missing labels). This helps with training models even in environments where there are no labels, by utilizing a labelled source data set in addition to the un-labelled target data.

PoC #B.2: The above described distributed training and inference is implemented in split neural network setting as a proof of concept (PoC). The implementation is performed on a Kubernetes cluster where the pods are emulated as isolated but inter-connected network elements with compute and storage capabilities in the same namespace. These pods are communicating and exchanging model parameters and signalling between each other during training and inference. More details on the PoC and corresponding results will be presented in next deliverable D3.3.

4.2.2 Privacy-aware data collection and learning

The emergence of heterogeneous networks and distributed data applications in 6G requires distributed data-driven decision-making approaches, such as distributed ML and federated learning. Enabling distributed ML among cellular nodes requires data and model exchange between the cellular network and UEs. This implies new requirements related to data collection, training, and inference such as privacy preservation and coordination of data and learning among cellular nodes.

Moreover, privacy preservation imposes a critical challenge when sharing privacy-sensitive UE data with the network and other UEs. Depending on privacy sensitivity levels, UE data can be shared directly, aggregated, or not shared at all. Moreover, the coordination mechanism is required to keep the models and data up to date and synchronized among the cellular nodes, which affects the network architecture and protocols.

On-device/UE machine learning training and inference could be partially offloaded in a distributed fashion to network or other UEs, while preserving user privacy requirements.

UEs can take advantage of network information along with on-device contextual information (user activity, intent, and usage patterns) to assist the network in connectivity decisions, i.e., to improve the connectivity QoE.

In future distributed data-driven applications and network deployments, strict requirements, primarily on user privacy, accuracy, and power consumption will be imposed. The guarantees on these KPIs and Key Values Indicators (KVIs) could be achieved by carefully investigating the role of UE in functional framework to facilitate distributed ML.

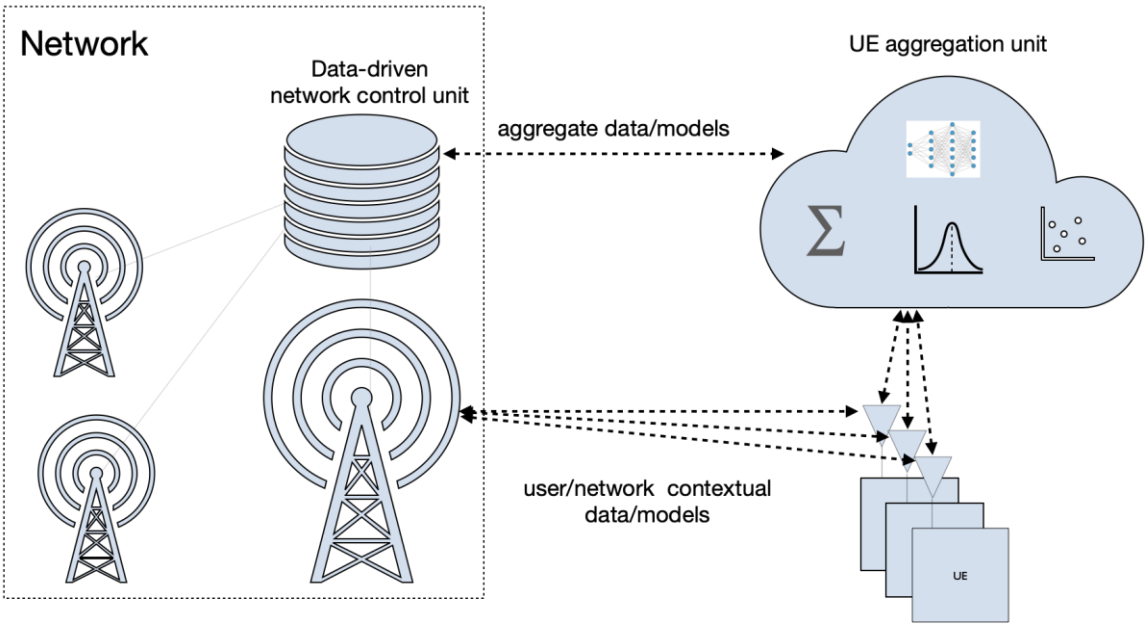

Figure 4-6: Privacy-preserving architecture for data collection, learning and analytics.

The privacy preserving-architecture for data collection, learning and analytics is illustrated in Figure 4-6. This study will be focused on the privacy aspects of the data and privacy preserving sharing mechanisms. Moreover, architectural and signaling adaptations of the cellular network will be investigated.
The main architectural component is UE aggregation unit, which performs privacy-preserving data aggregation. It uses secure aggregation techniques and/or data anonymization to collect user data and should not contain any confidential information about a specific user. Another architectural component would be data-driven network control unit, an entity at the network side that configures the base station by performing automation, optimization, and intelligence services, based on both network and UE data. UE-collected data (contextual data, measurements) is shared by the UE aggregator with the Data-driven network control unit and can be used by network to improve its control and decisions. The exact implementation of the aggregation mechanism and the servers split/deployment can be agreed between UE and network vendors to enable the privacy preserving coordination. For example, privacy-preserving cryptographic protocols like Prio [CB17], currently being standardised in IETF Privacy Preserving Measurement Workgroup [GPR+23] can be used.
Depending on the level of privacy of data, different training mechanisms can be applied. In case of privacy-sensitive UE data, a model is trained in the UE and shared with the network using private federated learning. Moreover, differential privacy, combined with federated learning, enables privacy-preserving learning on user data. Inference is done at the UEs given the network-shared (complete or partial) model. Furthermore, to keep the data and models up to date, the learning coordination is needed. It assumes a new signaling for data sharing, training and inference and life-cycle management of data and models which will be investigated in the study (e.g., how frequent are data changes and how often to update the model).

4.2.3 Wireless hierarchical federated learning: On the accuracy-energy trade-off

Federated learning (FL) is a collaborative ML model training method where the ML model is produced in a distributed manner by several end users. The end users locally train models using their data, and then model aggregation is performed by a central entity, usually located in the cloud. Nevertheless, direct communication with the cloud may cause increased backhaul network traffic while increasing the users' consumed energy when considering FL implementations over wireless networks.

To overcome these issues, Hierarchical Federated Learning (HFL) suggests adding an extra layer of intermediate model aggregation where several edge servers facilitate the aggregation and transmission of end

users' model parameters to the cloud [LZS+20]. The users are associated with different edge servers, and more efficient user-edge updates can be performed, resulting in reduced network overhead and consumed energy. An illustration of the wireless HFL network architecture is provided in Figure 4-7.

Therefore, in the context of wireless HFL networks, the problem of user-to-edge-server association and wireless resource allocation emerges to control the achieved local model accuracy and incurred energy consumption of the users. Different objectives can lead to contradicting outcomes for the model's accuracy and consumed energy. A user-to-edge-server association that balances users' data across edge servers to force an Independent and Identically Distributed (IID) data case may increase communication's energy consumption [LYC+22]. Inversely, an association mechanism purely based on favourable wireless communication conditions may yield poor model accuracy [LCW+20].

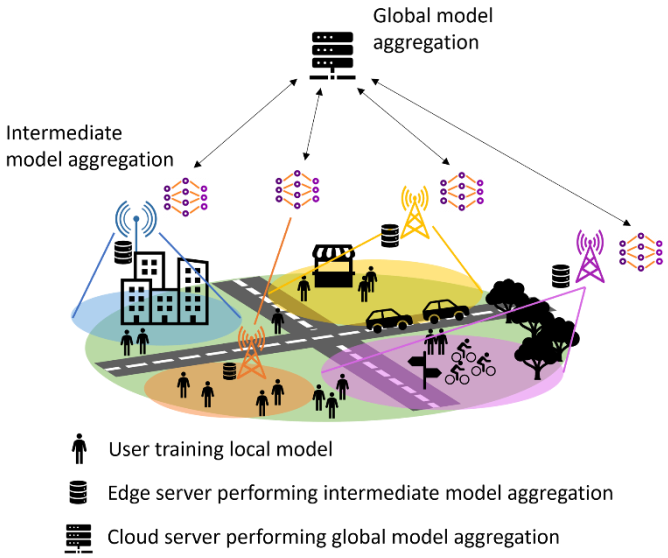

Figure 4-7 Overview of wireless HFL network architecture.

This study contributes to the energy-efficient provisioning of digital services that require training heavy ML models for the following two reasons. On the one hand, a wireless HFL network architecture is proposed to reduce the burden of large data centres by the introduction of end-user devices in the ML model training process, while their energy efficiency in terms of local computation and model parameters' transmission is pursued. The KPIs considered are the ML model's accuracy and the end users' energy efficiency, considering the energy consumed due to both local model training and wireless transmission of model parameters to the edge. In this way, crucial KVIs, e.g., sustainability and trustworthiness, are considered.

Specifically, the goal is to design an appropriate framework for the joint optimization of the user-to-edge-server association and the users' uplink transmission power for transmitting their model parameters to the edge. The framework will allow each user to self-configure via selecting its edge association and transmission power level, striking, in this way, its personally pursued local model accuracy and energy consumption balance. The interactions between the users' interdependent actions and decisions will be modelled as a non-cooperative game between them [LT11], and different game-theoretic equilibria will be studied that yield different accuracy-energy trade-offs [PTL+12].

4.2.4 Incentive mechanism design for wireless federated learning networks

FL provides an effective way to train global ML models by utilizing the large volume of data generated from diverse IoT end-user devices. Instead of centrally collecting the data and falling into security and privacy issues [LWW+23], in FL, the end-user devices locally train the corresponding ML model using their private data and exchange the resulting model parameters with a central entity to produce the global ML model. Therefore, FL relies heavily on the quality of the local model updates performed by the end users.

In this context, motivating end users to participate in the FL process and contribute accurate model updates by investing their resources is a challenging problem that needs to be addressed. Indeed, end users may be reluctant to invest their computing and communication resources to perform local processing and model parameter transmission when there is a multitude of other computation and communication tasks to be executed on their devices, and in general, they may be reluctant to consume their battery when talking about battery-powered devices [ZZH+22].

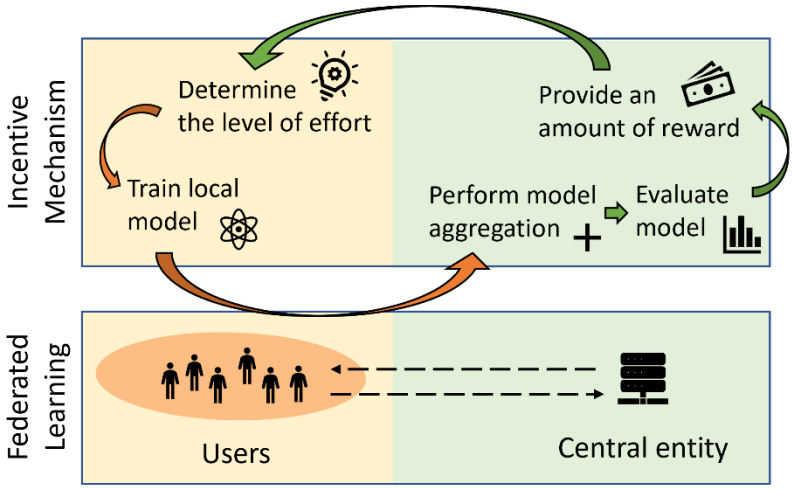

Figure 4-8 Overview of incentive mechanism design for FL

This study contributes to network sustainability in the sense that end-user devices of restricted communication, computing, and battery resources make it possible to participate in a federated learning process that they could not do otherwise. By indirectly contributing their data for ML model training, they allow unlocking new digital services or application features, which they can subsequently enjoy. The KPIs considered are the ML model's accuracy and the network entities' profit.

Specifically, the goal is to design an appropriate incentive mechanism to motivate the end users to participate in the FL process and settle their costs. The incentives will comprise a monetary reward, aka payment, to the end users, based on which the most beneficial investment of resources from the users' side will be determined. The payment may be revised after evaluating the local models produced by the users when employing the previously agreed amount of resources. Therefore, this results in an iterative process that can take place along with the FL process that is performed on a time-slot basis anyway. An overview of this iterative process is presented in Figure 4-8. The outcome of the incentive mechanism will be a market equilibrium point [NLB21] where the payment will be such that both the end users' costs are adequately settled, and the global FL model's accuracy reaches the required level.

4.2.5 Federated learning approach between different city verticals

It is generally believed that 6G will be established on ubiquitous Artificial Intelligence (AI) to achieve data-driven Machine Learning (ML) solutions in heterogeneous and massive-scale networks. However, traditional ML techniques require centralised data collection and processing by a central server. It is becoming a bottleneck due to the large amount of data ingested and processed by a single (logical) component.

Federated learning (FL) synergises very well with edge computing, allowing distributed client nodes (edge nodes) to contribute to the overall training of the algorithms by sending their learning and not the data used to train the models [LFT+20]. Federated Learning may also be one of the supported AI-as-a-Service (AIaaS) in the 6G architecture, through AI functions such as AI repository, training, monitoring, and AI agent that enable having AI as close as possible to the application and cross-domain AI service consumers and data producers [5GP22].

In order to realise the vision for a truly Smart City and drawing from our experience in the Smart Cities field, we built an integrated system to manage a city, however, the different verticals need to communicate and share information in a federated way. 6G has the potential to be an AI catalyst in several city verticals highlighted in Figure 4-9, such as tourism, waste management, energy efficiency, parking and mobility, where the AI

nativeness of 6G constitutes strong value propositions in deploying innovative applications. Looking particularly into mobility within cities as an example, providing the everyday citizen with updated information that can help in key decisions such as the best time to leave home or work, using a which transport method (car, public transport, bicycle, etc..), carbon footprint of the chosen path, in what can be referred to as Mobility-as-a-Service (MaaS). A simple route from A to B can have several parameters such as time, price, carbon footprint, burned calories. Combining data from weather forecast, cultural events, hotel occupation, the time of the year, construction work, live traffic and biometrics is something a human would struggle but AI excels at. Combining data from several sources and uses machine learning algorithms help citizens to decide more consciously their transport method and the respective impact in themselves, other persons, and the planet.

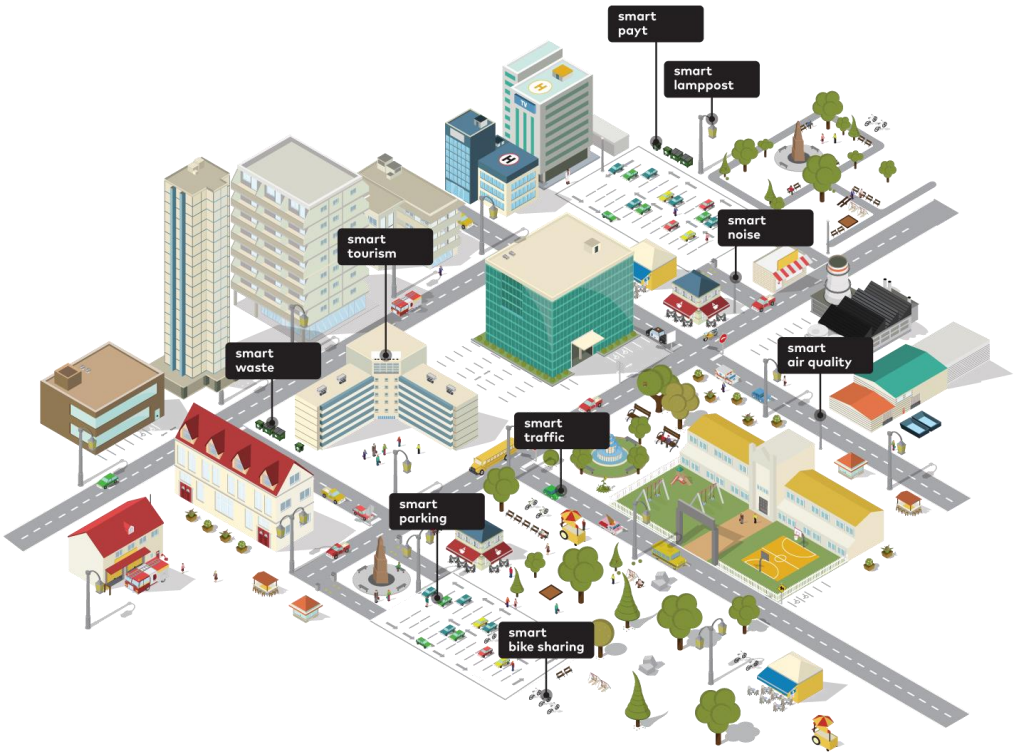

Figure 4-9: City as an integrated system.

The high availability, reliability, and coverage of services over 5G and 6G is key for municipalities to integrate all the necessary systems and collect data from CCTV cameras through computer vision, sensors and user applications to monitor and manage the city as a whole. Going back to the smart mobility example, municipalities can use historical data from multiple sources to plan mitigation measures to alleviate congestion during peak times, plan construction work to decrease its impact in overall city mobility or promote mobility alternatives that reduces carbon emissions.

The goal of the study is to leverage 6G AI driven architecture and some of its functions to use federated learning techniques across several edge nodes, in a city environment. The collected data will come from different sources such as sensors that measure humidity, noise, air quality, EV chargers, live video streams, parking sensors, etc. In the study it is intended to also explore the cloud-edge continuum depending on application requirements such as privacy or latency, where typically live video streams should be processed locally and only metadata and the results of the local training of algorithms, running in edge nodes, are sent to a central cloud monitoring platform.

4.2.6 E2E 6G Network Slice Instance Employing Distrusted Intelligence Solutions

There is a strong consensus within the research community that with the increasing demand of industrial applications and use cases, the number and types of standardized and operator-specific network slice (NS) instances will also be increased. The early research and development on 6G have demonstrated that the

futuristic network slicing framework is expected be equipped with novel automation and intelligence capabilities to fulfil the management and orchestration, among several other aspects, of a large number of heterogenous 6G NS instances. To enable the 6G slicing framework with AI/ML capabilities, there are two scenarios: (a) the data is collected in a centralized location, the model is trained, and the recommendations/predictions are generated; (b) the data is collected at domain-level or network – function (NF)-level, the model is also trained in such distributed locations, and subsequently applied to the network domain and NFs. Each scenario has its own advantages and limitations. Considering the privacy and security of data and end users, we anticipate that the second scenario, distributed intelligence, will be a novel enabling intelligence solution to 6G network slicing framework.

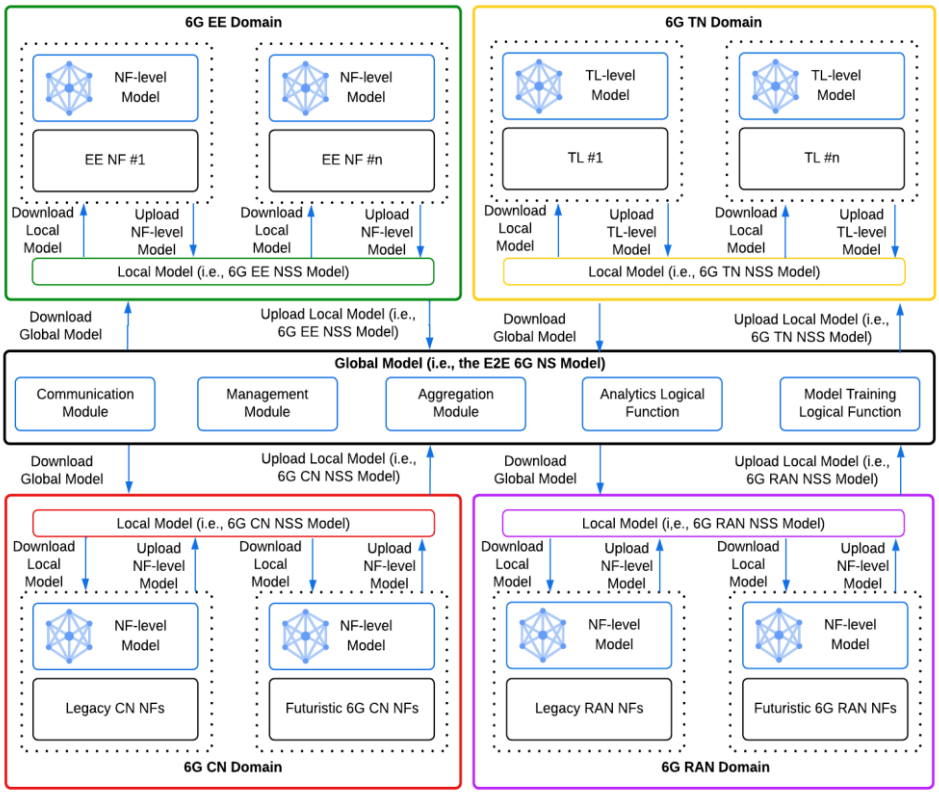

Figure 4-10 The proposed distributed intelligence-assisted architectural solution for 6G slicing.

There exist four network slice subnets (NSSs) in a 6G slicing framework [HEX-D62]: the core network (CN) NSS, the radio access network (RAN) NSS, the transport network (TN) NSS, and the extreme-edge (EE) NSS. Each domain may consist of legacy and futuristic NFs, except the TN domain which consists of transport links (TLs). These domains produce a large amount of data. The collection of this data in a centralized location is challenging from various aspects, including the security and privacy of collected data. Figure 4-10 shows that the 6G slicing framework also consist of a centralized location, where the main decisions with respect to a 6G NS instance are made. To help the centralized management entity federate the decisions and let only global model be created in this location, we propose the concept of distributed and federated intelligence in the 6G slicing framework. We believe that distributed intelligence can be one of the most efficient solutions that on the one hand improves privacy and security, and on the other hand increase the efficiency of deploying various 6G NS instances.

Within the context of "**From Robots to Cobots**" use case family, there can be mainly five use cases that can be studied within the context of the proposed distributed intelligence framework. The four use cases can be related to the building local models for each of the domain, and the fifth use case can be related to designing a global and distributed mode for an end-to-end 6G NS instance. Data Privacy, User Secrecy, Data Security, and NS Security and Privacy are the main KPIs that will be studied within this architectural solution.

We propose a preliminary architectural solution for the distribution of intelligence across the 6G slicing framework in Figure 4-10. The proposed architecture will consist of four domains, as we described them in

the above. Each domain will have its customized local model, called the NSS model, which can be obtained through collecting NF-and TL-level models in the respective network domain. Additionally, the proposed framework will also consist of global model, which is obtained through collecting NSS models. In this solution, only the trained models will be shared among the domains, avoiding the data sharing due to security and privacy concerns. The global models can be downloaded by the domain-level intelligence entity to allow the local model to provide more efficient, collaborative, and real-time intelligent solutions to its underlying NSS instance.

4.3 AIaaS

The importance of Artificial Intelligence as a Service (AIaaS) in the context of 6G networks arises from the significant potential of AI and the transformative capabilities offered by 6G. AI has the power to revolutionize numerous industries by enabling intelligent decision-making, automation, and enhanced user experiences. As 6G promises ultra-high-speed connectivity, minimal latency, and extensive device connectivity, the demand for AI applications is projected to soar. However, harnessing the power of AI in the 6G landscape necessitates efficient and scalable infrastructure, which is where AIaaS assumes a crucial role.

AIaaS is a comprehensive framework that offers a wide range of AI functionalities and tailored inference capabilities for applications, services, and AI-driven management and orchestration decision logics. Alongside pure analytics, prediction, and classification capabilities, it provides diverse AI functionalities and services to facilitate closed-loop network and service automation. Figure 4-11 illustrates the framework four primary functions. The AI model repository function serves as a catalogue of AI-trained models that are either deployed or ready for deployment in AI agent instances. The AI training function is responsible for training AI algorithms and generating executable models. The AI monitoring function evaluates the performance of AI models and triggers training and retraining operations in the AI training function based on the results. The AI agent executes models and delivers inference capabilities using the available trained models, ensuring compliance with necessary data pre-processing requirements. The AIaaS framework exposes dedicated APIs and interfaces for managing and controlling the various AI functions. These APIs and interfaces facilitate deployment in cloud-native virtualized infrastructures, initial and runtime configurations, lifecycle management, and more.

Within the context of 6G, AIaaS becomes increasingly vital for several reasons. Firstly, 6G networks are expected to generate an unprecedented volume of data from diverse sources such as IoT devices, sensors, and edge computing nodes. AIaaS offers the computational power and scalability required to process and analyse this massive influx of data, extracting valuable insights and facilitating real-time intelligent decision-making. Secondly, the dynamic and intricate nature of 6G networks demands AI algorithms that can continuously adapt and optimize network performance. AIaaS provides a flexible and scalable platform for the development, deployment, and management of such AI algorithms. This empowers network operators and service providers to effectively leverage AI capabilities as a service, thereby unlocking the potential of intelligent network management in the 6G landscape. Furthermore, AIaaS addresses the challenges associated with AI model training and deployment within the 6G environment. Training sophisticated AI models necessitates substantial computational resources and expertise. By utilizing AIaaS, organizations can offload the computationally intensive tasks to the cloud, significantly reducing the overhead and time required for model training. Additionally, AIaaS streamlines the deployment and scalability of AI models across the 6G network, facilitating rapid and efficient integration of AI capabilities into diverse applications and services.

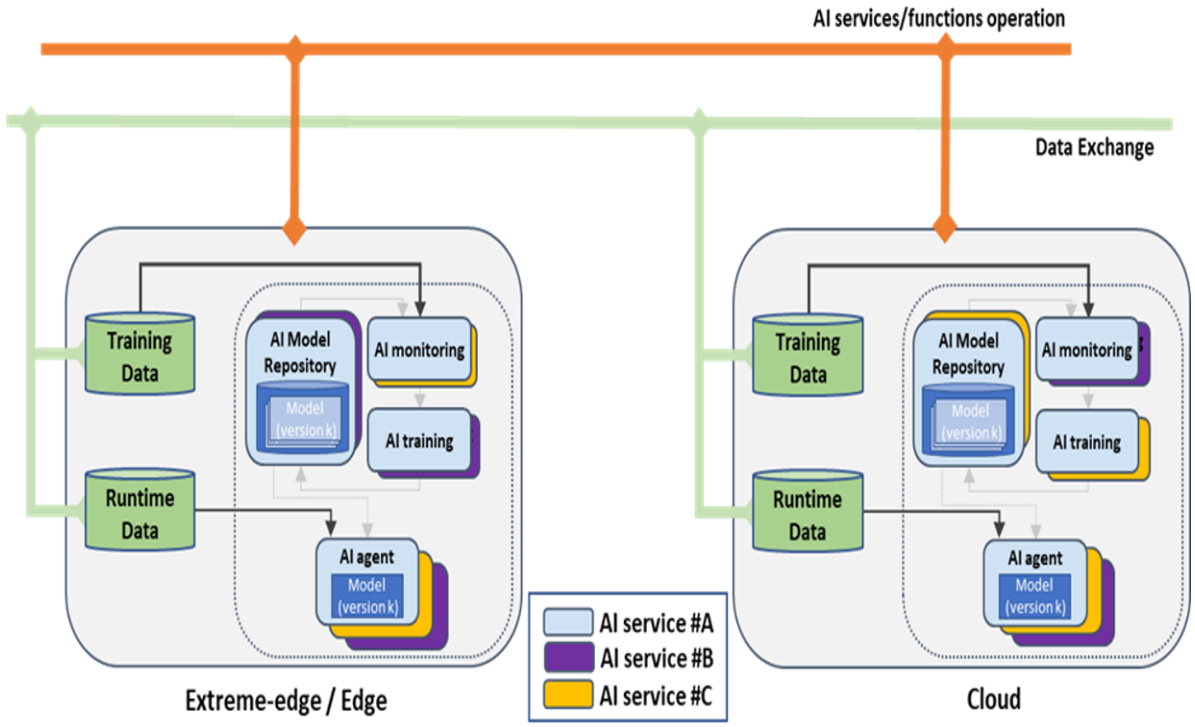

Figure 4-11 AIaaS functionalities.

Within the context of AIaaS, the studies focus on data and ML model management procedures in distributed and cooperative scenarios. Specifically, they address the requirements for data and ML model transfer as outlined in Sections 4.3.1, 4.3.2 and 4.3.3 These sections explore the trade-offs involved in the extreme edge/edge/cloud continuum, considering data exchange considerations, and determining the optimal timing and location for training and inference processes to minimize the amount and type of data exchanged across the network. Additionally, the studies delve into AIaaS exposure APIs for AI services and functions operations, which are discussed in Sections 4.3.2 and 4.3.3.

4.3.1 Distributed AI Services

AIaaS and Compute-as-a-Service (CaaS) have been identified as crucial architectural enablers for dynamically deploying distributed AI services, as emphasized in [HEX-D52] and [HEX-D53]. The study investigated the optimal placement of these frameworks within the anticipated 6G architecture to facilitate AI service allocation, instantiation, and operation, resulting in a proposed architecture. This study focuses on exploring efficient methods for monitoring, anomaly detection, and assisted troubleshooting of distributed intelligence. To verify these methods in a specific use case, an architecture comprising AIaaS and CaaS components is employed.

The evolution from traditional industrial robots to cobots and AI-enabled robots with high or full autonomy characterizes the shift from robots to cobots. Cobot systems possess the ability to independently sense, perceive, plan, and control towards a shared objective without explicit human instructions. These cobots are equipped with video cameras that stream data to a local compute server for real-time processing. Additionally, they leverage advanced sensing and positioning capabilities and harness the connected AI capabilities offered by 6G to facilitate situation-aware cooperation, collaboration, and assistance. On the human side, interaction with machinery or mobile robots can be either direct or indirect. Through task interactions, robots can learn from humans, leading to optimization of execution steps and improvement in error mitigation and prevention. This collaboration enables machinery to perform highly customized on-demand tasks, facilitating lot size one production and maximizing the utilization of innovative manufacturing techniques such as additive manufacturing.

In the context of Distributed AI Services, the use case of transitioning from robots to cobots highlights several KPIs that are crucial for successful implementation. Firstly, in terms of communication, achieving high

reliability, availability, low latency, and a high data rate is essential. This ensures seamless and efficient data transfer between the cobots and the connected network, enabling real-time collaboration and decision-making. Secondly, the KPIs related to AI and computation play a vital role in the effectiveness of cobots. High agent availability ensures that the cobots are consistently ready to perform their tasks, while high agent reliability guarantees their dependable performance. Moreover, achieving high inferencing accuracy in AI computations enables precise decision-making and execution of tasks, enhancing overall system performance. Lastly, in the area of localization and sensing, high service availability ensures that the required services for localization and sensing are readily accessible whenever needed. High service reliability guarantees the consistent and accurate functioning of these services, enabling cobots to accurately perceive and interpret their surroundings. Additionally, achieving high location accuracy is crucial for precise positioning and navigation, allowing cobots to operate efficiently and safely within their environments.

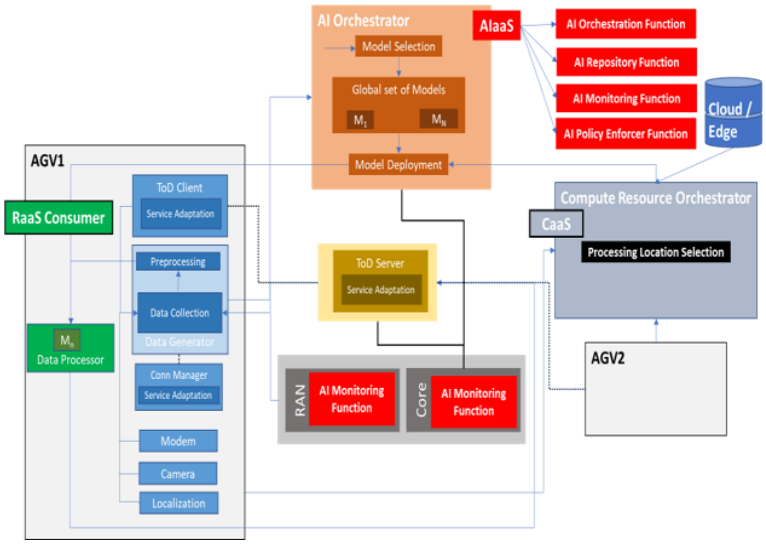

Figure 4-12 Architecture for realizing Predictive Quality of Service (pQoS) in Interacting and Collaborative Robots Use Case

The objective of this study is to calculate the predictive QoS to determine the motion pattern of a cobot using a proactive approach in robotics scenarios. During the cobot's movement towards a specific target with defined speed and direction, there may be instances where a decrease in the QoS value is observed at time $t+\Delta t$, indicating potential issues such as shadowing. To ensure worker safety, avoid accidents, and successfully reach the target, the cobot utilizes ML models and leverages data collected from other cobots, humans, and the network. Based on this information, the cobot can make informed decisions such as halting or delaying operations, reassigning tasks, or choosing alternative movement trajectories to achieve the desired pQoS value at time $t+\Delta t$. This proactive approach should be adopted by other cobots and system users to maintain a seamless and safe operation of the cobot system.

The system demonstrated in Figure 4-12 comprises several essential components including data collection mechanisms, modems, cameras, a localization module, a ToD client and server, and a data generator. These components work in conjunction with the network infrastructure to facilitate the exchange of analytics-related information. To ensure efficient monitoring, anomaly detection, and assisted troubleshooting of distributed intelligence, it is imperative to examine the interactions between the AIaaS functional entities, CaaS, cobots, and the network. Through a comprehensive analysis of these interactions, the utilization of information derived from the network can be optimized for enhanced system performance and operational effectiveness.

4.3.2 AIaaS Operation

With the advent of 6G and its distributed intelligence capabilities, the role of AI is set to expand significantly. Hexa-X-II has identified key drivers for AI in 6G which are new opportunities leveraging the 6G infrastructure flexibility, coping with network and service management complexity, and supporting new revenue streams via novel services with benefits both for society and industry. To support these drivers, Hexa-X-II aims to establish a data-driven architecture framework for resource and service management, as well as vertical application mechanisms. This framework will define the necessary functions and interfaces to facilitate efficient data

exchange. Within this context, Hexa-X-II will design a data-driven architecture that enables verticals, such as AIaaS, to access and utilize the capabilities of 6G. The study involves developing enablers and architectural concepts, supporting functional entities of AIaaS operation, and identifying the required APIs for AIaaS customers to seamlessly deploy their AI services.

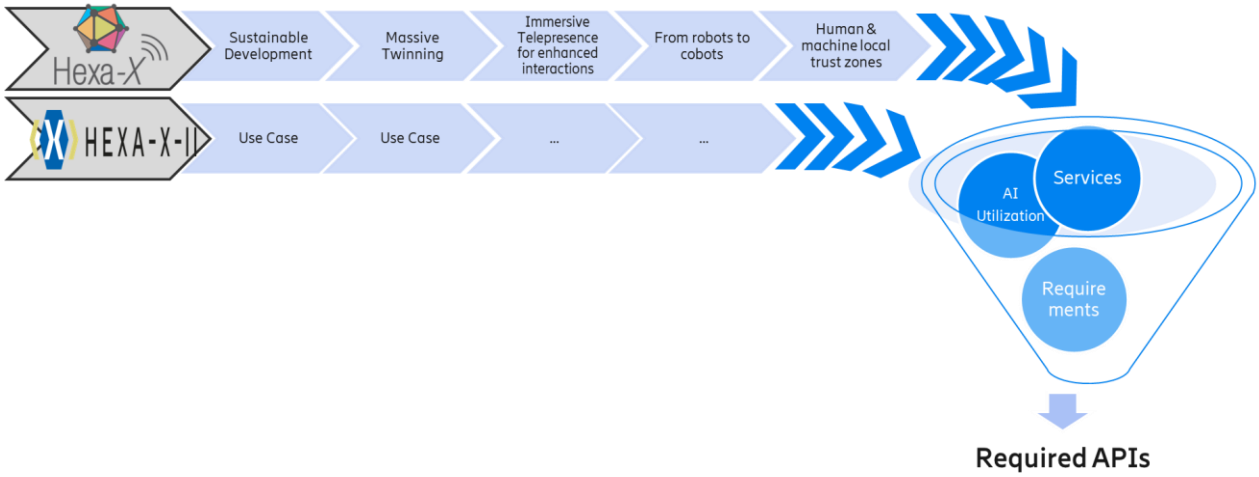

Figure 4-13 A Use case-based approach to define AIaaS APIs

To gain a comprehensive understanding of AIaaS APIs and their functionalities, a use case-based approach is being employed as shown in Figure 4-13. This involves creating and investigating a diverse range of use cases. The investigation focuses on use case families derived from [HEX-D12], [HEX-D13], [HEX-D14], [HEX-D71] and [HEX2-D11].

The investigation process involves analysing the services required by each use case and determining how AI can enhance their execution. This includes considering network-wide reconfiguration to meet QoS specifications. By thoroughly understanding each use case, the subsequent step involves exposing the necessary input and output through APIs to obtain the desired services. In essence, a systematic approach is followed, breaking down the high-level use case to identify the required data, services, traffic, network functions, and how AI can contribute. This process helps in determining the specific APIs needed to facilitate the exchange of inputs and outputs. To begin with, the approach will be elaborated on the robots to cobots use case family, and then the same approach will be applied to all use case families specified in the deliverables [HEX-D12], [HEX-D13], [HEX-D14], [HEX-D71], and [HEX2-D11]. In this deliverable, the Communication-related and AI-related requirements of robots-to-cobots use case family services have been stated.

In the context of AIaaS, the transition from traditional robots to cobots brings forth significant use cases with several KPIs to consider. Firstly, in terms of communication, the KPIs of high reliability, high availability, low latency, and high data rate are paramount. It is crucial for cobots to have reliable and available communication channels, ensuring seamless and uninterrupted data exchange. Low latency facilitates real-time responsiveness, while a high data rate enables the efficient transfer of large volumes of data. Secondly, in the realm of AI and computation, high agent availability, high agent reliability, and high inferencing accuracy are essential. Cobots should have consistent access to AI capabilities, ensuring their continuous operation. Reliable agents and accurate inferencing enhance the reliability and effectiveness of AI-powered functionalities. Lastly, in terms of localization and sensing, high service availability, high service reliability, and high location accuracy are critical. Cobots rely on precise localization and sensing capabilities to perform their tasks effectively and safely. Ensuring high availability and reliability of these services, along with accurate location information, enhances the overall performance and productivity of cobots within the AIaaS framework.

4.3.3 Strategies and mechanisms for distributed AI and AIaaS functions management

The realization of a data-driven architecture with native and in-network AI capabilities is critical for enabling full automation in how 6G networks will be managed and operated to satisfy the challenging requirements of the 6G use cases in terms of pervasiveness, mobility, network performances, sustainability. Specifically, common services and functions for AI are required to facilitate a seamless integration and use of AI and ML functionalities in the 6G network management and operation frameworks.

Hexa-X has proposed an AIaaS solution which is implemented as a stand-alone framework built by the integration of few AI functions covering specific capabilities and offering a set of services [HEX-D53]. In particular, four main AI functions have been identified so far: AI training, AI repository, AI agent, AI monitoring, see Figure 4-14 [HEX-D52]. However, these Hexa-X initial studies on AIaaS did not yet cover deep analysis on few critical aspects, including deployment and operational models for the AI functions, and detailed definition of exposed APIs and services towards external consumers. Specifically, more studies and investigations are required to properly support decentralized and cooperative AI services and functions. These include:

- defining how AI functions need to cooperate to enable distributed and federated learning,
- understanding if there is a need for more types of AI functions,
- identifying how many deployment models are required to be supported within the AIaaS framework.

The relevant use case to the proposed study is Interacting and Collaborating Robots. Indeed, this use cases poses specific challenges for the implementation, deployment and operation of AI functions and services in a distributed and decentralized way, especially considering edge and extreme edge constrained resources for the execution of AI workloads in a delay-sensitive application scenario.

In terms of performance improvements, the proposed study aims at finding optimal solutions for data distribution among the involved AI functions, considering the various functionalities and their data requirements (e.g., for training, inferencing and monitoring functionalities). Similarly, the study targets an optimization of computation resources for AI, by analysing the AI placement across the compute continuum (thus considering extreme edge, edge and cloud locations). In both cases, the aim is to perform a qualitative analysis among different options (for both data distribution, i.e., what type of data to exchange and when, and AI and computation).

To align and evolve the implementation, deployment, and operation of in-network AI functions to the 6G requirements and trends, this study aims at providing support for heterogeneous cloud-native deployments of AI functionalities and workloads across the whole compute continuum (i.e., cloud, edge and extreme edge). At the same time, given that 6G networks are expected to integrate highly distributed and heterogeneous domains and technologies, it becomes essential to enable decentralized and cooperative AI techniques, through tailored sharing and cooperation models of AI functions and services. These include enabling per-domain, per-slice and per service AI services, in combination with cross-domain, cross-slice, and cross-service AI services to realize multi-domain and multi-layer intelligent 6G networks. Therefore, the study will focus on defining AIaaS deployment models in support of decentralized and cooperative AI, enabling continuous AI monitoring, validation and models re-training. Specifically, qualitative analysis of trade-offs and computational gains in deploying and operating the AI functions at the extreme edge, edge or cloud will be carried out. For this, the definition of operational workflows for AI functions interactions (casted to the specific deployment model) will be provided. In addition, data distribution and AI/ML model management mechanisms at the edge and extreme-edge will be defined to facilitate distributed and federated learning solutions.

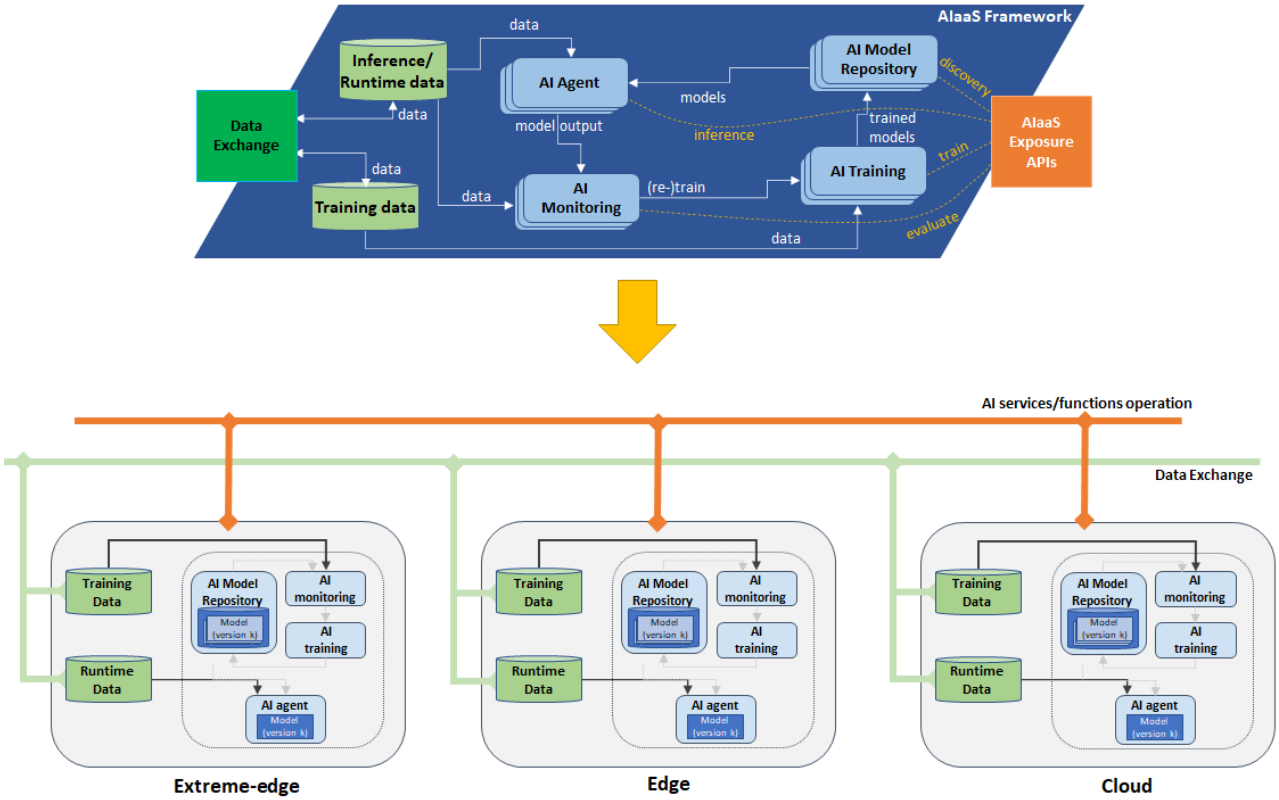

Figure 4-14 Evolution of AI functions deployment and operation to fully distributed approach.

The goal is to carry out a qualitative analysis of trade-offs and benefits in terms of data exchange optimization (i.e., where and when to train, validate and monitor the models). Moreover, to achieve a consolidated AIaaS framework design, common and unified exposure, APIs for AI services and functions are planned to be defined and details, with the aim of generalizing AI operations and management and accommodate the requirements of multiple AI solutions. In summary, the study activity will include a conceptual design of the new AIaaS capabilities (e.g., new functions, procedures, and mechanisms) to support.

4.4 DataOps

DataOps is an approach that focuses on designing, implementing, and maintaining a distributed data architecture. It encompasses various disciplines within information technology including data collection and, data transformation, data extraction, data quality, and more. Figure 4-15 illustrates a data pipeline comprising two main functional areas: data ingestion and data refinement. The harmonized solution for the data ingestion architecture, referred to as DataOps, ensures that the collected data is made available for further processing in the data refinement functionality and can be consumed by various applications. In other words, instead of locking data into a single pipeline or application, the harmonized data ingestion architecture enables authorized application suites and their associated data pipelines to access the data. This approach promotes data discovery, quality control, and effective management of the data lifecycle, fostering trustworthiness for every application that utilizes the data.

In the 6G landscape, DataOps plays a crucial role in handling the enormous volume, velocity, and variety of data while ensuring data quality. It enables real-time insights and decision-making by facilitating the efficient processing and analysis of data. By implementing DataOps practices, organizations can maximize the value of data, enhance operational efficiency, and unlock innovative applications and services in areas such as smart cities, autonomous vehicles, and personalized experiences. Furthermore, DataOps enables the seamless integration of AI and machine learning algorithms into the data pipeline. This integration empowers organizations to leverage advanced analytics, predictive modelling, and actionable insights. DataOps ensures that the data pipeline is optimized for AI and machine learning, facilitating intelligent automation, and enabling real-time, data-driven decision-making.

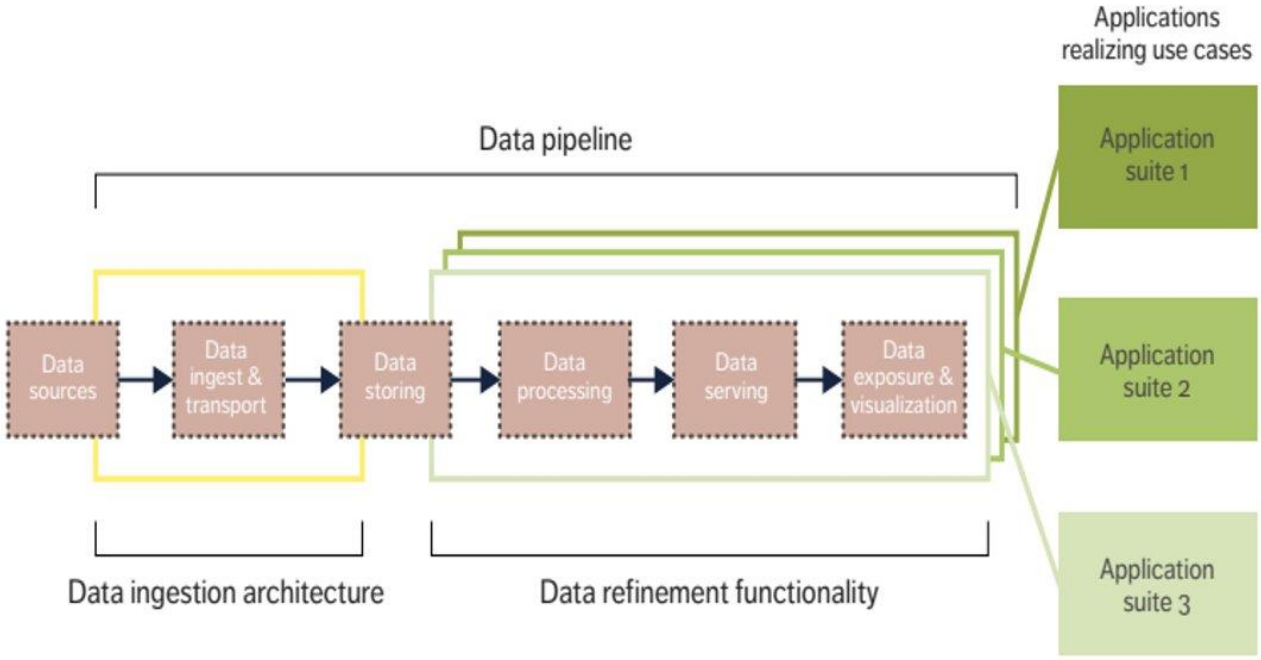

Figure 4-15 Data Ingestion Architecture for Telecom

Within the context of DataOps, the studies focus on defining and developing data and model management schemes. These schemes encompass various aspects such as data collection and pre-processing in Sections 4.1.2, 4.3.3, and 4.3.1, distributed model and feature selection in Sections 4.1.2 and 4.2.1, and privacy-aware data classification in Section 4.2.2.

4.5 Intent Based Management (Zero-Touch)

The importance of Intent-based Management (IBN) in the context of 6G networks stem from the increasing complexity and scale of future communication systems. 6G is envisioned to bring revolutionary advancements, including ultra-high-speed connectivity, low latency, and massive device connectivity. With billions of devices and diverse applications operating in dynamic and heterogeneous environments, traditional network management approaches struggle to cope with the complexities and demands of 6G networks.

Intent-based Management offers a new paradigm that aims to address these challenges. It focuses on capturing high-level intentions or desired outcomes from network administrators and translating them into automated and dynamic network configurations. By leveraging artificial intelligence, machine learning, and automation techniques, Intent-based Management can analyse the intent, monitor the network's state, and autonomously make adjustments to ensure the desired outcomes are achieved. This approach simplifies network management tasks to express objectives in terms of service requirements, performance goals, and security policies, rather than dealing with low-level configuration details. Intent-based Management promotes agility, flexibility, and adaptability in managing the complex network infrastructure of 6G.

Figure 4-16 illustrates the main interactions of the intent management function within a layered operational infrastructure on the left side. The intent management function receives all intents directed towards its autonomous domain and provides feedback to the intent origin regarding the successful fulfilment of the intent, completing an intent-based control loop. On the right side of Figure 4-16, an example is given for three levels of intent-based operation. The service-level intent manager is part of a SMO system. In this example, the solution at the service level impacts both the network function management and the RAN.

In the realm of advanced system management, the integration of IBN and AI catalyses dynamic self-adaptation and optimization across different dimensions. One facet of this synergy lies in distributed solutions and user self-adaptation, where AI-driven intent-based management seamlessly merges with distributed systems, allowing user intent to guide system behaviour. This integration empowers AI to interpret user preferences and translate them into real-time adaptations, such as automatically reallocating resources within a distributed

cloud environment to optimize application performance. Meanwhile, the convergence of AI and intent-based management also leads to intelligent, self-monitoring systems in the realm of AI-based solutions with automatic monitoring and management. Here, AI's continuous monitoring of system performance against defined intent becomes the foundation for automated corrective actions, maintaining system equilibrium in alignment with intended objectives.

Furthermore, this holistic approach extends to KPI monitoring and decision support. Intent-based Management complements KPI monitoring by aligning system behaviour with predefined objectives, a synergy enriched by AI's prowess in continuous data analysis. By constantly evaluating KPIs against intent, AI provides insights that guide well-informed decision-making. Collectively, the integration of Intent-based Management and AI fosters an adaptive, efficient, and informed approach to system management, propelling advancements across a spectrum of contexts and industries.

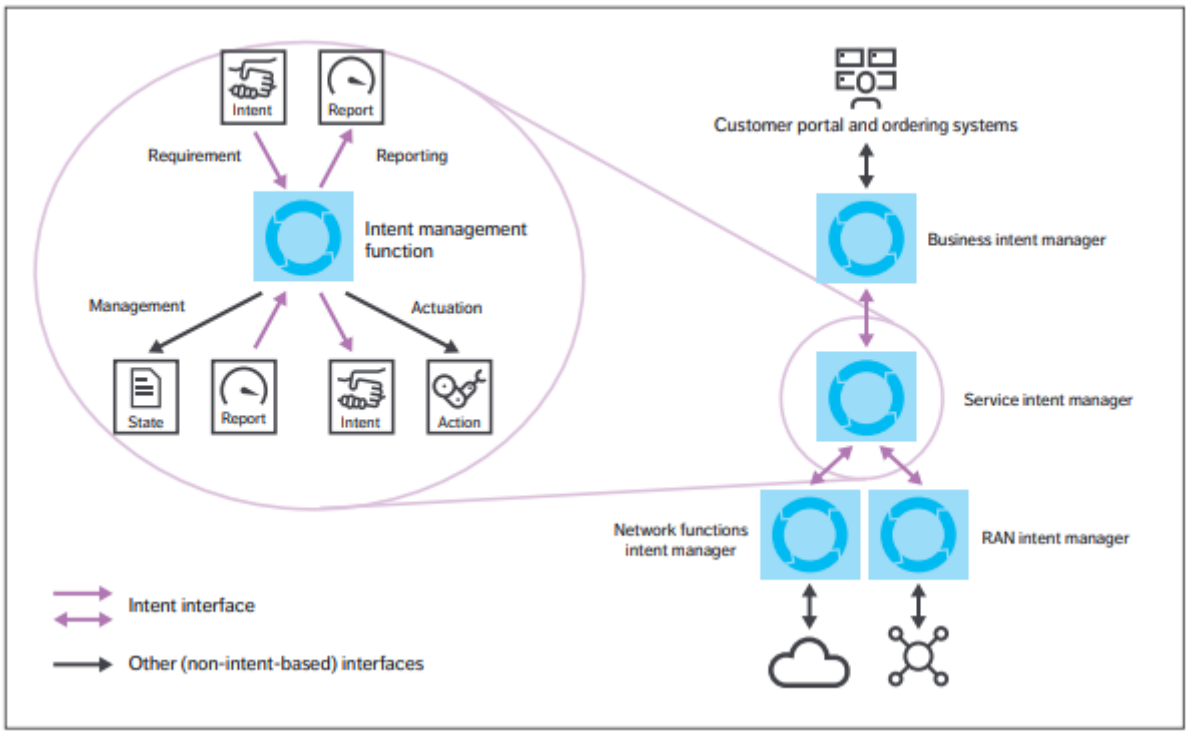

Figure 4-16 Creating Autonomous Networks with Intent-Based Closed-Loops

Within the context of IBN, the studies focus on distributed solutions that enable user self-adaptation, as described in Section 4.2.3. Additionally, AI-based solutions, including automatic monitoring and management, are discussed in Section 4.2.5 and KPI Monitoring in Section 4.1.3.

5 Network modularisation

6G needs to support a wide range of use cases and deployment scenarios with various requirements. For enabling flexibility without increasing complexity, 6G system (6GS) needs an easily deployable architecture of modules that can grow, and change based on current needs [23.501]. Network modularity targets to decompose the 6GS into orthogonal building blocks (i.e., network functions, services and interfaces) with the right level of granularity. Modularisation of the network functions needs to be performed with an E2E vision, considering not only the network function granularity but also the necessary interfaces and deployment options to incorporate existing and new use cases such as NTN, programmability and Everything as a Service (XaaS). Moreover, the security implications of the network modularization need to be detailed including but not limited to the management of trust among different entities, i.e., network modules, layers or slices [HEX2-D21]. Figure 5-1 demonstrates the envisioned modular network architecture where the modules can be customized for the procedures or specific KPIs.

The following sections detail the network modularisation enablers and respective study areas of focus. The enablers are divided into five main clusters (cf. Figure 5-1), namely (1) optimized network function composition, (2) streamlined network function interfaces & interaction, (3) flexible feature development and run-time scalability with modular network functionality, (4) network autonomy / multi-X orchestration and (5) network migration. In 6GS, maintaining a balance between network function granularity and the number of required interactions between the network functions or the modules have a pivotal importance. *Optimized network function composition* will investigate the trade-offs of network function composition in 5G and analyse the advantages and disadvantages of different decomposition options. Extending the first enabler's findings on NF composition, *streamlined network function interfaces and interaction* will demonstrate how the architectural modules and their external interfaces need to evolve for different use cases as well as distributed and centralized deployments. Built upon NF composition and interface optimization, *flexible feature development and run-time scalability with modular network functionality* investigate how the modular structure should vary within the context of RAN disaggregation and slicing where the NFs are customized according to the requirements of the specific services. *Network autonomy/multi-X* orchestration focus on determining the extend of NFs for a tenant to control a network slice and how these functions may vary depending on the level of control allowed by each tenant.

Hexa-X-II brings out a large set of new features for 6GS that are including but not limited to network modularisation, improved cloud platforms, flexible topologies and subnetworks. The final study on *network migration* will cover migration aspects from 5G to the proposed 6G architecture including interoperability of new 6G features with already defined functionality and architecture design principles.

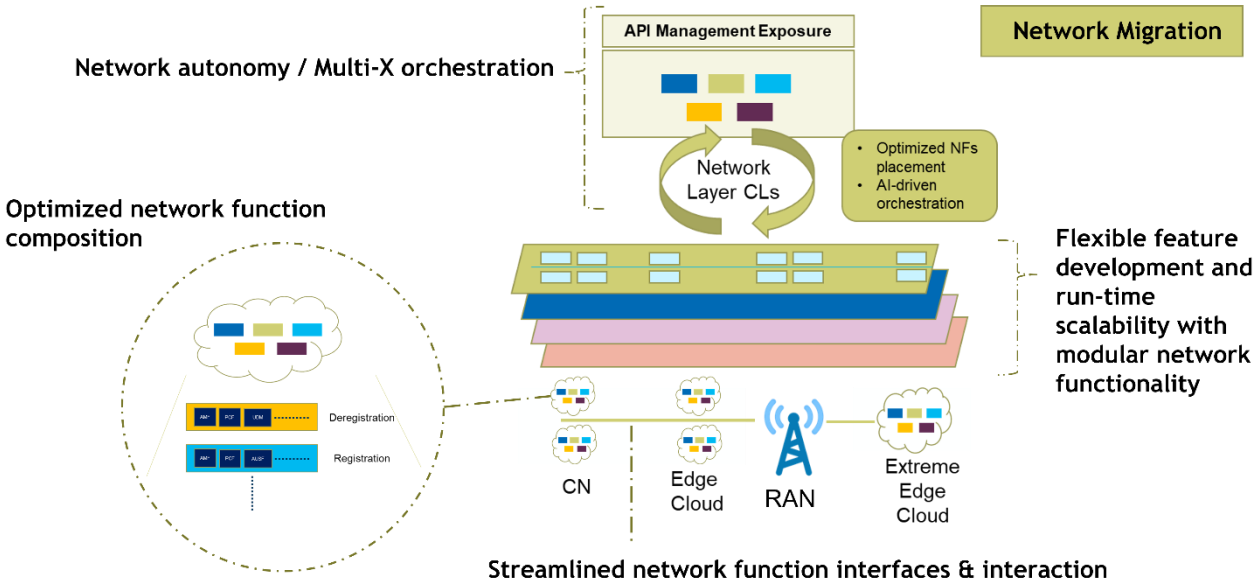

Figure 5-1 Overview of modular network design and enablers

5.1 Optimized network function composition

To overcome flexibility and reliability limitations of the previous generations, the 5G core is built upon the Service Based Architecture (SBA) approach, where the network functions have a high degree of functional decomposition. In this architecture, network functions (e.g., AMF, UPF, etc.) need to interact with each other using predefined interfaces to complete procedures [23.501]. However, in SBA, the high inter network function (NF) dependency and inter-NF interactions (i.e., via http/2) can cause low signalling performance and increased latency in 5G. In addition to the inter-function dependencies, 6G network will have an increased infrastructure complexity due to the need to integrate the near-edge and the far-edge with the Cloud Continuum [HEX-D63]. To be able to improve flexibility, overhead, scalability and efficiency of 6G networks, there is a need to analyse the different network functions within the 5G core architecture and understand the dependencies between these functions, how they interact with each other to deliver end-to-end services, and how the architecture needs to be evolved. One aspect of this is that the 6G network functions should be designed to be as self-contained as possible and with as little dependency as possible among network services. By reducing dependencies there will be fewer interfaces and processing points and, thus, possibility to improve the procedures and signalling. Also, this approach provides a smoother introduction of new functions and services in the future. The analysis should also focus on recognizing real-time KPIs to develop mechanisms that are able to dynamically adapt network functions to changing network conditions (i.e., changing traffic patterns, user demands and network state). This enabler will identify the trade-offs and dependencies between various network functions to optimize their composition for improved KPIs, i.e., designing network modules.

Modular design of the network functionality brings the opportunity to place the network functions or functionalities that are dependent on each other in the same module. By eliminating the inter-function dependencies and maximizing the relevance of the NF functionalities within a module, modular design can optimize the signalling and latency. It is also possible to further optimize the modules for specific services or deployment options to meet specific KPI targets. However, the functionality should not be removed from the module design only to reduce complexity [CTM+22]. As a general design principle, in an effective modular architecture the dependencies between modules are expected to be low, whereas the relatedness of the NFs within a module should be high [VK21]. One end of this design spectrum is a big monolithic design where all the network functions are placed within the same module, which would minimize the relatedness of the network functions and the inter-module dependencies. Although this big monolithic design would minimize the signalling costs, it would also increase the possible failures, limit the flexibility, have higher maintenance complexity and be difficult to adapt to changes. In such a design, a failure in one service would affect the entire system [VK21]. The other end of the design spectrum is a full disaggregation of the network functions, having one module per functionality. In this design, both the relatedness of the functionalities within a module and the inter module dependencies would be maximized. Fully disaggregated network functions can provide a large flexibility, but it can also lead to a high signalling cost and management complexity. Therefore, for the modular design to optimize the performance of the predecessor generations while providing a high level of flexibility, it is crucial to have a clear definition of relatedness and determine the trade-offs of different modularization options. The relatedness of network functions can be defined in various ways. This enabler focuses on two major ways, i.e., procedure-based and performance-based. In a procedure-based structure, the network functions are within the same procedure in the same module, cf. Figure 5-2 Optimizing the network function composition. This way the inter-module dependencies would be minimized, whereas the relatedness of NFs within a module would be maximized. For the performance-based design, the modules would be created to minimize a set of KPIs, such as a high level of flexibility and multiplexing gain. Therefore, depending on the considered KPI metrics in the design process, the inter-module signalling of the performance-based design might be higher than the procedure-based design.

The optimization of NF composition aims to achieve the KPI and KVI requirements that are defined for 5G and extending those with the ones that arise from new use cases, such as energy consumption and social sustainability among others. The KPI selection should consider the future 6G network composition over the Cloud Continuum and its capacity of scaling differently based on the different hierarchical level. To achieve the increased flexibility, optimized signalling and resource efficiency, several factors need to be considered during the module design, including the number of hops, network latency and availability, the availability of the functionalities for particular use cases, processing time, variability in demand [SRH17]. The optimization of the network composition provides high level of deployment flexibility as well as performance improvements

(e.g., efficient signalling, latency). By customizing the network function composition to particular use case requirements and specific deployment options, it is possible to achieve enhanced reliability and resilience. Although it can support various use cases, this enabler is especially focusing on immersive telepresence for enhanced interactions and from robots to cobots use cases.

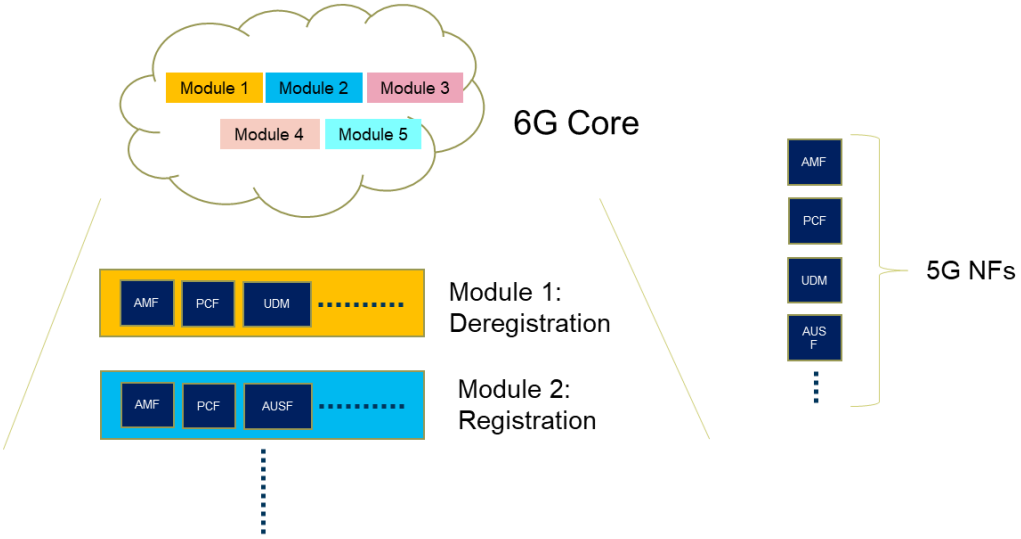

Figure 5-2 Optimizing the network function composition.

5.1.1 Procedure-based functional (de)composition for core NFs

The 5G core is defined today following the SBA approach, which increases the flexibility and agility in introducing new functions to the core network. In this architecture, a set of NFs (e.g., AMF, SMF, etc.) are defined such that each function has a specific logic to execute. Each NF in this architecture produces services that can be consumed by other NFs. The core network supports handling different essential procedures such as UE registration, UE deregistration, PDU session establishment, etc., by defining the interactions and information exchanged between these different NFs [23.502]. However, this interaction between the distinct NFs to execute certain procedures results in an increased volume of signalling traffic between the different NFs to exchange messages and information elements. In addition, the Procedure Completion Time (PCT), i.e., the time needed for a procedure to be fully executed, increases because of the inter-NF communication between the different involved NFs [GSH+22]. Alternative designs can be studied for 6G where a new set of control plane core network functions can be defined to reduce the inter-NF signalling in the system as well as the procedure completion time.

This study item will focus on the design and implementation of a new set of core NFs called **procedure-based NFs**, where each procedure-based NF includes the logic required to execute one full procedure such as UE registration, UE deregistration, etc., unlike the current architecture where the logic needed to execute a full procedure is distributed among different NFs (cf. Figure 5-3). To develop this solution, the processing logic and services offered by different NFs involved to execute a complete procedure call will be grouped together to create the self-contained procedure-based NFs. For example, one procedure-based NF is the UE registration NF which is made up of the processing logic needed to execute the UE registration procedure that is now distributed in the following 5G NFs: AMF, AUSF, PCF, NRF, UDM, and UDR.

The new design should be compared to the 5G core NFs that serve as a baseline. Different metrics (e.g., PCT, signalling overhead, etc.) should be evaluated to assess the advantages and disadvantages of the proposed procedure-based core architecture. The overall performance of NFs based on this new design is determined according to the volume of signalling traffic and the procedure completion time.

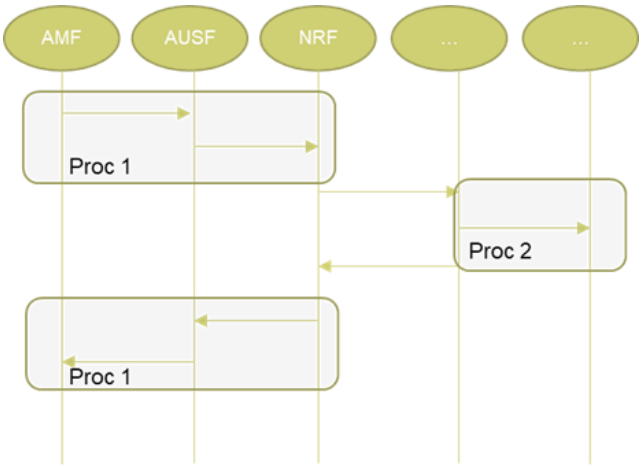

Figure 5-3 Procedure-based definition for the control plane functions of the core network.

5.1.2 Efficient signalling – separation of concerns

The support for SBA in 3GPP networks was introduced with Release 15. At that point in time the "service based" part comprised NFs for the part of 5GS referred to as the core network (CN). Since then, there have been proposals [HEX-D52][HEX-D53] to extend SBA to also incorporate the RAN. Part of this change involves adding service-based interfaces (SBI) to enable signalling directly between NFs. With SBI there might be less need or no need at all to use functional proxies such as the AMF, depending on how services and system procedures will be designed. Note that depending on relationships between entities, some level of proxying might be useful to avoid exposing a too large set of services.

The high-level model of the 6G architecture [HEX-D53] consists of two parts: radio network functions (RNF) and shared network functions (SNF).

The SNFs are generally responsible for larger areas in the network (or the whole network). As such, relocation of these functions, e.g., due to mobility, is less critical. The SNFs include reusable, and preferably, self-contained services/NFs, allowing independent scaling, and striving to use mainstream solutions (e.g., for security).

The RNFs are typically responsible for a smaller area in the network, e.g., covering a "single base station" area (e.g., like D-RAN) or a larger area (e.g., like a C-RAN deployment). An assumption regarding RNFs is that they are not divided further for multi-vendor inter-working purposes, with one exception being the interface towards the radio unit (RU). The RNFs could of course have an internal implementation architecture with multiple parts.

RNFs are responsible for time critical procedures in the network e.g.:

- Handovers (HOs), for ensuring that a UE is connected to the best cell, minimizing risk of HO failures, and reducing service interruption;
- QoS modification, to ensure that new flows with different priorities are treated correctly;
- State transitions (from sleep to active), which affects end user performance;
- Radio resource reconfiguration including additions of component carriers, e.g., to ensure that UEs get access to full BW with minimum latency.

In 6G it is still important to maintain and improve the performance of time critical procedures. The RNF needs to support multi-vendor HO.

Some details of the RNF are introduced with emphasis on time-critical procedures. The RNF entity in the SBA that handles these CP time-critical procedures is called UE handler. Mobility in this case involves change of UE handler. The change of UE Handler due to mobility implies that other network functions which established services with the source UE Handler will be now contacted by the target UE Handler, and this generates unsolicited notifications plus creates coupling as services should be "transferred" among UE Handlers.

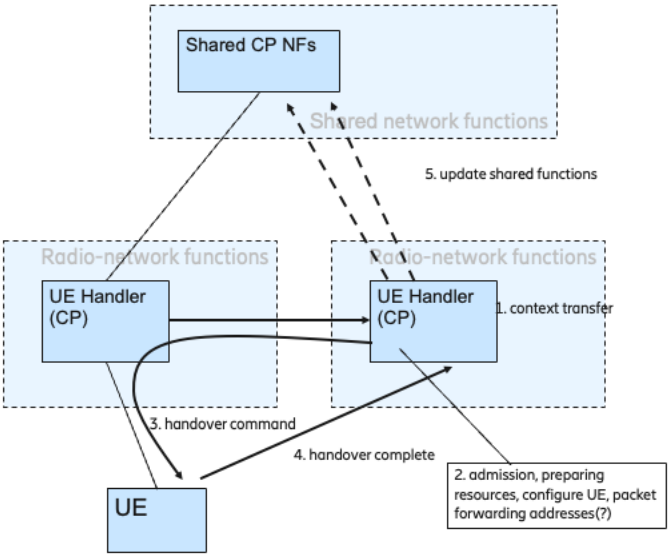

Figure 5-4 Example of HO with UE handler.

An advantage with this architecture, demonstrated in Figure 5-4, is that other CP functions do not need to be involved in handover preparation. The other CP functions can be updated after the handover, enabling service modifications, etc. The process builds on one of the design objectives established in [HEX-D51], namely, to avoid duplicated functionality. Further, the study presents more details on this example.

In this study the idea is to introduce improvements to the architecture. An important task is to try to quantify the outcome of these improvements by comparisons between new and existing solutions. Assuming that different NFs have different requirements, e.g., some are radio near (so called RNFs mentioned above) and (latency) critical while some NFs can process data whenever, it seems likely that to show that our proposed changes to the network are really efficient we need KPIs with more than one dimension, e.g., a spider diagram. Hence, in Hexa-X-II we will provide a "KPI map" where the designed NFs can be evaluated in several dimensions. The KPI-map discussed in Hexa-X [HEX-D53] will be used as baseline.

Here the idea is to study further how to design efficient network functions and how these functions interact in the architecture, in terms of combined KPIs such as latency, failure points, dependencies and number of messages. Each KPI represents a dimension or axis of the KPI map described in [HEX-D53] and can for example be the following:

- Latency to execute a procedure is still an important KPI; latency was discussed and demonstrated in [HEX-D52];
- The number of functional dependencies indicates how many times a certain entity depends on another entity to complete a task. This measure impacts latency and error handling, i.e., failure to signal between NFs. The KPI is discussed in [HEX-D52], as "good separation of concerns";
- The number of functional processing occasions or points indicates how many times a functional entity must process messages received from another entity. Once again latency is affected by the individual processing times;
- The number of failure point indicates how many times a functional entity requires a re-start of a procedure resulting from a failure to send/receive a message. Note that the number of failure points is not only an indication of the number of dependencies between NFs but also an indication of the likelihood that a process is interrupted. This is also a measure of resilience.

Resilience is the ability to provide and maintain an acceptable level of service in the face of faults and challenges to normal operation. The importance of network resilience is continuously increasing, as communication networks are becoming a fundamental component in the operation of critical infrastructures. Thus, when designing a new architecture, it is important that the level of resilience is maintained and, if possible, increased. To ensure that this is the case some measure of resilience should be included when new design is evaluated. This contribution is an attempt to quantify resilience of new solutions.

5.1.3 Optimised composition and placement of 6GC functions

The 5G Core (5GC) uses a message bus between control planes entities that allows for programmability of the 5GC Control Plane and thus creation of customised CP behaviour, for example, implementation of the context-aware CP services. The NF of CP in 6G are expected to be virtualised (cloud-native), which in the case of the cloud continuum, allows flexible orchestration and placement that breaks the current split of a system into domains. For example, virtualised CN functions can be placed in hosts (or DCs) where also RAN functions are placed and vice versa. The decomposition of CP into services provides the opportunity for a service-centric approach in contrast to function-centric approach used in previous generations of mobile networks. The approach allows for the runtime optimisation of KPIs of predefined CP service using dynamic placement or cloning of highly granular CP functions and their composition.

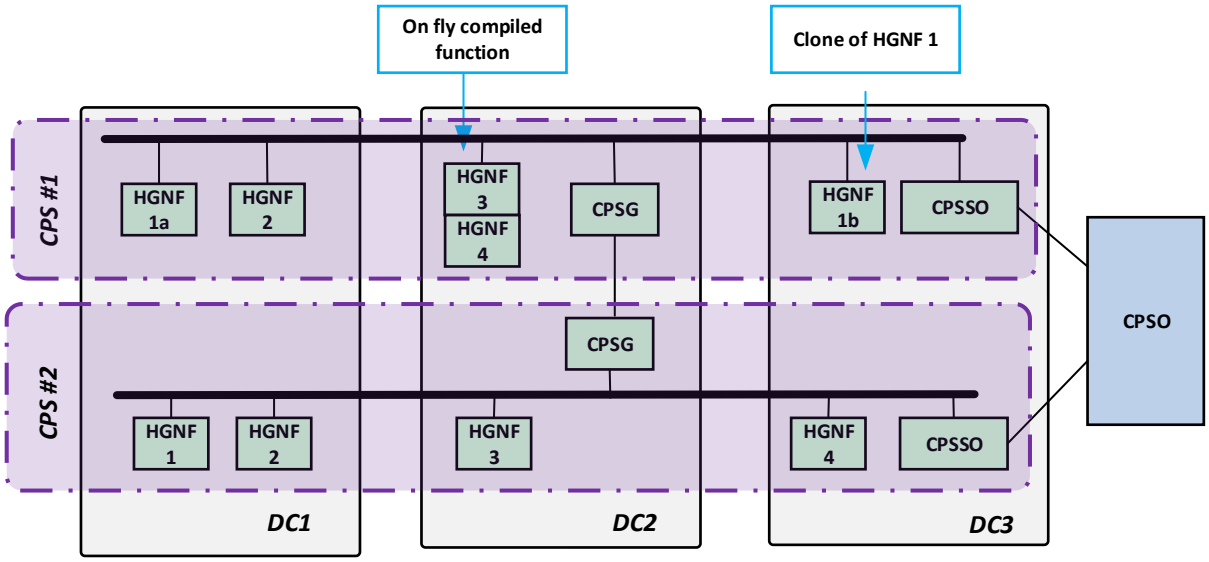

Figure 5-5 Service-Centric Control Plane concept.

The proposed Service-Centric Control Plane (SCCP) is composed of multiple, Highly Granular Network Functions (HGNF) that can be grouped, forming Control Plane Service (CPS).

A HGNF functionality is defined in a way that allow it to be a member of a specific service (security, mobility management, etc.) and its placement impacts the service KPIs (reaction time, traffic overhead). Each of the CPSs (for example Mobility Management) is self-managed, exposing externally information about its status and performance. CPSs Gateways (CPSG) allow interactions between different CPSs. The primary value of the proposed concept is the optimisation of HGNF placement in a way that maximises the performance of the CPS that they are forming. For the evaluation of the CPS may include KPIs as they are seen externally (e.g., message processing statistics) and internally by the assessment of the CPS traffic or CPS-related energy consumption. Such evaluations are made by the CPS Status Observer (CPSSO) entity which also analyses the impact of different placements of HGNF on the CPS performance. Such evaluation, probably made by the AI-driven algorithms, will deal with the possibility of a new placement of a specific HGNF, its cloning or termination. As a result, the CP traffic will be optimised in the context of CP reaction time and traffic volume. Using the cloud continuum approach may lead (in extreme cases) to creating a nearly complete 6G network in a specific area only. The orchestration of CPSs has to be done by a dedicated orchestrator (CPSO), and the process should take into account not only intra-CPS traffic optimisation but also inter-CPS traffic optimisation This process can be optimised using by proper placement of CPSG. For the sake of performance and/or reliability, the placement of HGNF can be driven by affinity or anti-affinity rules. The affinity rules can be used to group some HGNFs, using run-time compilation to increase their performance.

Please note that the initial placement of HGNF can be done on the basis of some estimation of the CP traffic; however, during run-time, the CP topology will be modified. An important problem related to CPS run-time

orchestration is providing CPS service continuity during run-time orchestration. Multiple techniques can be used for such a purpose; one of them is to design HGNF as stateless.

5.2 Streamlined network function interfaces & interaction

Streamlined network function interfaces and interaction aims to optimize and simplify the inter-module interactions with an E2E vision as shown in Figure 5-6. The 6GS is challenged by: (1) tight coupling and timing constraints between the control and data plane, and (2) the increasing control plane activities that impact data plane handing, i.e., driven by the emerging use cases [JPQ+22]. The increase in the inter-module communication (e.g., process calls), also increases the response time. In addition to this inter-module latencies, the communication between the gNB and the core network increases the round-trip latency (i.e., between 10 to 20ms) with often little control from ISP [LBZ+21]. Therefore, further optimization of network modules based on deployment decisions and the implementation of SBI for state synchronization are critical to accelerate the control plane procedures [JPQ+22]. Built upon the findings of Section 5.1, this enabler will revisit the interfaces and the interactions between network elements (cf. Figure 5-6 Streamlined network function interfaces and interaction), and optimize the procedures, by enhancing the existing interfaces/procedures if needed, and removing the redundant or absolute interactions. Moreover, this enabler identifies the network functionality that can be optimized by co-locating or disaggregating the network function.

In runtime, the NFs need to interact "internally" and externally to perform management tasks. Their internal interaction and placement allow for a ‘runtime’ definition of interfaces, i.e., each function may have a different version of interfaces depending on their use. For example, in the case of a common placement of several NFs in the same location (cloud), a macro-function can be created using runtime NFs compilation to avoid a communication delay. Moreover, to achieve an optimal network modularization, it is of paramount importance to have clearly defined communication interfaces/APIs between the modules comprising the slices and, besides, clear levels of capability exposures on each network module. Traditionally, procedures related to network management and orchestration (M&O), NF inter-communication and network control have been developed as reference-point-based setups. In those kind of setups optimizations tend to be open-loop (i.e., there is no feedback among the network modules), and they have a limited configuration scope [GKM+22]. Therefore, new interface approaches are required in order to cope with the expected 6G use cases, regulating communication among the different network modules and even across administrative domains. The API Management Exposure architectural block presented in [HEX-D62] might represent a potential baseline to develop new interfaces that would support not only inter-module but also across administrative domain interaction. This block is able to replicate the functionalities of the ETSI ZSM cross-domain integration fabric [ZSM002] and, therefore, all of the network modules in the various layers can interact and communicate with one another using this block at a variety of granularity levels while adhering to a unified pattern, i.e., by exposing and consuming a subset of services and associated management APIs that can be controlled by access control policies. This model may be applied with a larger scope to reflect future federation-based interactions in addition to communication among M&O resources. Furthermore, it also adds functionalities within a single administrative domain such as: (i) API/endpoint registration, (ii) API/endpoint consumption and (iii) API/endpoint access control. Finally, the interfaces designed for this enabler should remain open, in order to be able to integrate standardized frameworks such as the 3GPP Common API Framework [23.222] or the CAMARA initiative [OD22].

A critical aspect of enabling inter-modular and cross administrative communications is to ensure the inter-module and inter-domain synchronization. Entanglement distribution in the future generation of classical-quantum communication network can be pivotal to enable several protocols that can enhance aspects like synchronization, security, and reliability. As presented in the model in Section 5.2.2, distribution of entanglement enables time synchronization which can provide femtosecond level accuracy. Apart from this, distribution of entanglement also facilitates the usage of highly secure cryptographic protocols such as quantum key distribution, secret sharing, etc. In an entangled system, man-in-the-middle attacks can be easily detected by observing the changes made to the qubit.

The network modularization and related interactions need to be customized based on deployment locations. In the context of modularity within extreme edge, the data centric networking within the cloud, edge, extreme

edge continuum should be revisited. Starting with 5G, mobile networks have moved away from the point-to-point model used by previous generations towards an SBA focused on a Cloud-native design. In release 15, 3GPP introduced a common control protocol (e.g., HTTP) for implementing two communication models in the SBA for core CP: request-response and subscribe-notify. Network Functions can communicate directly using the Network Repository Function (NRF) for discovery, or starting from Release 16, they can use the Service Communication Proxy (SCP) for mediated communication. The SCP allows centralized signalling monitoring, separating the discovery and selection processes. It offers advantages such as better control over service routing, stricter access control, and increased robustness to the SBA. For example, it can apply different traffic distribution schemes (e.g., round robin) based on capacity and availability. However, while the SBA is a step towards a microservice-based architecture, it does not fully achieve the desired levels of distribution, decentralization, and atomicity expected in 6G systems. Although not obligatory, the SCP simplifies SBA integration in highly distributed Edge deployments by serving as the single entry point for a cluster of Network Functions. Nevertheless, its centralized operation, like any other proxy, makes it vulnerable as a single point of failure, a limited attack surface, and a potential scaling bottleneck.

This raises the need for flexible service routing capabilities that allow the utilization of the distributed resources available over multiple infrastructures. This work embraces this vision and explores the utilization of data-centric and dataflow mechanisms for a seamless realization and interaction of composable services in a fully distributed, Edge-native 6G architecture. In this line, to efficiently cope with this new approach towards a 6G architecture, the foundation of the underlying SBA must be rethought from scratch. Data-centric networking [SYN+21] appears to be a promising networking model, instead of today's host-centric model, to better fulfil requirements of an Edge-Native 6G networks' SBA. By combining data-centric with dataflow concepts, 6G networks can be seen as a dynamic chain of serverless atomic Network Functions dynamically orchestrated in optimal balance between the consumed and available resources over the continuum. Ultimately, it will foster the adoption of serverless computing into 6G where Network Functions are loosely coupled and stateless by default. Such concept will prove to be essential to support a Network of Networks framework vision [ZXM+19], and to provide enhanced connectivity and services in a variety use cases, such as Public Protection and Disaster Relief (PPDR) scenarios [MAT+23] where the 6G fabric must adapt to extreme performance and global service coverage requirements but only when such communications are in place.

Finally, in the design of E2E network function decomposition and inter-module interactions, the native AI/ML function interaction across RAN-core interfaces should be defined. Momentum towards the utilization of AI/ML in mobile networks has been increasing, e.g., 5G core networks, in 3GPP, introduced analytics framework (i.e., NWDAF) that can train, employ and re-train ML models for the generation of various network analytics. Similarly, starting from 3GPP Rel. 17, AI/ML support within RAN for defined use cases, such as energy saving, load balancing, and mobility is being discussed. In this regard, AI/ML currently is an add-on feature and used in the scope of enhanced automation. However, to be able to support envisioned services and applications in 6G with careful considerations regarding system complexity, moving toward integrated AI/ML functionality, i.e., native-ness, both intra-domain and cross-domains gains ever increasing importance. To this end, this enabler identifies use cases that can benefit from such AI/ML framework extended to operate in a cooperative way and study their optimized enablement in the network. A pros/cons analysis of the resulting architecture will be presented based on determined performance metrics.

Streamlined network function interfaces and interactions optimize the inter-module interactions for various deployment scenarios considering the E2E requirements. Through this optimization, this enabler extends the modular network design flexibility to different deployment options and network topologies. It further optimizes the KPIs (such as latency or bandwidth utilization) by tailoring the inter-module interactions and interfaces to the physical limitations i.e., raised from deployment decisions. Therefore, this enabler can support a wide range of use cases, but within Hexa-X-II, it will be focusing on immersive telepresence for enhanced interactions and from robots to cobots use cases.

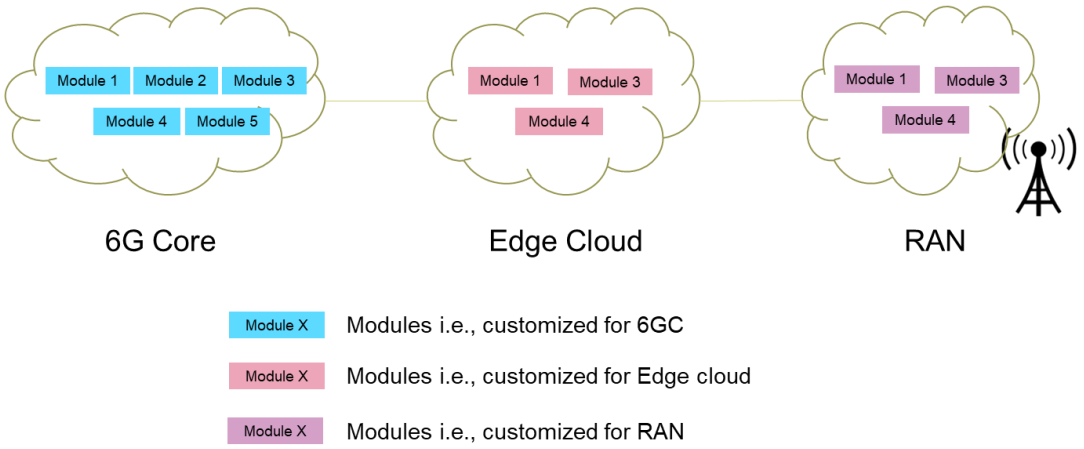

Figure 5-6 Streamlined network function interfaces and interaction

5.2.1 CN-RAN Refactoring

In 5G, RAN and CN have been developed and standardized separately, and therefore, following different design paradigms. RAN utilizes the traditional point-to-point reference interaction model whereas CN adopts SBA where services are defined and exposed/consumed using service APIs. Interactions between RAN and CN are done via a single-entry point in both domains, namely gNB/CU-CP and AMF, respectively [38.401]. The interface between RAN and CN (NG interface) applies control/user plane separation, specifically control plane (NG-C) and user plane (NG-U/N3) parts [38.413]. NG-C is further divided to N1 and N2 reference points; where N1 corresponds to the UE-CN interaction and N2 corresponds to RAN-CN interaction [23.501].

The separate development and standardization as well as the support for different use cases and scenarios (e.g., latency sensitive communication, latency tolerant communication, NPN, PN, roaming, etc.) resulted in duplicated functionalities (e.g., Handover, UE context/UE state handling, paging, etc.) in both RAN domain and CN domain in 5G. 6G can target further simplification, as well as performance optimization (e.g., latency) as per design objectives. CN and RAN functionality can be revisited to match 6G use cases and performance, and other key KPIs & KVIs while ensuring cloud compatibility. However, the trade-off between different modularization options (e.g., granularity) must be studied. For instance, on one hand increased granularity can lead to better scalability management, faster failure recovery and flexibility. On the other hand, it would lead to higher complexity and single point of failure. RAN can be disaggregated in terms of latency sensitivity of operations as well as user and control plane operations. The disaggregated logical entities can be deployed at different sites in the network. The CN can also be distributed and located in distributed cloud close to RAN nodes.

Enablement of disaggregated RAN and distributed CN functions in 5G has already initiated a reconsideration for the boundary between RAN and CN, as demonstrated in Figure 5-7. This is further accelerated by the growing trend for cloudification and edge cloud deployments as well as demand for flexible networks to offer service innovations with a plug and play approach. Consequently, this work starts by modelling the trade-off between design granularity and the performance of each model and based on this model identifying the network functionality that can be optimized by co-locating or disaggregating. Built upon the identified optimizations, the horizontal distribution of NFs through cloud/edge/RAN continuum to optimize KPIs & KVIs will be investigated. In particular, the evaluation of the concept will be done based on various evaluation criteria classified into two categories, namely design objectives and performance metrics. The design objectives can be listed as: upgrade flexibility, deployment flexibility, scalability, resiliency and robustness, security, simplification, modularity, and transition from and interworking with legacy systems. The considered performance metrics are latency, amount of signalling, number of hops, and energy efficiency. Finally, an analysis of the native AI/ML function interaction across the developed RAN-Core interface will be provided.

The refactoring of RAN-CN is applicable to deployment scenarios with cloud RAN implementations. Some considerations for the 6G RAN-CN architecture and interaction are beneficial for edge cloud communication and low latency services (cf. Figure 5-7).

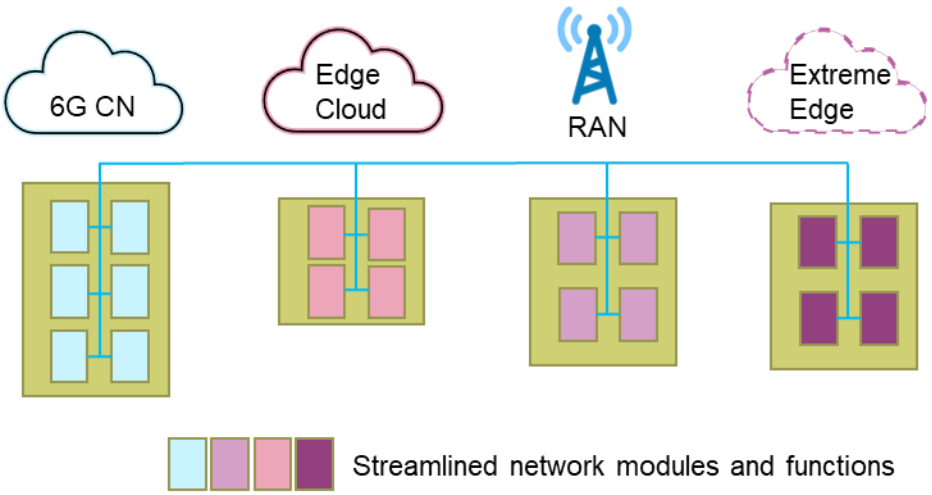

Figure 5-7 RAN – core functional split in the distributed cloud continuum.

5.2.2 Network Modularisation in Hybrid 6G-quantum Architecture

As softwarization of network is on a rise to cater various functionalities, it will affect aspects such as latency and computational overhead. In the context of classical communication there are limitations in terms of network-theoretic intrinsic limits [Sha48] which will hinder further development of future generation of networks, especially 6G. Introduction of quantum principles such as superposition of quantum states and entanglement, in the already present protocol stack can overcome the classical limitations. As opposed to classical bits, quantum bits or qubits stores information in a superposition that is a two-dimensional complex vector space. Measuring a qubit destroys this superposition and condenses the entire state into one bit of information, either '0' or '1'. In a multi-qubit scenario, there exist states that are mixture of multiple states, intertwined in a manner which cannot be written as a product state of its component systems. These special states are termed as entangled states. Qubits and its probabilistic nature along with entanglement facilitates several quantum protocols which makes it an appealing tool that can be used to go beyond certain classical limits. The management of a hybrid classical-quantum will require to define the interfaces of 6G network modules catering to both classical and quantum communication resources [FBD+21]. The development of such interfaces will pay attention to both distributed and centralized deployments, depending upon their specific application. The network displayed in Figure 5-8 is a continuum considered by the softwarized layers and by the management and orchestration.

The interfaces presented in Figure 5-8 enables the coexistence of classical and quantum technologies which maintains the segregation between data and control plane, as envisioned by SDN. Here, control plane manages the classical protocol stack as well as the integrated quantum physical-link layer resources. The southbound interface on quantum communication enables the control plane to issue commands and provide status information. Whereas the inclusion of northbound interface enables classical software applications to leverage quantum effects at the physical and link layer. With this hybrid architecture, KPIs such as latency and resilience can be targeted.

Additionally, usage of quantum synchronization can enable distribution of entanglement between various network modules for precise synchronization further progressing the distributed network aspect. This way, use cases such as 'cobots to robots' can be facilitated by such a hybrid network.

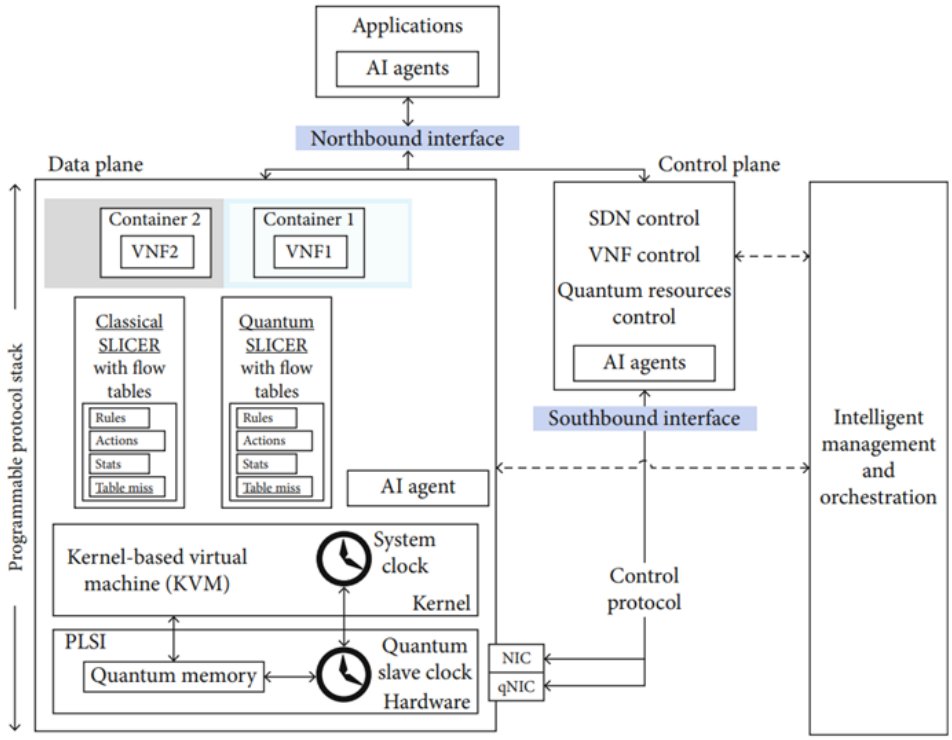

Figure 5-8 High level schematic of quantum-classical network architecture with network softwarization.

5.2.3 Data centric SBA for EDGE Native 6G Networking

The introduction of 5G networks brought about a major shift in system design, with the adoption of SBA and modular, Cloud-native approaches in the core network. This departure from point-to-point interactions between Network Functions towards standard service APIs has greatly enhanced the scalability and extensibility of the network when compared to previous mobile generations. However, with 6G research underway, the SBA model is being optimized to extend end-to-end orchestration to the Far-Edge domain and adopt serverless computing.

To efficiently cope with this new approach towards a 6G architecture, the underlying SBA must be rethought without ties to any particular communication architecture. Data-centric networking [ZXM+19] appears to be a more suitable networking model for an Edge-Native 6G network's SBA. Combining data-centric with dataflow concepts will enable 6G networks to be seen as a dynamic chain of serverless atomic Network Functions orchestrated in optimal balance between consumed and available resources over the continuum.

Ultimately, this concept will be beneficial to support a "Network of Networks" framework vision [SYN+21] and provide enhanced connectivity and services in a variety of use cases. This may prove crucial in meeting extreme performance and global service coverage requirements in scenarios such as Immersive telepresence, or sustainable development scenarios.

This study area presents the foundations for a data-centric interaction, modelled as a dataflow, of serverless atomic functions as the redesign of current core network concepts and procedures. The most promising path to address the aforementioned issues, as already identified by 3GPP in Release 18, is to adopt data-centric solutions for the SBA and to leverage on its name-based routing capabilities. We will study how the application of such techniques may be used to re-architecture the functions and services for the upcoming 6G network design.

5.3 Flexible feature development and run-time scalability

Flexible feature development and run-time scalability with modular network functionality aim to achieve two KVIs of 6G networks, i.e., flexibility and efficiency [HEX2-D11]. 5G technology provided a large degree of

flexibility with respect to the predecessor technologies through virtualization and customization. In particular, the network slicing concept in 5G enables the logical separation of the physical network resources. In these isolated logical network slices, the resources can be customized according to the requirements of the underlaying use cases. However, the network operators have limited options to customize the network functions of different slices as the 5G network functions have a well-defined set of functionalities and the operators can only enable or disable certain functionalities of the specific NF which can decrease the E2E performance of the network slices for particular use cases. These unnecessary functionalities cause the consumption of limited resources, that includes not only storage but also energy. The network modularity, i.e., analysed in Section 5.1, needs to be extended in the context of RAN disaggregation and slicing where the minimum set of functions required for a specific slice or use case are provisioned to a tenant on the data and the control planes. This enabler will analyse the modular enhancements to network slicing to meet the KPIs as well as the modular disaggregation in RAN, cf. Figure 5-9.

In addition to the customization of the network modules, further gains can be achieved by the horizontal distribution of the network modules within the cloud continuum. However, this distribution brings out the need to revisit the module design as the KPI metrics that are used to create the specific modules may change in this distributed deployment. For example, a modular design to minimize latency that contains a multitude of network modules may not be feasible for horizontal deployment due to propagation delay. Similarly, a massive module that contains a multitude of services may not be feasible to be deployed at the extreme edge/edge. Therefore, the horizontal distribution of the network functions and its implications on the network modules and the observed KPIs must be well understood. This enabler also focuses on the horizontal distribution of network functions or modules and the implications of different deployment options on the optimality of the module. The research activities will also outline how these deployment options should be integrated into the module creation process.

At the RAN side, user centric cell free design can enable runtime scalability. In a cell-free scenario under a disaggregated RAN, the RAN architecture needs to support that users are served by Radio Units (RUs) that might be connected to different Distributed Units (DUs), in turn possibly connected to different Centralized Units (CUs). This is to mitigate the poor performance for users that are at the boundary areas between RUs managed by different DUs, as it is the case of cellular networks. Users should not perceive any connection establishment when moving and being served by different RUs, i.e., the underlying RAN architecture should be as transparent as possible to the user, therefore eliminating any reminiscence of a cellular network. In 5G, PDCP dual connectivity is a partial solution, in which a user might be served by RUs connected to two different DUs [AKP+21]. However, this solution does not eliminate inter-cell interference between RUs that use the same carrier frequency, the fundamental cause of poor performance at the cell-edge. The solution to achieve a complete cell-free experience to the users, therefore, needs to involve lower-layer RAN protocols that address the cell-edge problem. A certain level of cell-freeness can be achieved by allowing RUs at the boundary of the coverage area of a DU to be connected to the neighbouring DU. In this case, each DU controls several groups of RUs. Radio resources are assigned, at the network level, to these groups. By choosing the RUs that compose each group, and allocating the resources accordingly, it is possible to achieve a scenario in which there are no users suffering from strong interference due to being at the cell-edge. Viewing cell-free as the next step from multi-connectivity, in the context of a flexible and reliable disaggregated RAN, it can provide the required reliability for applications of immersive telepresence, in which multi-connectivity provides an added level of resilience. Furthermore, and since the user mobility is made as seamless as possible, use cases that require high user mobility are also allowed for by cell-free networks.

Flexible feature development and run-time scalability ensures that the tenants can customize their slice according to the KPIs of the specific slice or the use case with an E2E vision. This customization provides a functional support for extreme requirements, including but not limited to low latency, high data rates and high reliability. As detailed in Section 3.1 and Section 3.2, the ultra-high flexibility to dynamically support QoS/QoE and the run-time scalability are key requirements for both immersive telepresence for enhanced interactions and from robots to cobots use cases. Therefore, this enabler will mainly focus on these two use cases and their respective KPIs/KVIs.

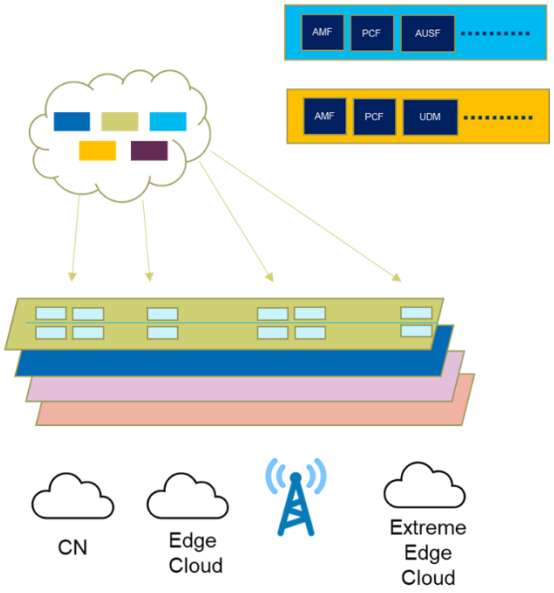

Figure 5-9 Flexible feature development and run-time scalability with modular network functionality

5.3.1 Flexible UPF design

In the 5G system, while the Control Plane (CP) functions are defined in a modular way following the SBA, the user plane (UP) processing is still handled in one monolithic function called UPF with many tasks to handle as listed in [23.501]. This monolithic design slows the development cycle of this function and hinders its scalability, reliability, and flexibility [Res22]. Modularizing the UPF is desired to improve its scalability, speed of development cycle, third party innovation and reliability by splitting the software to microservices where each one can be developed independently from each other as shown in Figure 5-10.

Modularizing the UPF should be carefully handled because of the performance-flexibility trade-off. In other words, while defining fine-grained modular UP functions brings higher flexibility in managing and controlling these functions, it also comes at the cost of impacting the packet forwarding delay because of the overhead added when traversing multiple hops in the UP path. Accordingly, a well-thought design that considers this performance penalty should be recognized.

The modular UPF (mUPFs) design should identify a set of subfunctions that enhance the flexibility, development cycle, scalability, and management of the UP in the core network without a big impact on the packet forwarding performance. The composition of the UPFs can be tailored towards specific deployment scenarios (e.g., Operator vs. Enterprise) and enables the independent scaling of UP functionalities (e.g., in the case of asymmetric up/downstream). This study item explores the modular UPF design in detail and analyse the UPF composition for different scenarios. An analytical evaluation of the overhead induced by the new design compared to the current baseline will be conducted in terms of the latency, throughput, degree of flexibility and scalability.

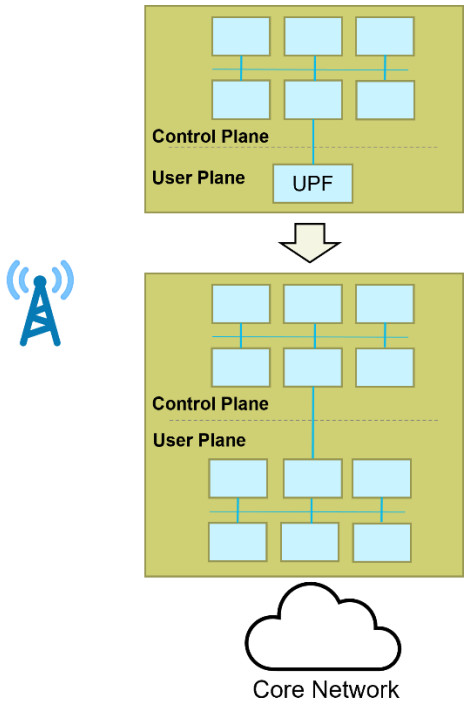

Figure 5-10 Modular user plane design in the core network

5.3.2 Split management of network slices

In 5G network slice management is done by a centralised OSS/BSS. The 3GPP has, however, defined some capabilities of the delegation of slice management to a slice tenant. The tenant may use the publish/subscribe mechanism regarding management data and management and orchestration operations. There are, however, already defined interfaces that allow interactions between tenants and system operators but they are based on operator OSS/BSS only. The present approach based on a single OSS/BSS raises scalability and security issues. The ITU-T has defined a solution which allows interactions of slice operator OSS (so-OSS) and slice tenant-OSS (st-OSS) – see Figure 5-11. In this approach, the tenant may have a dedicated management platform. The st-OSS is generally much simpler, and its role is to provide some synthetic information about a slice status (KPIs) supported by visualisation tools. The reconfigurations made by a tenant also should be high-level, and the preferred solution is to use declarative policy-based management (intents). In the 3GPP and ITU-T approaches, the implementation of shared management has not been described.

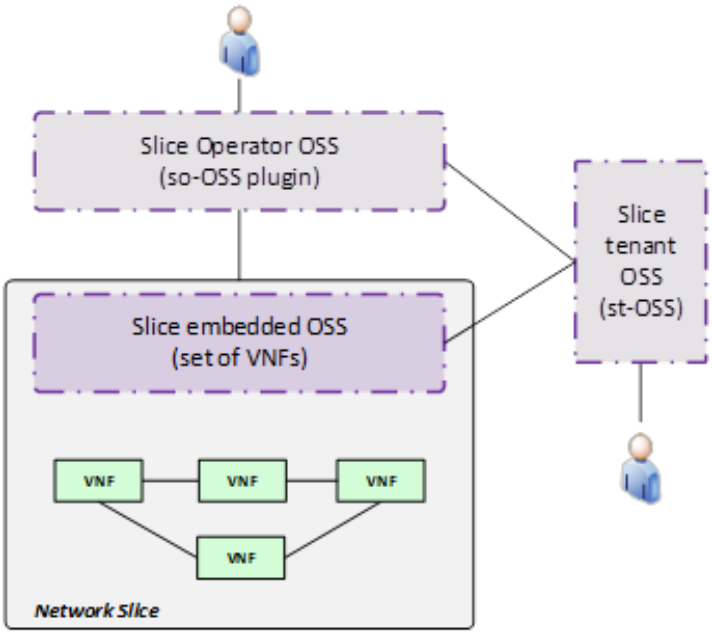

Figure 5-11 Split management concept.

A proposed approach lies in implementing essential management functions as a part of a slice template, i.e., in the form of programmable NFs (VNFs). It is also proposed to implement st-OSS and so-OSS as orchestrated cloud applications deployed with the slice. They can be seen as st-OSS and so-OSS plugins which may have a common, generic part. The st-OSS operations are monitored by so-OSS, which takes corrective actions in necessary cases. The split of functions between st-OSS and so-OSS can be made programmatically by creating several options of st-OSS and so-OSS twins or by defining both templates before deployment using dedicated tools. The management part of a slice template should use autonomic, i.e., control loops driven operations for the sake of management performance, simplifying the human-based management and increasing the network slice autonomicity.

5.3.3 Cell-free massive MIMO in disaggregated RAN

Current RAN components still lack certain levels of openness. While the interfaces between nodes offer the possibility to utilise handover and dual-connectivity features, operators still face difficulties finding options to combine hardware and software from different vendors. This has led to limited options for operators to deploy and optimally configure their RAN equipment [PBD+23].

The Open-RAN Alliance (O-RAN) [ORAN] promotes disaggregated and virtualised RANs with open interfaces connecting the different network elements. These networks offer multi-vendor interoperability, allowing for more flexible and heterogeneous networks. Furthermore, O-RAN claims capital and operational expense savings by lowering the entry barriers to new competitors, breaking out from the vendor lock-in situation in which operators depend on a limited number of vendors [GC21]. To overcome the limitations of the current RAN, the solution is in making networks more open.

The major bottleneck of network-centric cellular implementations is inter-cell interference [IBN+19, LHA13], which hinders densification of massive MIMO deployments. Cell-free massive MIMO aims to eliminate the inherent problems of cellular networks. One solution is changing from a reality in which access points are surrounded by users, to a reality in which it is the users which are surrounded by many access points, in a paradigm shift in which, from the users' perspective, there are no cell boundaries during data transmission [NAY+17]. This aims to eliminate the cell-edge problem of cellular networks, where certain users suffer from considerable levels of inter-cell interference.

The flexible architectures resulting from a disaggregation of RAN components need to support the operation of cell-free massive MIMO. To adapt to the new physical layer technologies brought by cell-free massive MIMO, the RAN control framework needs to be designed accordingly.

A study of current challenges regarding the implementation of cell-free massive MIMO in disaggregated RAN will be conducted, with the final aim of providing users an experience in which there are no locations in which the performance is too low, as an improvement of the experience cellular networks can provide. The study item addresses possible solutions of disaggregated RAN to support user-centric operation and provide users with a cell-free experience. It will simulate potential scenarios, focusing on the role played by the different RAN components, and determine how the physical resources are shared among the different (possibly overlapping) coverage areas [WDY+23, WSH+21]. Figure 5-12 depicts a possible RAN architecture to support cell-free operation. Interfaces between nodes shall be defined and their roles identified, in a way that does not greatly constrain the flexibility of the network.

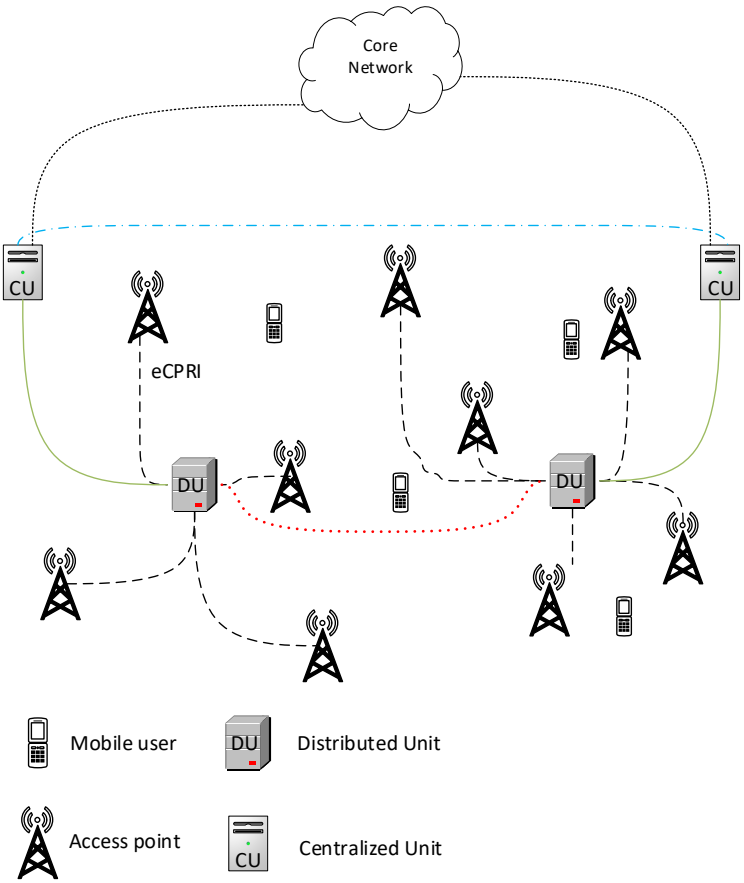

Figure 5-12 Cell-free RAN architecture

5.4 Network autonomy & Multi-X orchestration

Network autonomy and multi-X orchestration extends the findings of Section 5.3 to increase the flexibility and the efficiency of the 6G networks by introducing enhanced control and orchestration of network functions, cf. Figure 5-13. As detailed in the earlier section, network slicing has been a key enabler in 5G to accommodate multiple services with various requirements within the same infrastructure by creating logical networks. The virtual network resources allow the operators to tailor the network resources to meet the exact needs of the services. In this design, the management and orchestration of the network slices are partly built upon open-loop slice configurations and rely upon well-defined and (semi-)static parameters derived from service level agreements (SLAs) and network slice templates. However, these operations may lead to low resource utilization due to dynamic changing of network condition and ever-increasing number of devices and device types in mobile networks [AMC19]. In 6G, the control mechanisms need to be reconsidered to increase the autonomy of the decisions in network operations and orchestration. Beyond the capabilities of conventional 5G networks, future 6G network automation refers to the automation of network management and orchestration activities. To optimize network performance, reduce latency, and boost overall network efficiency, it is envisaged necessary to use improved AI/ML techniques compared to previous approaches to autonomic network management (e.g., to correlate and extract relevant data from different network domains). Also, to be

able to bring the development and operational teams closer together in such multi-stakeholder and multi-domain scenarios incorporating DevOps practices in this new telco-grade environment should be considered in order to get a higher automation level in the processes for developing, deploying and operating the network services. Due to the unique aspects of telco settings, this is a difficulty since it calls for collaboration between the Mobile Network Operator (MNO) and outside software suppliers, so it becomes essential to have a solution capable of dealing with those environments as reflected in [HEX2-D21].

Extending the concept of (SBA) introduced in 5G network, 6G network should rely on multi-level, multi-cloud technologies to achieve the maximum efficiency in executing network functionalities. The overall infrastructure will include multiple operators, each one with resources belonging to different domains of the Cloud Continuum. This increased complexity requires an additional intelligence which will be in charge of orchestrating the deployment in a seamless way in regard to the service provided to the final user. Since no assumptions can be made on the underlying technology powering the operators' infrastructure, the orchestrator should also handle this type of integration. Additionally, the goal of intent-based networking (IBN) is to develop into an access point for users who have scarce or even any knowledge of controlling and interacting with resources that are accessible in layers underneath the application layer (e.g., certain operational teams withing the MNO scope, or external vertical industries deploying their services on the MNO infrastructure, that will not be typically aware of the underlying network complexity, but will need to efficiently configure and customize their services). IBN can let tenants manage and interact with resources more easily because they offer an abstraction layer that separates the application layer from the underlying infrastructure. In this sense, Intent-based mechanism is considered a key enabler to achieve a higher degree of automation and programmability on future 6G networks.

The network autonomy and multi-X orchestration is built upon the idea of exposing the capabilities/data between network entities, i.e., network functions, layers or slices. It enables the close loop control of network functionality as well as providing the functional support to meet extreme requirements. The following subsections focus on the specific functionalities to enable network autonomy and multi-X orchestration. Although this enabler can support majority of the use cases, in Hexa-X-II, it will be focusing on two major use case families, i.e., from robots to cobots and immersive telepresence for enhanced interactions.

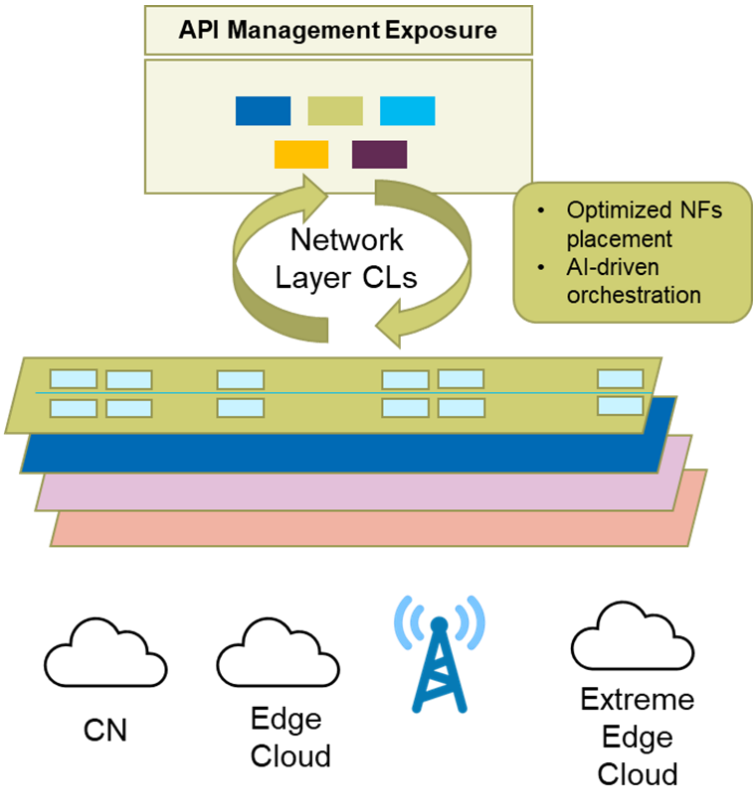

Figure 5-13 Multi-domain orchestration

5.4.1 Slice as meta module to aggregate separate modules

As detailed in Section 5.4, network slicing is a key enabler in 5G to support different services using the same underlying mobile network infrastructure. The verticals of a slice (i.e., slice customer) can customize the network according to the needs of the service. The isolation between slices ensures that the operation of one slice does not affect the other slices. Life cycle management of a network slice consists of preparation, creation, operation and termination phases. Preparation phase includes the network slice template design, capacity planning, evaluation of slice requirements, network environment preparation and any other process needed before the creation of the slice instance. Slice creation consists of allocating and configuring the needed resources to operate the slice according to the provided requirements. Slice activation, deactivation, monitoring, reporting, and reconfiguration of a slice instance is done during the operation phase. Termination phase terminates the slice instance after which the slice instance does not exist. Slices are pre-determined and configured with certain requirements, i.e., *semi-static slice templates*. Tenants do not have the option to control the slice configurations dynamically. This semi-static slice control causes lower resource utilization and QoS/QoE (e.g., delay, UE satisfaction, error rate etc.)

The current network slice preparation is rigid regarding the semi-static service requirements and consequently the network configuration parameters. This may lead to sub-optimal support for various services and their varying service characteristics considering the dynamic changing of network condition and ever-increasing number of devices and device types in mobile networks [AMC19]. Lack of handling slice instances in a more dynamic way, that may require distributed and/or hieratical orchestration, can also cause sustainability problems as some allocated and configured resources may remain unused. Although distributed and hierarchical orchestration options exist, central orchestration maintains its widely used status due to the short-comings of cross-domain orchestration. Addressing the issues with the cross-domain orchestration, slice operation related decisions can be made more efficiently and intelligently based on domain-specific information. Additionally, the trend towards AI/ML integration into mobile networks as well as the availability of immense amount of data envisioned for the 6G era can be leveraged by network slicing framework to offer differentiated services better.

This study item starts by determining the module requirements per slice and their configurations in a multi-tenant environment which also includes the horizontal NF placement per slice and tenant as demonstrated in Figure 5-14.

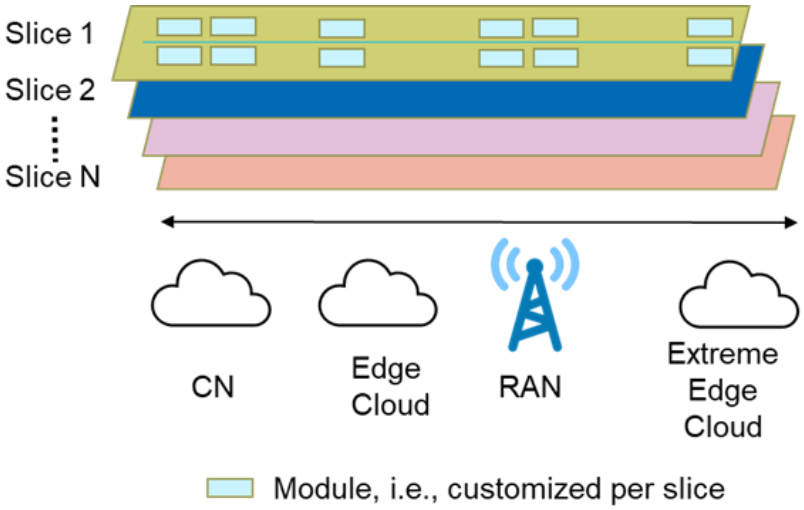

Figure 5-14 Illustration of slice as a meta module

The research will particularly focus on how the module design and location should change to support different slices. Extending this modular architecture, this study item works on formulating the enablers for dynamic slice control & management, that can lead to dynamic policy management. The dynamic management decisions and module placement should be seamless from the end user's perspective. Finally, this study item will work on enabling seamless E2E service provision, that integrates various resources and disaggregated &

distributed NFs, cf. Figure 5-14. The evaluations measure the performance based on the resource utilization, slice utilization and energy efficiency.

5.4.2 Network functions capability exposure and communication

Management and orchestration, NF inter-communication, and network control systems have been typically developed as reference-point-based setups. However, those kinds of setups have a limited configuration scope [GKM+22]. An answer to those limitations is the adoption of a cloud-native approach, as it has been anticipated in [HEX-D14] and [HEX-D62]. In cloud-native contexts, the exposure of capabilities of the different network components through well-defined interfaces can play a relevant role in improving the above-mentioned systems. Here are some key reasons:

- Service Composition: Cloud-native applications are typically built as a collection of loosely coupled microservices (which would be used to implement the different network functions in telco-grade environments) which work together to deliver functionality. By exposing well-defined interfaces, each microservice can communicate with other services enabling seamless service composition and promoting modularity, scalability, and ease of development and maintenance.
- Flexibility and Agility: Cloud-native services are designed to be highly flexible and agile. Exposing interfaces allows the different service components to be developed, deployed, and updated independently, e.g., by adopting agile development methodologies such as continuous integration and continuous deployment (CI/CD).
- Portability and Interoperability: Exposed interfaces provide a standardized way for different components and services to interact, promoting portability and interoperability. By adhering to common interface specifications, cloud-native services can be deployed and executed on various cloud platforms or container orchestration systems, which can help to avoid vendor lock-in.
- Cloud-native architectures emphasize scalability and resilience. Exposing interfaces enables horizontal scaling, where multiple instances of a service can be created and orchestrated to handle increased load. Exposed interfaces facilitate load balancing, fault tolerance, and fault isolation, ensuring that the overall system remains available and responsive even if individual components fail.
- Decoupling and Independent Evolution: Exposed interfaces act as boundaries between different components or services within a cloud-native application. This decoupling enables independent evolution and deployment of individual services. Each service can evolve and scale independently without affecting the overall system, promoting agility, maintainability, and scalability.
- Developer Experience and Collaboration: Well-defined exposed interfaces enhance the developer experience and enable effective collaboration. By clearly specifying the inputs, outputs, and behaviour of each interface, developers can understand how to interact with a particular service. Additionally, interfaces serve as contracts between different development teams, allowing them to work independently on their respective services ensuring compatibility and integration.
- Open ecosystem. That interface-as-a-contract feature mentioned in the previous bullet makes possible also to easily integrate with the MNO the different stakeholders from the different domains typically participating in the telco-grade environment, e.g., different external public and private network providers, software providers (e.g., network service developers), vertical industries, data providers (e.g., to train or deploy AI/ML models), etc.

In summary, exposure of capabilities in cloud-native contexts promotes service composition, flexibility, portability, scalability, resilience, decoupling, and collaboration, empowering stakeholders to build and maintain cloud-native applications that can be very modular, adaptable, cross-domain, and highly responsive to changing business needs. As it can be seen, the functionality provided by this enabler is quite general, so potentially applicable to all the use cases in Section 3.

The technical approaches that could be considered to implement this work could be diverse: e.g., the usage of well-known state of the art technologies such as RESTful APIs, message brokers and event-driven architectures, service meshes, or API gateways, among others. However, besides these general technical approaches, the study approached here would be more focused on the adoption of comprehensive API *Management* Platforms in the Hexa-X-II architecture. These API management platforms provide complete solutions for managing and exposing interfaces, offering features like API lifecycle management, security and

authentication, rate limiting, analytics, and developer portal capabilities. They can simplify the process of exposing interfaces and provide additional functionalities to monitor, control, and secure the APIs. API Management Platforms can play a relevant role in exposing interfaces enabling organizations to build scalable, secure, and developer-friendly API ecosystems. There are already some specific implementations in the state-of-the-art, such as Apigee [APIG], AWS API Gateway [AWSAG], or Kong [KONG], which could be used as a reference.

On the other hand, there are also several standards and formal specifications regarding this feature of exposing interfaces in cloud-native and API-centric contexts, which can help to establish best practices, promote interoperability, and ensure consistency across different systems. Here some of them:

- The Common API Framework (CAPIF) [SCT+22]: This is a set of 3GPP specifications to standardise some common capabilities exposed by the 5G core northbound APIs. The goal of this initiative is to establish a consistent and standardized northbound API framework that spans various 3GPP functions, enabling the utilization of the capabilities offered by 5G networks and ensuring uniform access to the exposed 5G features. Within this standardization process, the 3GPP have addressed different aspects, including the onboarding/offboarding of network functions, services discovery and management, events subscription and notification, as well as charging and security.
- The GSMA Open Gateway [GSMA]: This is a framework consisting of standardized network APIs aimed at granting developers a universal access to operator networks. With the backing of 21 mobile network operators, this initiative is a significant shift in how the telecom industry can design and deliver services in an API-driven context. It intends to enable faster enhancement and deployment of services for developers and cloud providers across operator networks by offering centralized access points to the world's largest connectivity platform. The APIs in this framework are developed and published in CAMARA [OD22], a project hosted by the Linux Foundation [LNXF] which provides the API definitions and reference implementations, which are freely available under the open-source Apache license.
- The ZSM integration fabrics: This is in the context of the ETSI ZSM architecture [ZSM002], which is envisaged to automate network and service management in multi-domain environments. This architecture defines different so-called “integration fabrics” in different scopes, which could be well aligned with this enabler: the “domain integration fabric” (specific to communicate components within a single network domain) and the “cross-domain integration fabric” (with the same functionality, but with a wider scope, covering different domains).
- OpenAPI [OAPI] (formerly Swagger): This specification allows defining and documenting RESTful APIs, providing a standard format to describe API endpoints, request/response formats, input parameters, and error codes. OpenAPI promotes interface discoverability, self-documentation, and code generation.
- Reactive Streams [RESTR]: This initiative provides a standard for asynchronous stream processing. It defines a set of interfaces, protocols, and libraries enabling reactive programming in applications. Reactive Streams can be also used to expose interfaces that involve the processing of streams of data, allowing for efficient and scalable communication between services.

Figure 5-15 shows a simplified high-level representation of the capability exposure concept compared to the legacy approach, where this concept would not apply. As it can be appreciated, in the legacy approach, common functionalities are duplicated across multiple services, which is not the case by using the API Management Exposure concept. Also, new common functionalities are also included, such as the APIs life-cycle management or the developer’s portal.

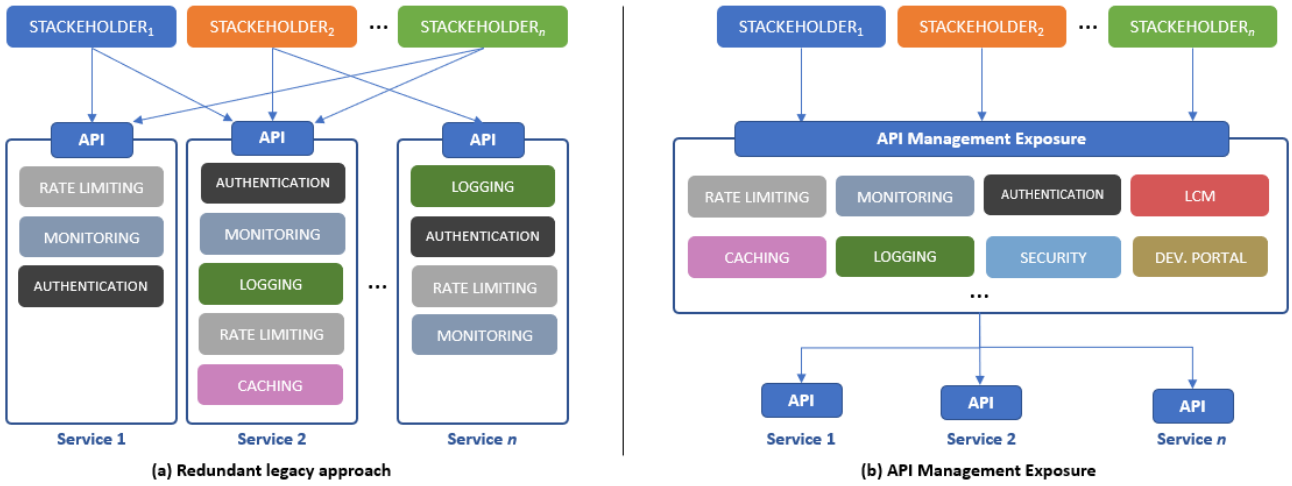

Figure 5-15 The API Management Exposure concept.

5.4.3 Network modularization over the Cloud Continuum

5G network standardization proposed a new approach based on the principle of SBA: monolithic functions were divided into microservices, increasing the granularity of micro functions. Moreover, enhancing on this concept, the slicing was introduced to define reserved and dedicated levels of services.

As regards 6G, we expect that it must natively support new generation technologies, in particular the Cloud Continuum. The Cloud Continuum is a seamless integration of various types of clouds that extends from the centralized cloud to the on-premises equipment, passing through the far-edge and the near-edge. This extended cloud needs to be coherent on every level and should be able to scale based on the different hierarchy level that needs more resources in a specific moment.

The cloud continuum owned by different operators should be federated among each other, facilitating seamless extensions across operator boundaries and national geographical borders, thereby promoting dynamic collaboration among MNOs.

Technologies that host 6G networks need to natively support multi-cloud scenarios, in which there are multiple levels and cloud domains whose underlying technologies may differ from one another. In this way, applications provided in next generation networks could easily benefit from a native coupling between network capabilities and services provided by cloud technologies, in particular on the edge.

Potentially 6G network functionalities should be deployed coherently on all the Cloud Continuum, while being effectively deployed only where the specific function, micro-function or module is needed and only for the time it is strictly needed as shown in Figure 5-16: Visualized NF composition and deployment over the Cloud

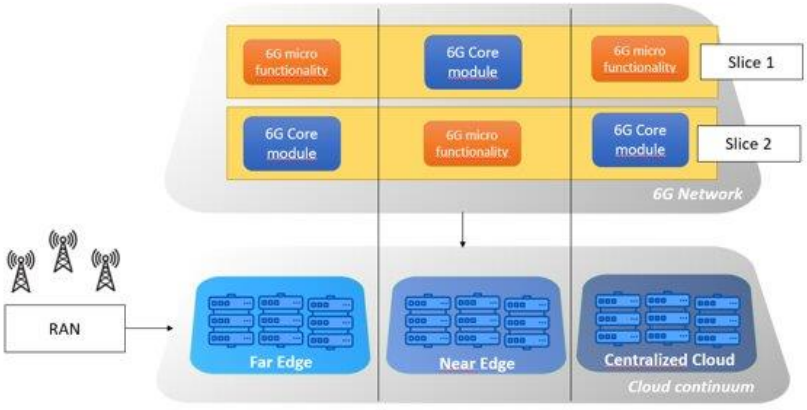

Figure 5-16: Visualized NF composition and deployment over the Cloud Continuum

Continuum. The main objective should be to minimize the cloud resources which have an increased cost at the edge for the operators. Thus, the role of the orchestration becomes of paramount importance to optimize the deployment of the functionalities/modules on the Cloud Continuum.

5.5 Network migration

When 5G was developed, there were several ideas [MET17-D24] that the new 5G deployments needed solutions that allow a gradual transition from 4G RAN, i.e., using 4G as a connection anchor for reliable connection and idle mode coverage, and to use 5G as a capacity booster when there was 5G coverage. This also includes the possibility to use 4G CN (EPC) as CN via the 5G RAN. This was achieved by leveraging on the 4G Rel-12 feature known as dual connectivity, where a UE could be connected to two different base stations at the same time with LTE or NR as master and secondary in any combination connected to either EPC or 5GC [37.340]. The final result in 3GPP was a standardization of multiple non-standalone (NSA) and stand-alone (SA) architecture options with different combinations of 4G radio, 5G NR, EPC and 5GC. However, in the implementation phase only two options (i.e., option 3x non-standalone, NSA and option 2 standalone, SA) were realized and deployed. The split into NSA and SA led to a phased introduction of 5G, causing fractional transition to the SA 5G architecture which resulted in delays on exploiting the full 5G capabilities (e.g., URLLC, slicing). In addition to the delays, this also increased the complexity for network vendors, device vendors and network operators.

The 5GA from Rel-18 will form the foundations for 6G migration. Migration from 5G to 6G should not repeat the complications that are faced in the previous generations, i.e., stemming from defining multiple migration options that would not be used. 6G should target at a smoother migration process which would preferably have a single step migration that would optimize the performance and limit the complexity of the systems. Procedures shall be streamlined where justified and the number of options within the 6G system as well in combination with existing systems shall be reduced whenever possible. Features and related parameters should be designed for practical deployments in such a manner that the number of options for the same or similar purpose and use case are avoided and technologies and features will be practical for real deployments and implementations. This study is intended to prevent the outlined 5G migration challenges from occurring in the migration to 6G and foster 6G usability and limit costs for all phases from standardization, development, deployment to operation. As a starting point determining role and alternatives to the non-standalone architecture 6G in the migration to the next generation needs to be determined to ensure timely and complete introduction of the 6G functionality. This study deals with how to perform an efficient migration from 5G to 6G from both operator, vendor and user perspective. In particular, it will focus on the pros and cons of different migration options.

The activities in this study will also point the coverage issue, e.g., how to deal with the limited coverage, throughput and capacity for 6G in the initial phase. The interworking between 5G and 6G (e.g., handover etc.) will be further elaborated, cf. Figure 5-17. Interworking with core network and RAN for 5G and 6G, e.g., spectrum sharing such as MRSS (cf. Figure 5-17) as well as spectrum considerations will be analysed.

The evolution from 4G to 5G brought major revolutionary innovations. A key innovation was the Service Based Architecture (SBA) in the core that was a step towards a cloud-native network architecture. The SBA in 5G core enables the independent integration of new 6G functionalities, which gives the opportunity to introduce new services and new functions for 6G use cases or services within the 5GC. Nevertheless, the functions with non-backward compatible changes may require an independent evolution of 6GC. Therefore, it is crucial to have a better understanding of the need for core evolution and investigate the new 6G functionalities that might justify non-backward compatible changes.

6G is envisioned to reutilize 5G components and principles as much as possible where applicable. The old network functions (NFs) would still be able to utilize common and shared functionality. This reutilization supports backward compatibility between 5G and 6G. However, as the research and development phase of 6G is still in the early stages, there is a major uncertainty in the 6G evolution and the related 6G functionalities. This uncertainty is reflected in the backward compatibility in 6G functionalities. The selected architecture option and the related backward compatibility can limit the innovation level of 6G. Therefore, this study will navigate these uncertainties and investigate the new 6G functionalities that might justify non-backward compatible changes, cf., Figure 5-17.

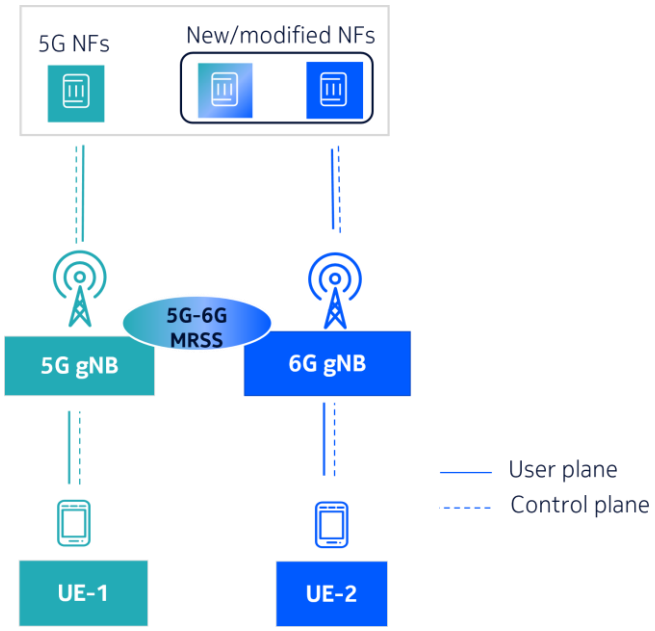

Figure 5-17 Overview of network migration from 5G to 6G.

5.5.1 5G-6G MRSS and 6G RAN coordination.

The introduction of radio access technology generations prior to 5G was driven in part by the availability of new spectrum and static refarming of existing spectrum resources towards the new generation. Bearing in mind that no new low band spectrum is expected to become available in the key 6G markets by 2030, it becomes evident that the ability to leverage existing 5G spectrum will play a pivotal role in the successful and cost-efficient migration to 6G. While traditional static spectrum refarming is, in theory, always an alternative, it might not be feasible in practice due to the disruption to existing services to support the initially low uptake of 6G devices. The Multi-Radio Spectrum Sharing (MRSS) approach should strive to attain maximum spectrum sharing dynamicity with 5G, accompanied by minimal overhead.

Figure 5-18 shows one option on a possible migration path, with 6G intra-RAT carrier aggregation (CA) used to combine capacity and coverage bands that are dynamically shared with 5G via MRSS. In an initial phase of 6G CA is expected to be used. Dual Connectivity (DC) may be used in a later phase when aggregated 6G carriers are not co-sited with similar coverage area.

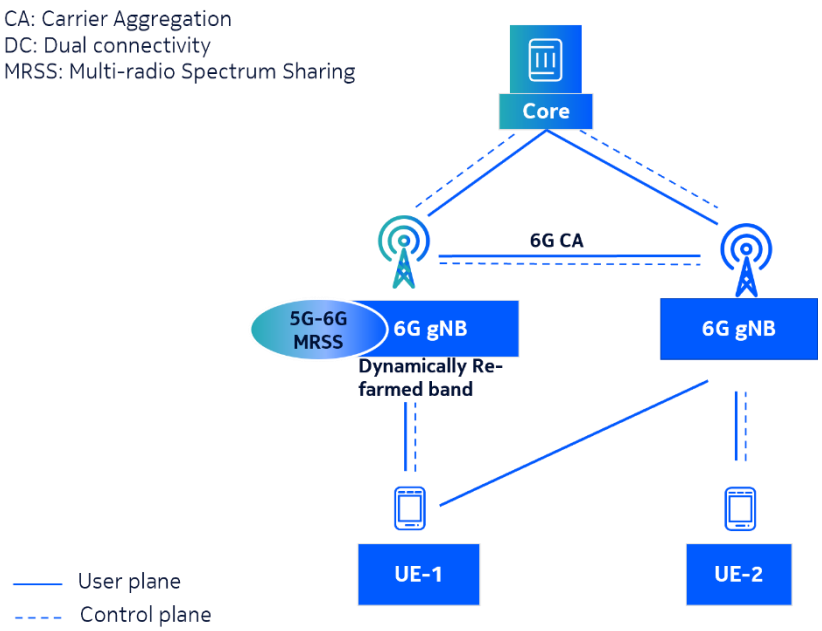

Figure 5-18 Migration to 6G system architecture with RAN coordination in 6G.

This study focuses on the migration options from 5G to 6G. More specifically, this work will aim to develop a simple, single step migration preferably based on SA. Research activities will outline how the capacity and the coverage can be increased dynamically. The further investigations will analyse if 6G core can be based on an evolution of 5G core and which 6G functionality would justify non-backwards compatible changes. The evaluation of different migration options is performed based on KPIs such as the number of migration steps, number of interfaces affected, impact to UE, impact to energy efficiency.

5.5.2 Evolved Core network and lower layer split

One question to investigate is if the migration from 5G to 6G should be performed in incremental steps, or more as an evolutionary development. This is depicted in Figure 5-19, where the core network may be more of an evolution than a new core network. The SBA introduced in 5GC decouples the network functions and allows a flexible introduction of new functionality and features, without having to change existing functions. This enables a gradual introduction of new 6G NFs (or modification of existing NFs) towards an evolved 5G CN.

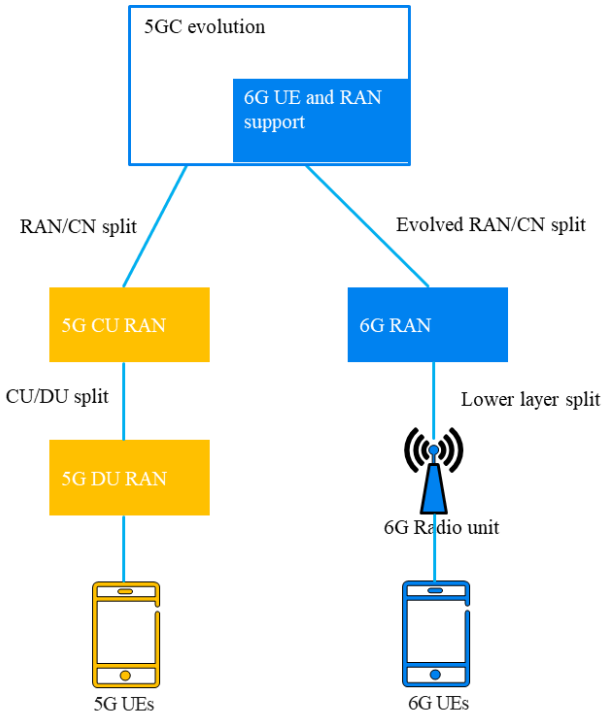

Figure 5-19 Possible 5G to 6G migration path for the Core network and LLS.

However, the 6G RAN/CN interface should be investigated so it performs well with a cloud environment. The use of a point-to-point interface between CN and RAN (based on SCTP) may not be optimal for cloud environments where NFs can be added, moved and removed dynamically and not be bound to a particular IP address. 5G NR introduced the higher layer split which cuts the network side's RRC function into two parts (with a multi-vendor interface between them) in the CU and DU respectively, see Figure 5-19. This complicated the decision logic and overall complexity in CU and DU substantially while leading to worse decisions, delayed actions and increased processing load. Therefore, this study will investigate if a new lower layer split can be used instead, with full control of the RAN node. This can be suitable for e.g., D-MIMO deployments.

6 Architectural enablers for new access and flexible topologies

This chapter identifies the architectural enablers in 6G that are related to new access and flexible topologies.

Section 6.1 describes the "network of networks" enabler, which includes studies on the architecture of subnetworks, on the architecture and coverage of Non-Terrestrial Networks (NTN), on the trustworthy and flexible topologies and their predicted coverage, as well as on the digital continuum. Subnetworks may enable the immersive telepresence use case family, which was described in Section 3.1, where devices of various capabilities and/or owners, coordinate with each other and with the overlay 6G network to achieve higher throughput and lower power consumption. Other potential use cases of trustworthy and unstructured networks, i.e., networks with flexible topologies, could involve cobots, precision farming services in remote areas, or emergency response and disaster relief applications. Trustworthy flexible topologies-related studies will also be involved in PoC #B.3, as described in Section 7.1.2. The use case of sustainable development explicitly addresses the Connectivity-Everywhere paradigm and hence understanding of large-scale coverage areas in the context of network architectures is needed. NTNs contribute to network sustainability in the sense that NTN deployment is considered for network coverage and capacity expansion and, thus, extended digital service provisioning, while optimization techniques to efficiently allocate the resources are employed.

Section 6.2 describes the "multi-connectivity" enabler, which includes investigations on the evolution of multi-connectivity from 5G to 6G, on an abstracted approach to multi-connectivity and on Terrestrial Network (TN) – NTN dual connectivity. The possibility to establish multiple paths for data transmission and reception among multiple co-located devices makes Multi-Connectivity (MC) an enabler of the telepresence use case family. Moreover, MC will be integral to subnetworks, which may themselves enable this use case. In addition, MC may enable the cobots use case family, due to the higher resilience that it provides.

Section 6.3 describes the "E2E context awareness management" enabler, which includes studies on context-aware transport, connectivity and RAN, on a tunnel-free user plane that supports mobility at both ends of a path, as well as on the delayed computing paradigm. E2E context awareness management may enable cobots, immersive telepresence use cases, as well as sustainable development, since the network will have to dynamically adapt to robots' context, to the user's sensory context and the application context. In addition, massive twinning may be enabled since the system behaviour may adapt to the environment of the specific use case.

6.1 Network of networks

Network of networks enables the integration of multiple subnetworks, including terrestrial and NTN in order to create a seamless and ubiquitous communication system, as illustrated in Figure 6-1. Terrestrial subnetworks may consist of multiple user nodes that are communicating with each other, with or without the aid of the network. In case of a network aided subnetwork, at least one user node in the subnetwork shall be connected to a network node. Other devices may also be known by the network node. This could be further enabled by capitalizing on sub-THz, which may be used for the links within the subnetwork, and D-MIMO networks (i.e., using L1 mobility) among others. NTNs comprise drones and satellites, which are traditionally used to offer communication services such as emergency management, navigation, and television broadcasting.

The proliferation of user devices in the network may entail a significant increase in the processing bottlenecks of multiple procedures, such as mobility, configuration, and scheduling. Hence, the necessity of creating subnetworks increases, to help in reducing those processing bottlenecks, while extending the network coverage. Those subnetworks would comprise of terrestrial and/or non-terrestrial nodes to which certain cellular procedures can be offloaded from the traditional network owned base stations and processes may become more efficient in a network of networks architecture. This may not only reduce the processing requirements in the main network nodes of the system, but it may also increase the system performance.

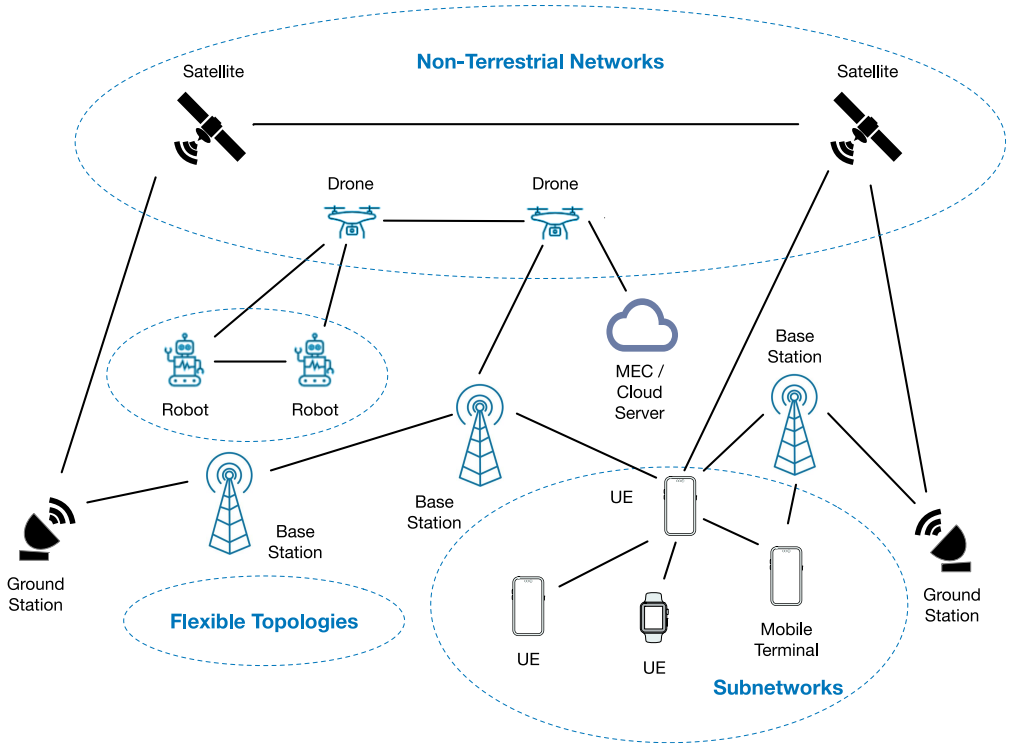

Figure 6-1 Network of networks with flexible topologies.

Recently, the existence of coverage holes or hotspot are of increased data traffic in conjunction with the advancements in mobile networks rendered the integration of NTNs as part of the broader terrestrial network infrastructure as a need for the provisioning of continuous and ubiquitous connectivity. NTNs can be deployed in an ad-hoc manner owing to their flexibility and reasonable cost, and act as either access nodes or relays to the core and cloud network, enabling the network to scale on demand and augment the performance of the underlying terrestrial network infrastructure. At the same time, NTNs would highly contribute to extending the network coverage, especially in underserved environments, while increasing the network reliability by providing an additional path to the core network. Concerning the integration of NTNs with the existing terrestrial network infrastructure, a multi-tier communication network emerges, spanning the ground, air, and space. Drones can play the role of access points or relays to the core network. Going one step further, drone-mounted edge servers can be used to realize in-network computing, resulting for a dual role for the NTNs. In this way, a multi-tier joint communications and computing network architecture is created.

One of the critical aspects related to the integration of multiple subnetworks (terrestrial, aerial, satellite, etc.) towards a ubiquitous communication system raises the challenge of trusting the diverse network nodes that will comprise such a network of networks structure, which in some cases may be opportunistically participating in network formations. It will thus be of utmost importance to define the novel functions and architectural components that will enable the timely assessment of the trustworthiness of various elements.

The architecture of flexible topologies, such as subnetworks formed by multiple user nodes and/or network nodes (terrestrial, non-terrestrial) will be studied. The control plane and user plane procedures may differ compared to the default cellular control plane and user plane. These procedures can be redesigned for the subnetwork architecture as motivated in the following subsections, to address signaling efficiency, as well as trust and sustainability aspects.

The network of networks enabler can contribute to the goals of 6G networks, such as extreme coverage, reduced complexity, increased reliability, and more efficient management of network resources. This enabler fulfils the 6G design principle #3 of [HEX2-D21], which aims to increase the flexibility to different network scenarios. The following paragraphs summarize the study areas on network of networks that this task will focus on:

Section 6.1.1 focuses on the architecture of the subnetworks, the capabilities of each node and the new roles that the user and network nodes may assume. Based on the role of each node in the subnetwork, new procedures may have to be defined and the functionalities that each role should support shall be investigated.

Section 6.1.2 investigates architectures with efficient inter-satellite-link hops to enable true global service coverage.

Section 6.1.3 focuses on the design of unified decision-making and resource allocation frameworks for the subnetworks, managing concurrently multiple types of communication and computing resources. Distributed decision-making and resource allocation procedures are designed, based on which network owned nodes steer the corresponding procedures via appropriate signalling, while distributed decisions are autonomously taken by different user owned nodes in the network.

Section 6.1.4 describes the need for data-driven ML tools that predict future large-scale coverage developments and include proposed solutions in the prediction models. History has taught that mere availability of technology or a standard does not guarantee coverage everywhere and it is therefore relevant to understand and model the potential of new network architecture solutions in different regions with various and different characteristics. The developed models predict whether remote regions will be covered or not when new network architectures/new technologies become available to operators. These tools will incorporate on one hand technological possibilities and, on the other, data-driven historical deployment developments.

Section 6.1.5 introduces the motivation and proposed way forward on developing novel architectural enablers for AI/ML-driven assessment of various flexible network formations, depending on the traffic requirements, as well as trust and cost optimization aspects. This will require specifying the way that node information and capabilities are exposed to the network, along with the novel network functions that will assess this information and provide insights for flexible and/or unstructured network formations.

6.1.1 Subnetworks: Architecture, new roles and responsibilities of the nodes

There is currently a rapid growth in the number of connected devices, which is expected to continue through the deployment of 6G. Traditional networks may not be able to efficiently handle this increasing number and diversity of devices and applications. Additionally, 6G will introduce requirements for increased coverage, lower power consumption, higher data rates, increased resilience, and increased trustworthiness / user privacy, compared to 5G. Subnetworks may aid in achieving these KPIs/KVIs by offloading functionalities from one node to another, by providing connectivity to devices that are not in network coverage and by managing the radio resources more efficiently based on information shared by the nodes. This would mean that the subnetworks concept would further enable the operation of devices of varying capabilities (e.g., compute power, battery, cellular, etc.). Subnetwork nodes would be comprised of user and network nodes. Moreover, subnetworks are designed to allow for more flexible and efficient communication between trusted devices. Their architecture can differ based on various criteria, such as device type, application type or geographical location of the nodes. The adoption of a distributed architecture would allow the Base Station (BS) to offload certain functionalities to the subnetworks. The specific set of functions that user and network nodes currently have may be insufficient and/or inefficient in the context of subnetworks. Therefore, new node roles and responsibilities in a subnetwork should be investigated, targeting a lower complexity for low-capability nodes (e.g., compute power, battery, cellular capabilities). Moreover, close coordination between nodes is required, to ensure efficient communication. Trusted nodes may need to securely exchange information about the devices' status and their applications to aid with the management of the subnetwork. Without effective coordination, communication may be inefficient, leading to increased latency, suboptimal radio resource management decisions, as well as link failures.

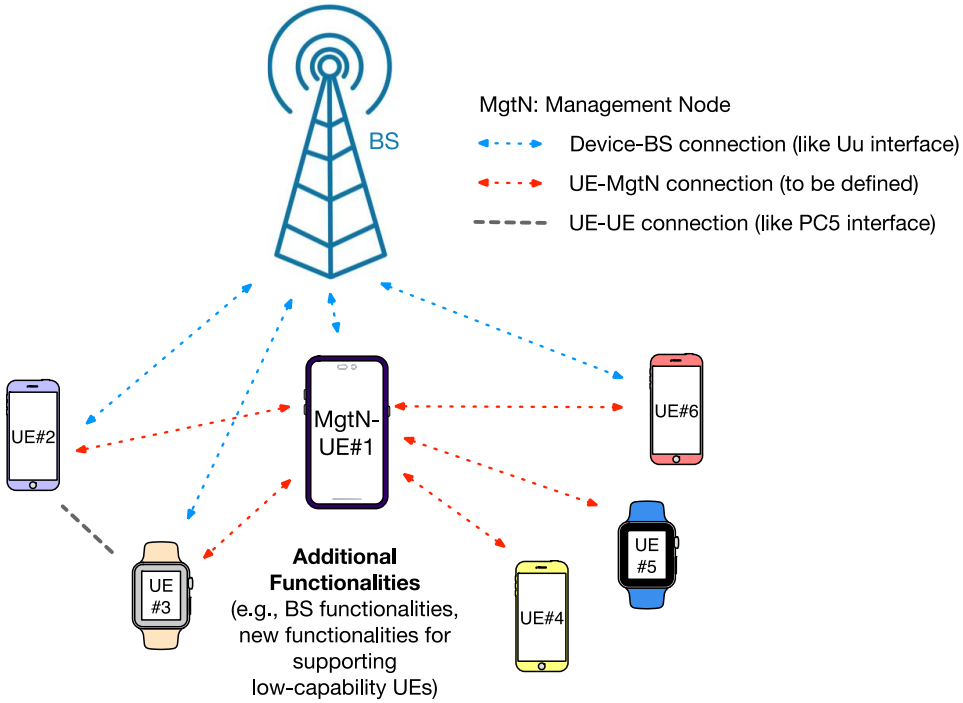

Figure 6-2 Subnetwork with a Management Node (MgtN) connected to five UEs and to the 6G base station. UE4 and UE5 are out-of-coverage of the BS, while UE2 and UE3 can also communicate directly with each other.

The Management Node (MgtN) is foreseen as a new UE role in a subnetwork. The MgtN is the primary node of the subnetwork, which can communicate with the BS and other UEs. The rest of the UEs in the subnetwork may be nodes with normal or reduced cellular capabilities. Figure 6-2 illustrates a subnetwork with a MgtN and multiple UEs. The MgtN may have a modified Control Plane (CP), which would include extra functionalities for supporting the control plane of the UEs in the subnetwork. At the same time, the rest of the UEs may have a leaner version of the CP without the full set of functionalities.

Various subnetwork architectures should be considered, based on the connections between the BS and the UEs. Based on the selected subnetwork architecture(s), the investigation will focus on how the MgtN may assist in reducing the complexity of various control plane (CP) and user plane (UP) procedures for the local devices. The list of procedures to be investigated may include:

- RRC procedures, such as offloaded RRC configuration and connection establishment
- Radio Resource Management (RRM) procedures, such as centralized mobility management and shared measurements
- Idle mode procedures, such as offloaded paging and tracking area update procedures
- Connected mode procedures, such as data flow management (UE-to-NW relay, UE-to-UE relay)

6.1.2 NTN architecture and global coverage

Hexa-X-II have an objective to find architectural solutions that enhances the global coverage. In Hexa-X, initial NTN simulations of the global coverage were provided. The simulation in [HEX-D53] mimicked established satellite operators' plans using the same number of satellites, orbits, inclination, altitudes etc. In addition, the settings for propagation, frequency, transmission power etc. are taken from 3GPP TR 38.821 [38.821].

The architecture needs to support efficient inter-satellite-link (ISL) hops to enable almost 100% global service coverage, i.e., a user in the ocean can also reach the ground stations and terrestrial network on the ground. This study will investigate different architecture options to enable an efficient ISL and connection to the terrestrial network. There is a need to investigate the architecture of NTN, i.e., whether transparent or regenerative architecture should be used. Transparent is when the Satellite serves as a relay of the signal between the UE and the base station on ground and regenerative payload is equivalent to have the base station onboard the

satellite. For both cases there is a need for a gateway on the surface to connect to the terrestrial network. In [HEX-D53], an NTN simulation was performed to evaluate the global service coverage assuming perfect ISL connection and regenerative payload see Figure 6-3 [HEX-D53].

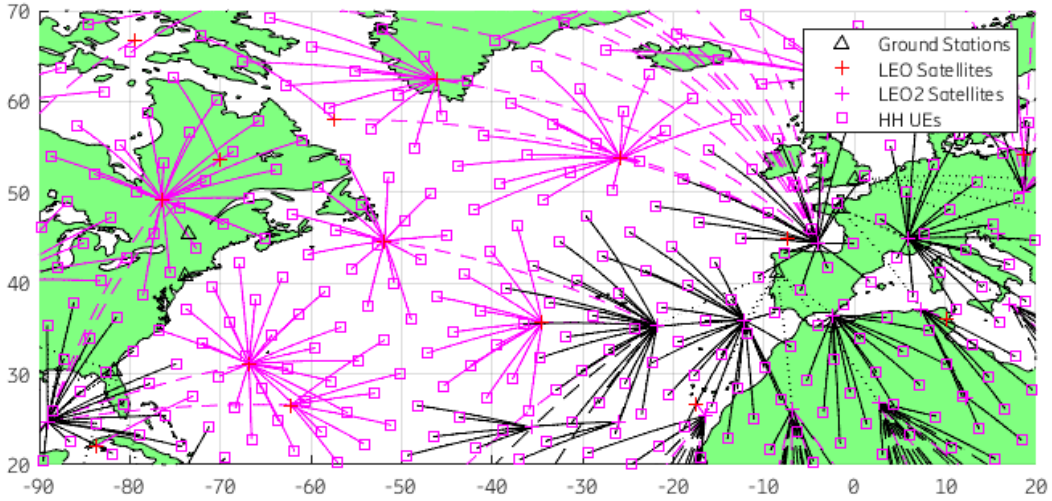

Figure 6-3 [HEX-D53] evaluated the NTN coverage for handheld (HH) devices over the Atlantic Ocean with coast areas.

The NTN simulator used in [HEX-D53] first computes the position of the satellites in terms of latitude, longitude and altitude of each satellite in the constellation around the globe over time. The users can be deployed in the region of interest. The ground stations are placed on the coastal areas of North America and west Europe and north-west Africa, see Figure 6-3. The solid lines in Figure 6-3 are UE to Satellite connection (service link), the dashed lines ISL, and the dotted lines are Satellite to ground station (feeder links). Magenta and black indicates the two different satellite orbit and altitude constellations used in the evaluations. Thereafter the simulator calculates the best UE to satellite connections and the pathloss and SNR of the connection. The radio and channel simulation parameters used in the simulation are based on [38.821]. The final step, is to evaluate the throughput per user per satellite, based on the available spectrum per cell and beam.

The results from [HEX-D53] showed that coverage over the Atlantic Ocean for a very low density of users is possible assuming a certain number of satellites and the use of ISL hops. However, the evaluations used in [HEX-D53] were rather rudimentary. To understand if the enhanced global coverage is possible, the next step is therefore to continue evaluating the same scenario but using more realistic models and assumptions, such as interference between overlapping beams, scheduling, ISL capacity and architectural aspects.

6.1.3 Digital continuum: Architecture design and decision-making

NTN infrastructure is an essential building block of 6G networks to support high-traffic and emergency communications in disaster-struck areas where persistent and robust information flow is needed. Especially when considering drones and Unmanned Aerial Vehicles (UAVs), features such as mobility, deployment flexibility, low cost, and strong Line-of-Sight (LoS) links enhance the development of NTNs and their utilization beyond pure communications to facilitate Multi-access Edge Computing (MEC) [ASC+22]. UAVs can serve as UAV-mounted servers allowing in-network computing. Therefore, the role of a UAV within the digital continuum is twofold since it can be considered as either a relay to provide connectivity to the core network and offload computation tasks to the cloud or as a standalone computing entity at the network edge. An overview of the resulting digital continuum is presented in Figure 6-4.

Introducing UAVs provides extra degrees of freedom to the communication and computing resource allocation problem, increasing at the same time the network orchestration complexity. For example, bandwidth splitting,

subchannel allocation, and transmission power control at the access and backhaul network parts [DCT+22a], computation offloading decision-making and computing power allocation at the different computing tiers [WHM+22] constitute only some of the various optimization problems to be concurrently considered to reap the benefits across the digital continuum.

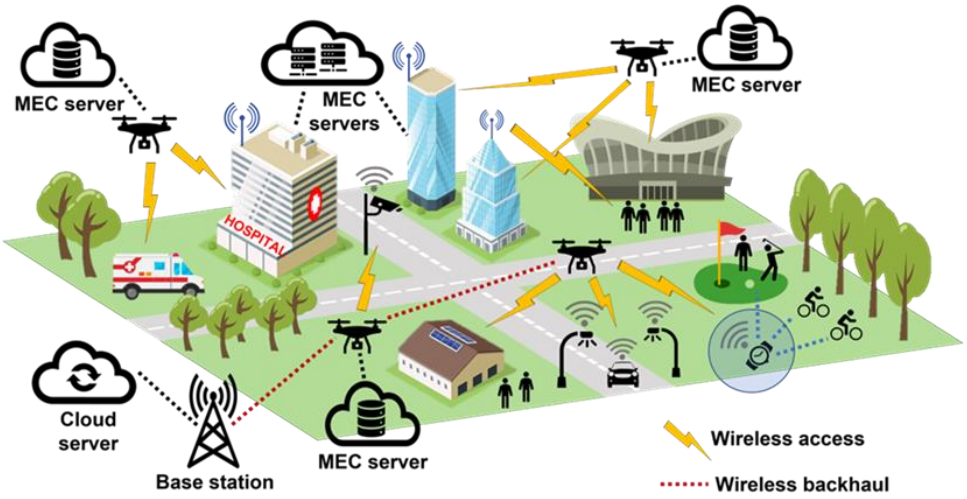

Figure 6-4 Overview of the digital continuum architecture.

In this study, the digital continuum will be considered, comprising multiple UAVs that serve as relays or edge servers. The goal is to design an appropriate framework for the joint optimization of wireless communication-related resources and computation offloading decision-making, targeting end-to-end network throughput maximization and service delay minimization. To reduce the algorithmic complexity of centralized solutions, a hierarchical optimization problem will be formulated based on the principles of game theory and Stackelberg games [DCT+22a]. The UAVs having a holistic view of the congestion in the access and backhaul network parts will determine the optimal wireless resource allocation across the digital continuum, and the users, based on this information, will decide on their computation offloading. This results in an iterative process that concludes at a Stackelberg equilibrium point where the most beneficial resource allocation and computation offloading decision is reached for the network and the user sides.

6.1.4 Large-scale coverage prediction for flexible topologies

A vast number of people are still deprived of means of connectivity, and it is no surprise that in rural and remote regions, characterized by low population densities, large distances to urban clusters and to societal service, such deprivation is most prominent. Despite many network-topological deployment solutions, operators do not seem to achieve global coverage. For 6G to provide coverage everywhere, a deeper understanding around the respective roles of technology and other societal parameters, such as population density, is needed. Figure 6-5 illustrates the presence of multiple available networks in a region.

A new cellular standard or a new technology will not automatically guarantee that rural or remote regions will be provided with coverage. Operators will have to actively choose to deploy networks based on these new standards and technologies. Whether or not an operator chooses to deploy in a region depends on other characteristics of that region and not merely on the technological merits of the standard or technology.

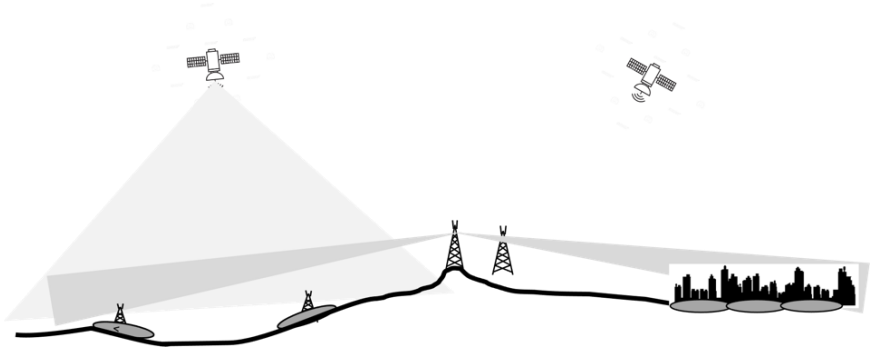

Figure 6-5 Various network topologies by network operators (providing coverage by, for instance, terrestrial medium and large-size cells, high-tower ultra-large size cells, NTNs, and potentially combinations thereof) are deployed in a large-scale region.

This study develops prediction tools that not only capture the technological characteristics of a future 6G standard, its possibilities and potential, but also historical societal characteristics of network coverage related to both population densities and presence of other infrastructure such as roads, railways, etc. It combines models that capture characteristics of radio-technology and architectural topologies with data-driven geographical and societal parameters. Based on historical cellular deployments and the associated population density maps, the areas where future deployments are likely to occur will be inferred. The study anchors in the assumption that the above societal parameters carry a lot of predictive power, independently and aside from the technological characteristics, and it develops learning tools where both dimensions are used to develop a suitable ML model. Provision of new standardisation or new technology and network architectures by itself is not sufficient to render large scale coverage. The mere availability of new technologies that provide coverage by drones, by Low Earth Orbit (LEO)-satellites or by high-tower ultra large cells is no guarantee. Operators have a choice to either deploy networks with these new technologies or not. Our coverage models combine the technological potential of new architectures with the historical evidence of operators choosing to deploy networks or not. On one hand the physical coverage capabilities and limits of available technologies will be available to train the ML model. On the other hand, data of population and road maps along with the historical development of the cellular coverage will be available to train the prediction model.

6.1.5 Trustworthy, flexible, unstructured networks

The digital era has brought with it an exponential growth in the number of devices and users, leading to a significant increase in data volume creation. This surge has created new communication needs and challenges that extend beyond human-centric communications and incorporate various applications such as extended reality (XR) and virtual reality (VR), massive twinning, and JCAS for the upcoming 6G network generation [JHH+21]. However, traditional mobile network operator (MNO) infrastructure often struggles with these evolving demands, particularly in handling traffic bursts or providing adequate coverage in remote or underserved areas. Therefore, to cater to these emerging requirements, it is essential to explore alternative solutions and develop flexible, trustworthy, low-cost networks that offer seamless communication and computation everywhere [BLG+23, 5GP22, LQW+20]. This study provides a high-level overview of a proposed solution that aims to address these challenges by creating flexible topology networks.

In cases of resource/coverage constraints, there is a need for a network solution that can quickly adapt and provide robust communication and computation capabilities. Flexible topology networks aim to address these challenges by developing an adaptable and efficient network structure [YTA+20], considering the identification of nodes, their cost, capabilities, resources, and trustworthiness, as well as the optimal formulation of the topology and service orchestration.

To achieve these objectives, the proposed solution (see Figure 6-6) utilizes a combination of components and functions, such as node discovery, trust/cost evaluation, and AI/ML-driven resource optimization. Firstly, the node discovery component gathers information about the nodes that will be utilized by the other components. Next, the trust manager and trust evaluation function components work together by gathering information from

both the trustworthiness assessment inputs component and AI/ML resource optimization component to assess the trustworthiness of nodes. This information is then utilized by the flexible mesh node selection function, which is the responsible entity that makes informed decisions on the optimal network structure. Note that the AI/ML resource optimization component interacts with both the flexible mesh node selection function and the trust manager/trust evaluation function components to optimize resources based on the available node information and trust assessments. The proposed solution aims to optimize 6G KPIs and KVIs such as sustainability, cost-effectiveness, energy consumption, inclusion, and system trust.

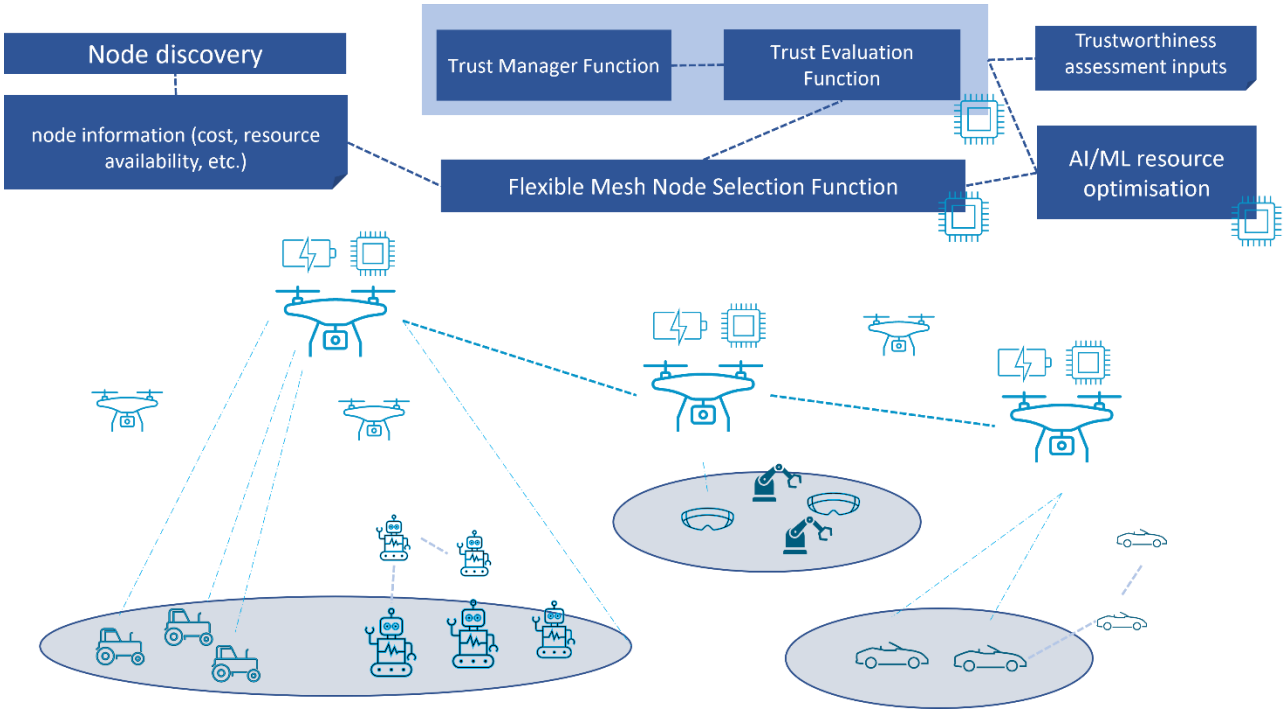

Figure 6-6 Overview of the trustworthy, flexible, unstructured networks concept.

As a key feature of 6G, the proposed flexible, unstructured networks are designed to adapt efficiently during periods of heightened traffic and limited resources. Besides terrestrial nodes, they can involve aerial ones, such as Unmanned Aerial Vehicles (UAVs) to provide the necessary relaying capabilities, along with sufficient bandwidth and processing power for extreme use cases. Responding to varying connectivity and computational needs across diverse scenarios, the solution places emphasis on cost-effective infrastructure, reduced energy usage, trust, and sustainable practices. This high-level overview provides a valuable tool for decision-makers to evaluate the feasibility and cost-effectiveness of applying such flexible network formations versus existing MNO static infrastructure in similar scenarios.

6.2 Multi-connectivity

Multi-connectivity (MC) enables the configuration and use of multiple frequency ranges by different physically co-located user and network nodes as well as the aggregation of different radio access networks and carrier frequencies. Moreover, multi-connectivity would enable the communication with both terrestrial and non-terrestrial nodes, as well as the integration of subnetworks to the parent network, as illustrated in Figure 6-7.

In 5G, carrier aggregation (CA) and dual connectivity (DC) have been adopted. Within a cell group, CA relies on the configuration of multiple component carriers (CC), which can increase the system throughput when activated. The CCs may be contiguous in the same frequency band and non-contiguous in the same or different bands. Even though CCs of both frequency range 1 (FR1) and frequency range 2 (FR2) are allowed to be configured in the same cell group, in practice only CCs of the same FR are encountered. The MCG has a primacy cell (PCell), on which some specific user procedures are performed, such as radio link monitoring and random access. Dual connectivity introduces an additional, secondary cell group (SCG), which may or may

not have carrier aggregation itself. The primary cell of the secondary cell group (PSCell) is responsible for receiving the configuration of the SCG and for performing the additional monitoring procedures of the SCG. Dual connectivity aggregation involves a radio bearer split or duplication at the PDCP layer, which means that the packets sent via the two cell groups are aggregated at the PDCP layer. Dual connectivity also supports MCG radio bearers and SCG radio bearers, which handle different traffic without routing the data packets via the other cell group. The CCs of the secondary cell group may lie in the same or different FRs than the CCs of the MCG. In practice, deployments with the MCG having FR1 CCs and SCG having FR2 CCs are encountered in NR SA architectures, where the MCG is used as an anchor for connection to the network.

In 4G, multiple solutions for aggregating different RATs, such as WLAN, to the cellular network were included. For example, LTE-WLAN aggregation splits and converges the WLAN and cellular paths in RAN. With the introduction of non-terrestrial networks (NTN) and subnetworks, it would be beneficial to investigate how the different access networks could be aggregated and integrated with the cellular network.

Contributions to 6G networks goals include extreme coverage, increased reliability, and increased flexibility. The multi-connectivity enabler fulfils 6G design principles #5 (Resilience and availability) and #3 (Flexibility to different network scenarios). Different types of multi-connectivity, such as cellular and subnetworks, cellular and different radio access technologies, dual connectivity to terrestrial and non-terrestrial nodes will be investigated.

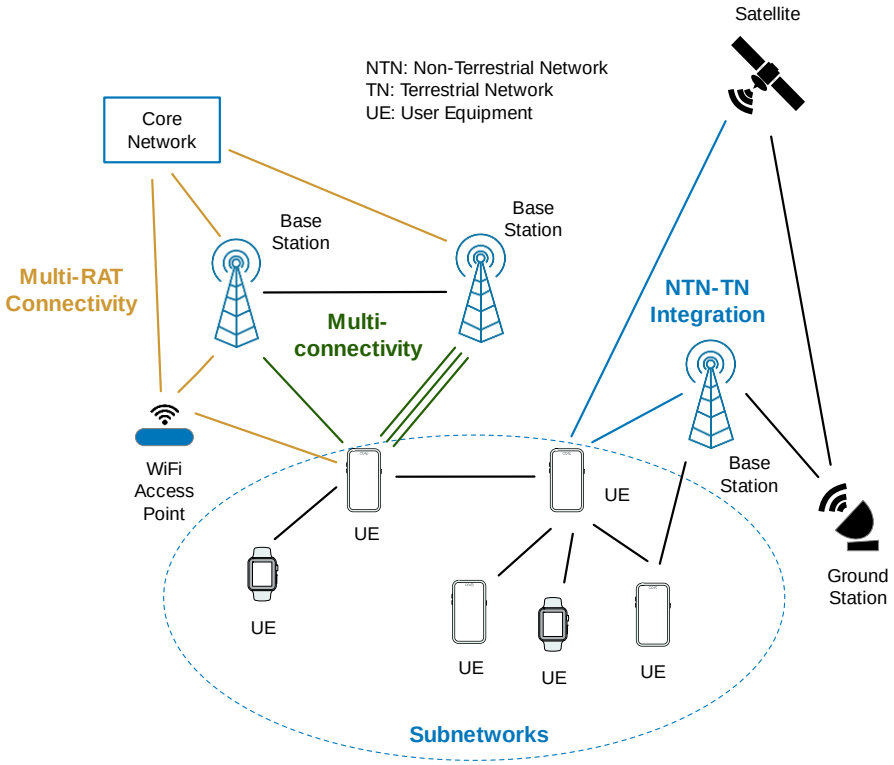

Figure 6-7 Different types of multi-connectivity, including TN-TN and TN-NTN dual connectivity, multi-connectivity within terrestrial subnetworks as well as with different radio access networks.

Section 6.2.1 explains the creation of subnetworks, which may include both user and network nodes, and their inherent need for multi-connectivity. In addition, the procedures that correspond to the newly defined roles of the user and network nodes should be investigated. At the same time, enhancements to the 5G dual connectivity and carrier aggregation mechanisms should be explored, as motivated in **Section 6.2.1** and **Section 6.2.2**. **Section 6.2.2** then focuses on designing a new multi-connectivity solution combining the positive aspects of the current DC and CA solutions, while **Section 6.2.3** investigates the DC between terrestrial and non-terrestrial nodes.

The aggregation of different radio access networks on different levels (e.g., CN, RAN) and in a connectivity domain-abstracted manner is motivated in **Section 6.2.1** and **Section 6.2.4**. The previously proposed solutions in LTE were not implemented in practice. In order to avoid this from happening, the investigated solutions that

integrate multiple access networks to the cellular network should be streamlined with the current evolution of 5G and should take forward compatibility into consideration.

6.2.1 Multi-connectivity for different technologies

Multi-connectivity can achieve improved reliability, coverage, and satisfy the demand for increased data rates and lower latency in various applications. The two cell groups may belong to the same or different Radio Access Technologies (RAT). MC between cellular and different RATs can improve network coverage, increase data rates, and enhance the user experience. However, there are challenges in integrating these heterogeneous networks, such as coordination between the involved nodes, resource allocation and handover management.

With the introduction of subnetworks, UEs that belong to a subnetwork may connect to multiple user and network nodes to further increase the network coverage and resilience or to enable new use cases. A MgtN may connect to multiple BSs as well as to other MgtNs. Trusted UEs in a subnetwork may use the connections within the subnetwork not only for transmitting data, but also for exchanging information available at other nodes, which would be leveraged for improving the decisions that the nodes shall make. In addition, it may enhance the network coverage since it may provide network access to an out-of-cellular-coverage UE.

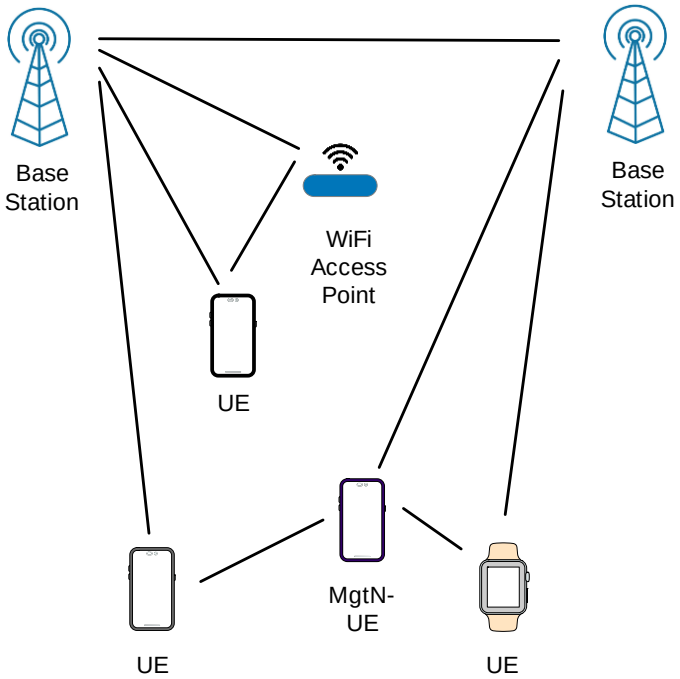

Figure 6-8 Dual connectivity, WLAN - cellular aggregation and multi-connectivity within a subnetwork.

Even though 5G DC supports split radio bearers, where data is transmitted over different paths and then aggregated at the receive PDCP entity, the aggregation of different RATs, such as WLAN, on the RAN level is not supported in 5G RAN. Aggregation of other RATs will be studied. This aggregation could result in increased reliability, higher data rates and seamless handovers, especially in indoor or out-of-cellular-coverage environments. Figure 6-8 depicts the multi-connectivity concept in terms of dual connectivity and WLAN-cellular aggregation.

Currently, the transition from RRC Idle mode to RRC Connected mode with DC suffers from high latency or high-power consumption. The UE would either commence measurements for the secondary cell group only after it transitions to Connected mode (i.e., high latency), or it would have to commence measurements while in Idle mode, which would be useful only if the UE eventually transitions to Connected mode (i.e., high power consumption with potentially no reward). Alternative mechanisms focusing on lower latency and power consumption will be studied. The Inactive mode has been introduced in order to reduce the latency required to transition to Connected mode in comparison to Idle to Connected mode transition latency. It comes with additional requirements of storing UE-related information in RAN and CN and comes with power and

signalling overhead for RNA updates compared to Idle mode. These may be some reasons why the Inactive mode is currently not widely used but it should still be taken into consideration for comparison purposes.

The MC concept may also be extended to multiple devices. As mentioned above, the 5G DC/CA concept refers to a single device connected to primary and secondary cells. However, MC in 6G may introduce the concept of primary and secondary devices. It may allow each cell group to serve a different device, where the primary and secondary devices are connected to each other and, in addition, the primary device (e.g., MgtN in subnetworks) is responsible for maintaining the connection of the secondary device.

6.2.2 6G multi-connectivity proposal

One of the main drawbacks of dual connectivity between 4G and 5G was the number of options specified in 3GPP, i.e., using either EPC (4G CN) or 5G CN and using different options where the connection was terminated. 3GPP also specified (normal) CA and DC within the 5G bands. These options increased the complexity of the specifications and increased the number of test cases to avoid multi-vendor interoperability issues to the expense of supporting fewer options.

Another drawback with DC (and with EN-DC) was that the master node (where the connection is terminated) will not have the most recent information about the secondary node performance since the backhaul (Xn) connection between the nodes may be too slow compared to the time scales of the variations in the radio channel. The communication protocol (i.e., the flow control, see [38.420], [38.300]) between the master and secondary node estimates the throughput based on the acknowledgements it receives from the secondary node. In some cases, if for example the secondary node is a cell with high frequency, the coverage may drop quickly and cause long packet delays for the connection over the secondary node. The master node may be unaware of the drastically decreased performance and still send data to the secondary node over the Xn.

Another feature of DC is that DL and UL are always coupled. This means that all connections in the UL shall be able to send acknowledgements of the Hybrid Automatic Repeat reQuest (HARQ) or Radio Link Control (RLC) packets. This can in some cases be beneficial, for example, if one of the connections fails, the remaining connection can keep the user from entering Radio Link Failure (RLF). However, since the secondary connection may have worse UL coverage than the master (the difference may very well be of several dBs, depending on the frequency range), the secondary node feedback may become so bad that this may cause a sharp increase in the round-trip times (or even a timeout), and this will cause a decrease in the TCP/IP connection throughput.

In CA, there is no need for a flow control as long as the backhaul is very fast and has low latency. Assuming low latency of the backhaul, a centralized scheduler of the Primary Cell (PCell) can therefore be used to schedule the data over both nodes. The DL and UL connections are not coupled in CA, the best UL (i.e., the PCell) can be used for UL response, which means that the UL coverage is often better for CA compared to DC, as the UE does not have to split its limited uplink transmit power between two concurrent UL.

Therefore, in order to improve the MC solution for 6G, as well as simplify the solution by reducing the number of architecture options, one option is to only allow MC between 6G enabled base stations and using one type of solution only. In [HEX-D53] a new 6G multi-connectivity solution was proposed, see Figure 6-9 for a high-level view.

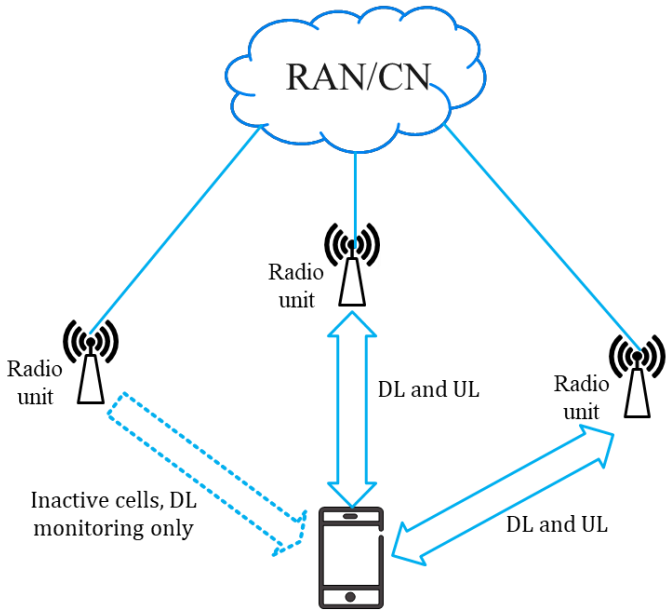

Figure 6-9 Proposed 6G multi-connectivity solutions overview.

The solution proposed combines the best features from CA and DC in order to provide both extreme reliability and excellent flexibility. The new solution aims to decouple DL and UL (e.g., two DL connections and one UL connection, see Figure 6-9) and inherent use of in-active connections. For the in-active connections, the UE only need to sparsely monitor the control signaling from the network, and the in-active connection should be able to be activated on a short notice. In the coming deliverables we will evaluate the need and benefits for the MC solution to understand in more detail how the new 6G MC solution should be designed. To analyse the type of solution that is required, simulations using different types of frequency bands will be evaluated, see Table 6-1.

Table 6-1 Overview of frequency bands for possible multi-connectivity evaluations

Frequency band	**Comment**
Sub-Terahertz (90…200 GHz)	Complimentary spectrum for extreme performance in very local areas
Millimeter wave (24…43 GHz)	High-speed, very low latency in local areas
Centimetric (7…15 GHz)	Potential new spectrum for 6G, good coverage and capacity
Mid-band TDD (2.6...<7 GHz)	Wide area coverage and good capacity
Low-band FDD (<2.6 GHz)	Nationwide coverage and deep indoor penetration

A promising combination to investigate is to use the Frequency Division Duplex (FDD) low-band together with the Time Division Duplex (TDD) mid-band. The main reason is that these bands do not differ very much in coverage and may therefore be co-located. Another interesting aspect to investigate is the benefits to use co-located nodes (i.e., using same site) for both bands and if it is more beneficial to use different node locations. The KPIs to investigate are user throughput and spectrum utilization.

6.2.3 NTN-TN integration and global coverage

Providing coverage and service continuity in “not-spots” areas, where it is commercially not viable to provide mobile coverage, is always challenging for mobile operators. NTN global coverage could be useful for users in “not-spot” areas. Figure 6-10 Coverage holes covered by NTN using dual connectivity between NTN and TN framework shows an NTN cell covering multiple TN cells and an area under the coverage of NTN cell where there is no TN coverage. A typical user may intend to stick with TN coverage as much as possible and select NTN network only in those coverage holes due to various reasons like e.g., user experience, latency,

billing etc. 3GPP Rel-18 [RP-223534] is already working on NTN-TN cell reselection enhancements in IDLE/INACTIVE modes. However, for NTN-TN connected mode mobility, NTN and TN operations might be controlled by different entities/mobile operators. 6G should provide seamless coverage and service interruption should be avoided during these transitions/switches between TN and NTN. The KPIs are service coverage, service continuity and lossless communication.

Dual Connectivity between NTN-TN could be the starting point but the main difference from dual connectivity is that traditionally in DC, a large cell takes over the role of Master Node (MN) and RRC procedures like handover are anchored in the MN node so that any mobility between small cells or Secondary Nodes (SN) does not involve the Core network and trigger handover like procedures. However, a typical LEO satellite constellation has multiple satellites covering a location on earth over a period of time and a handover takes place, for a UE, every few seconds due to movement of satellites in their orbits, even if the UE remains stationary. So, although feasible, it may be challenging, from the UE's power consumption point of view, to configure an NTN cell as the MN. If the TN cell acts as the MN node, then the NTN node could be configured and activated/deactivated as a Secondary Node (SN) even if the UE is out of the MN's coverage. An interface is required between the NTN earth station hosting base station function and the TN base station in order for DC to work. So, DC could be a starting point to provide coverage in "not-spots" and configuration & activation/deactivation of NTN as MN/SN could be enhanced further such that the service interruption is minimal.

Another challenge would be to meet the latency requirements of a service, especially in an NTN environment. Latency in NTN can be split into three parts: first being between the UE and satellite, second between the satellite and earth station or gateway (latency varies for bentpipe/transparent or regenerative satellite), a third between earth station and the application host. Normally, a satellite constellation connects to earth stations located at various locations spread over different continents and any connection to the application server/host can take place from these earth stations only. If earth stations are deployed sparsely and/or if application host is physically far away from the earth station, then second and/or third part of latency in NTN networks will increase. Whereas TN network may offer comparatively reduced latency for a service by offering breakout at geographically convenient places / points. It may be difficult to match the latency of TN networks, but some improvements should be possible for NTN network latency reduction in terms of interconnection between earth stations and application servers by deploying edge server closer to earth stations and offering breakouts at convenient places / points. Therefore, 6G may enhance procedures for NTN-TN DC configuration and improve the latency requirements.

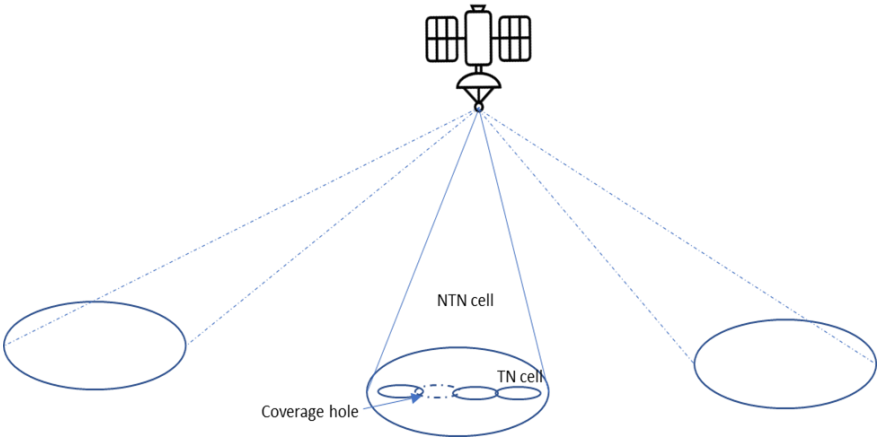

Figure 6-10 Coverage holes covered by NTN using dual connectivity between NTN and TN framework.

6.2.4 Abstracted approach to multi-connectivity

The multi-connectivity using different networking technologies is so far addressed only in some specific cases. For example, WiFi access can be added to LTE or 5G NR. The 3GPP specifications define the procedures required for such integration, mostly seen as static [23.501]. In 6G a new connectivity domain shall be added

dynamically, and such networks can be very different. For example, NTN, in-car or body area networks shall be smoothly integrated. According to the network-of-networks principle, different networking domains should be integrated, and fulfilling SDG, also legacy networks should be integrated. Integration of multiple networking solutions may also improve the network performance by simultaneous forwarding of traffic over several virtual links, or it can be used to enhance connectivity reliability.

The first way of such integration lies on the definition of protocols that are capable of handling all envisioned use cases. Such an approach is complicated and generates significant signalling overhead; however, it allows for the exchange of information from lower layers, which can be useful for making decisions regarding connectivity.

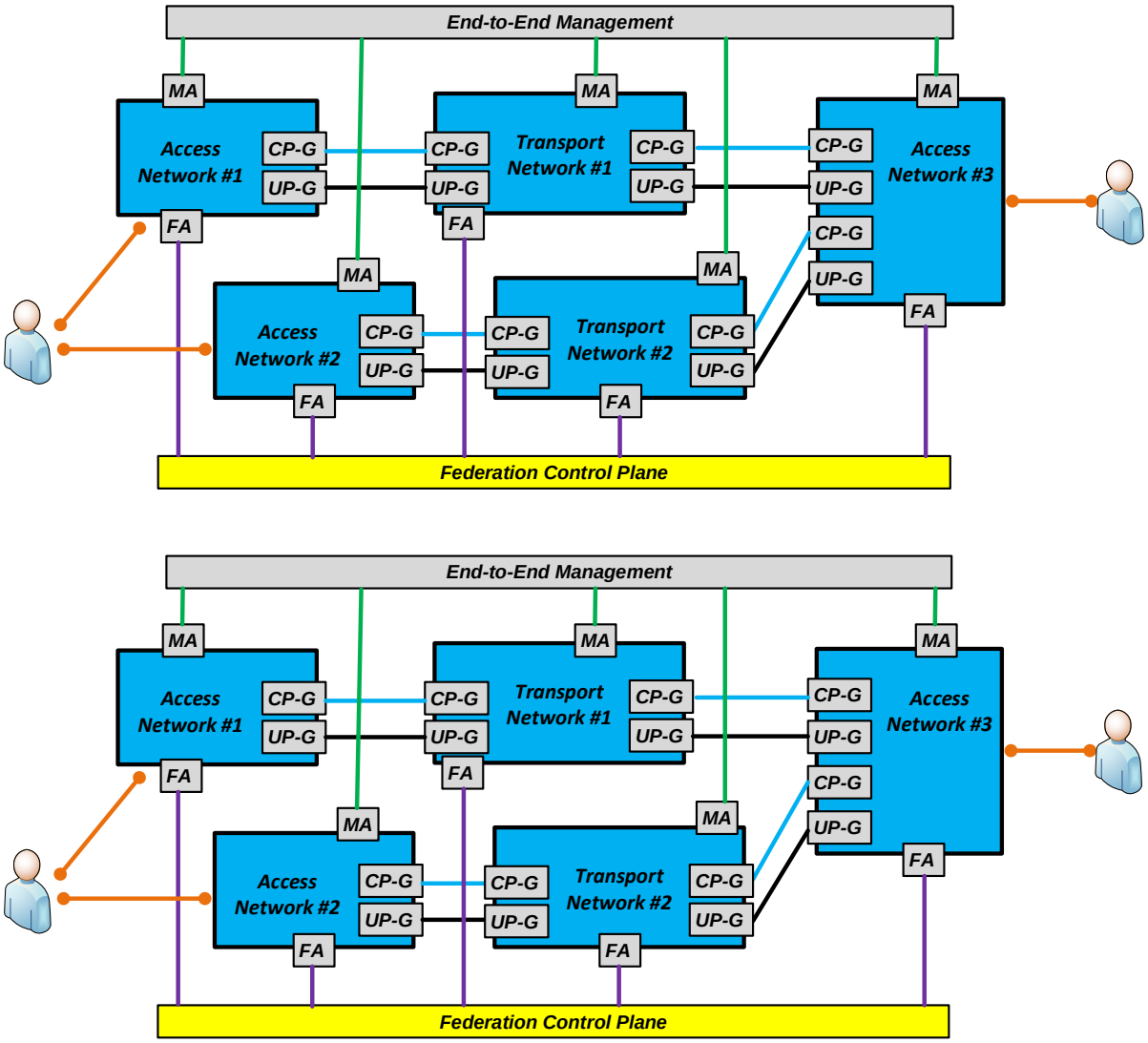

Figure 6-11 An example of interconnection of multiple connectivity domains.

The second approach is an integration of connectivity domains using generic abstractions. This approach requires the definition of functions that can be used for the federation of connectivity domains. To that end, a Federation Control Plane is proposed, as illustrated in Figure 6-11. Such signalling can be used to discover a new connectivity domain to be federated and the orchestration of procedures needed for the interconnection of specified domains. The integration of connectivity domains includes the definition of appropriate control plane and user plane gateways between two interconnected domains, i.e., CP-G and UP-G. The use of high-level signalling, in the form of intents, seems to be an optimal solution for the control plane. Each interconnected domain should, to a certain extent, be self-managed and expose vital domain parameters (KPIs) via a Management Agent (MA) to the End-to-End Management plane. A Federation Control Plane manages the whole federation interacting with Federation Agents (FAs) of each domain and may be responsible for traffic forwarding (load balancing) rules. Traffic load balancing can also be implemented on the source and

destination sides as an application. To improve resiliency, the federation request may include disjoint created path (node level, link level).

The proposed approach allows for integration of different domain in a relatively simple way. It may reduce the overall signalling, but it requires the definition of CP and UP gateways that can be a non-trivial task.

6.3 E2E context awareness management

The E2E context awareness management defines several mechanisms to allow each component of the network to dynamically adapt to the context (e.g., user requirements, layer specific characteristics) in order to ensure the expected end-to-end QoS for the services and the expected QoE for the users (humans or objects). The network components may include the RAN, transport network, core network, application, edge computing, etc.. The E2E context awareness management should guarantee an effective and optimized use of the network infrastructure resources. This management leverages on effective automation and orchestration mechanisms to facilitate the interaction among such components. Figure 6-12 illustrates the high-level scenario constituted by a "user layer", on the top, served by edge computing/cloud systems. A global connectivity/network infrastructure layer conveys the user traffic across radio and transport.

E2E context awareness management allows for the underlying network and edge computing layer's infrastructure dynamic adaptation as a means of (i) providing customized services to the users and (ii) extending its service capacity by optimally orchestrating its resources. This generates a diversity of user communication and computing requests in terms of delay or other KPIs that can be concurrently met. Regarding the former, additional parameters such as the users' service subscription payment availability can be considered as part of their QoE, and different communication and computing service options can be offered to them, comprising the targeted service quality and their subscription fee. This also allows for personalized and dynamic resource allocation, which in contrast with uniformly/equally allocating resources to the end users, is more efficient in terms of resource utilization and increased service capacity.

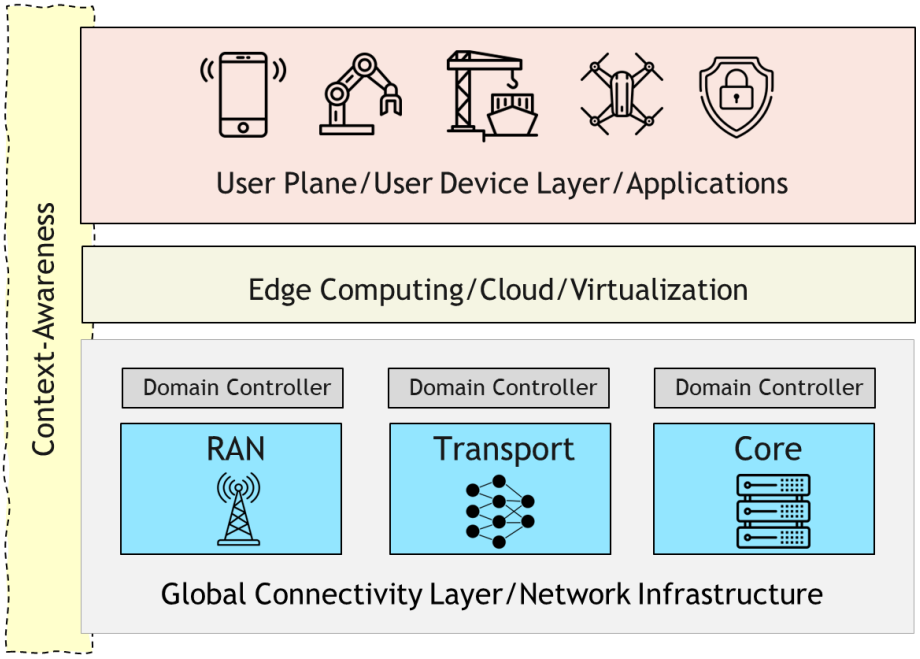

Figure 6-12 Context aware connectivity scenario.

To enforce context-aware connectivity several topics are addressed in this chapter as reported below.

Section 6.3.1 describes the concept of context-aware transport, where a resource orchestrator works with an abstracted view of the transport resources to facilitate the mapping of E2E QoS on the transport resources. This results in optimized usage of the resources.

In **Section 6.3.2**, the delayed computing paradigm suggests postponing the completion of specific tasks based on the requesting users' delay tolerance until the network experiences less traffic congestion as a means of

enhancing the network capacity. This technique contributes to tailoring the network usage to the specific network status while providing the users with reduced service costs as an exchange.

Section 6.3.3 focuses on developing situational awareness in a highly critical environment such as maritime ports, leveraging the network infrastructure to deploy automation mechanisms. Also, gathering information about the ports' environment at any given time and adapt behaviours accordingly.

Section 6.3.4 considers the semantics of the robot task to reduce network overhead and allocate edge resources flexibly, ultimately improving system performance by allowing for multiple edge allocations and RAN slices.

Finally, **Section 6.3.5** investigates a flow-aware transport, which considers the mobility of transport border nodes traversed by a path to support edge-based applications and for multi-connectivity.

6.3.1 Context-aware transport

A transport network provides the connectivity of the RAN/CN functions among the multitude of radio sites by ensuring the physical connections of the interfaces of a mobile 3GPP network. It is constituted by wired and wireless transmission media which interconnect switching and/or routing systems in various topological arrangements evolving the legacy point-to-point links. Research on transport shall ensure the support of the evolutions of radio generations and 3GPP releases.

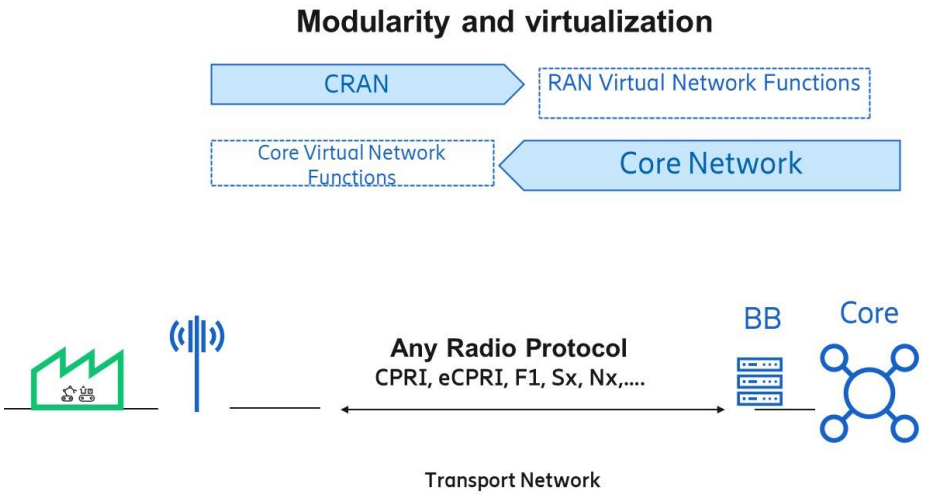

Figure 6-13 Transport network for RAN/CN needs.

Moreover, the RAN and Core Network functions will evolve towards a higher level of modularity and cloudification that requires a higher level of automation in configuration and handling. As shown in Figure 6-13, the transport network will have to support both tight requirements coming from the several radio protocols and from the E2E services.

The context-aware paradigm also applies to the transport network and it refers to its capability to dynamically adapt the connectivity to the mix of supported services with their respective E2E QoS characteristics stated in the SLAs. It includes services with tight requirements (e.g., latency, bandwidth) also including network reliability, availability, and resiliency (NRAR). The "context" concept also refers to the specific deployment areas, such as local, confined, and geographical areas, in which each service is expected to operate. The scope of context-awareness is to match the required QoS while limiting the waste of infrastructure resources (e.g., minimizing the over-provisioning). Thus, the transport infrastructure should automatically and dynamically reconfigure the connections with a smart and efficient distribution of the traffic load that, thanks to the use of suitable AI/ML techniques, best matches to the actual traffic needs considering the service requirements. Transport infrastructure includes several technologies such as wired (e.g., packet, optical, packet-optical), wireless and their combinations, depending on the installed based infrastructure and operator needs, and the capability of the transport domain can differ according to the specific technology. Such AI/ML techniques should consider several aspects such as the followings: they should be based on measurements data to be applied even in case of not availability of historical data. Moreover, the timing used for an AI technique should be compatible with the monitoring of the node to enable the use of such technique with existing transport node.

The rules that the optimization techniques could differ according to the several transport technologies (e.g., packet, optical, wireless) should be taken into account, hence the same AI technique should be able to be used for any transport technology. The heterogeneity of the transport technology creates challenges to mapping the service requests dynamically and automatically on transport requirements. Thus, a suitable abstraction of the transport domain can be very relevant to manage heterogeneous technologies and create an efficient decoupling between the service layer and the transport layer, to allow the translation of the service requirements in transport requirements and then to manage the transport domain accordingly. This is obtained by the use of transport abstraction technique that allows to expose at service layer a same view whatever the transport technology characteristics are.

To enforce the context-aware transport, a resource orchestrator creates an abstracted view of the transport resources and triggers the transport controller for resource handling to satisfy the QoS associated to a slice. The slice allows to organize the physical infrastructure resources in virtual networks, each one dedicated to each service. It also performs E2E admission control to ensure the expected QoS for active and incoming services. An E2E service orchestrator places all network functions on the abstract view to guarantee the QoS of the considered slice. In the abstraction technique, the infrastructure resources among transport edge nodes are abstracted as a set of transport logical links, each one characterized by a set of parameters like bandwidth, latency, resilience level for fault recovery. A network resource is exposed by the corresponding service parameters, hiding many details of the resource (physical details, real topology, etc.). For example, a path with related protected links in an optical network is reported as an E2E link with the amount of bandwidth that can provide and latency instead of reporting all the links composing the path and the supported wavelength channels.

A relevant aspect is to find a method to automatically translate the transport technology specific parameters (e.g., wavelength transmission rate, buffer size of a node, protected network links) into services parameters (e.g., latency, bandwidth, availability) to decouple the service layer and infrastructure layer and apply whatever the technology of the transport is. Several abstraction methods can be used with different level of details of the transport information. In any case, an efficient abstraction method should meet the following characteristics: i) applicable regardless of the transport technology (e.g., packet, optical, wireless); ii) exposed using suitable service parameters independently of the physical transport technology; iii) able to maximize the served traffic with QoS without affecting the scalability as amount of information to be stored and managed.

6.3.2 Delayed computing paradigm

Computation offloading of resource-intensive tasks has been extremely popular to facilitate the computationally and battery-constrained end-user devices to meet their applications' delay and energy QoS requirements. Among the different computing capabilities and options existing across the computing continuum, cloud computing and MEC have revolutionized the successful completion of computation tasks owing to the high computing power of the former and the proximity to the end users of the latter [FAS+21]. Especially driven by appealing attributes related to reduced transmission energy and response time, the prevailing literature considers end users selfishly subscribing to computation offloading services at the edge, overexploiting the edge computing network and gradually leading to its performance degradation.

Indeed, the diverse offloaded tasks and the varied user application requirements create a solid ground for using different computing options across the network. Different tasks are characterized by different levels of intensity, while the user applications pose different delay and power consumption requirements. In this context, motivating the end users to leverage their delay tolerance and allow for the network’s flexibility to smartly orchestrate computation tasks across the edge computing layer and the cloud is a challenging problem to be addressed [DCT+22b].

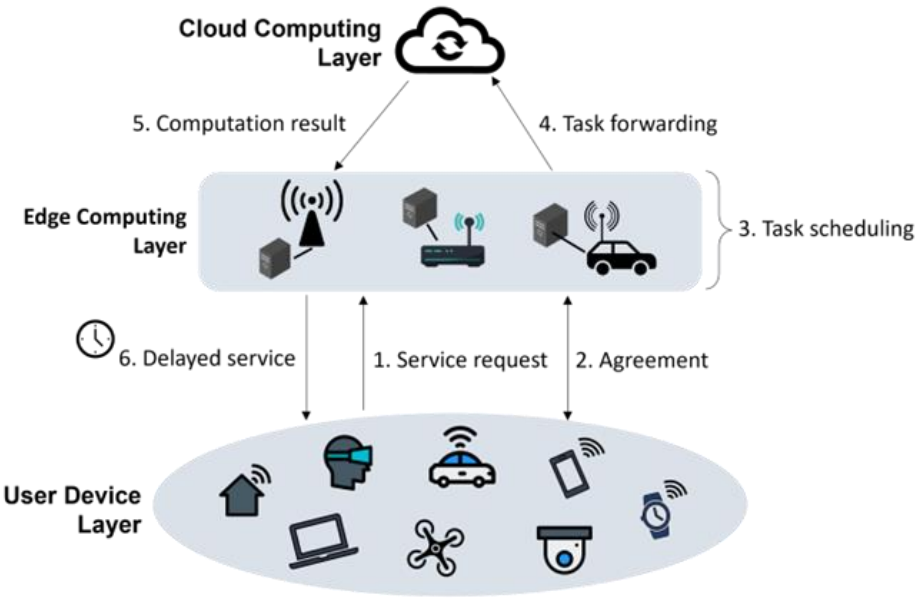

Figure 6-14 Overview of the delayed computing paradigm.

The delayed computing paradigm contributes to network sustainability in the sense that personalized computing services are designed to account for the different end users' QoS requirements and pricing capability, while optimization of the computing resources' orchestration across the computing continuum is performed for extended digital service delivery. The relevant KPIs are end-to-end offloading/communication and computing service delay.

In this study, the goal is to design an appropriate incentive mechanism to motivate the end users to leverage their delay tolerance and price sensitivity and allow for the network's flexible orchestration in exchange for some reduced computing service cost. In the proposed approach, the edge computing service provider will design a menu of bundles, comprising a subscription fee to the computing service and a corresponding response time approximation, by utilizing existing datasets about the end-user applications' potential QoS requirements. Each end user will autonomously select the one bundle out of the menu that best fits its application's characteristics and its payment availability. Having reached an agreement, the computation task offloading and scheduling will take place at the edge computing layer, and even task forwarding to the cloud will be considered, to increase the overall computing service capacity of the network. An overview of the studied delayed computing paradigm is presented in Figure 6-14.

6.3.3 Context-aware connectivity for maritime ports

Connectivity for maritime ports refers to the ability to exchange data and communicate effectively between various stakeholders and systems within the port environment. Most common use cases include:

- In real-time monitoring of vessel movements, sensors can be used to track the movement of vessels in and out of ports, providing operators with up-to-date information on the location and status of each vessel.
- In automated cargo handling, advanced robotics and automation technologies can be used to move cargo within port facilities, reducing the need for manual labour and increasing efficiency.
- In predictive maintenance, data analytics tools can predict when equipment and infrastructure within port facilities require maintenance or repair, allowing operators to plan and schedule maintenance activities more effectively.
- In supply chain optimization, analysing data on cargo movements, inventory levels, and other factors, context-aware connectivity can help optimize the operations within port environments, reducing costs and improving efficiency.

The demanding connectivity requirements from some applications in maritime ports like real-time monitoring of vessel movements, automated cargo handling, predictive maintenance and supply chain optimization, put

additional stress in local networks covering the port location. The constant moving of large metal containers can also pose challenges when it comes to Line of Sight (LOS) of local Access Points (APs) / Edge Gateways (GW), impacting network coverage, connection accessibility, reliability and performance.

This study intends to improve the resiliency and fault recovery of local networks by implementing techniques in the customer Operating System (OS) that we are currently developing for Edge GWs:

- Edge GWs to work cooperatively, improve communication resiliency
- Check and choose between the available connectivity options (adapt to changing conditions)
- Support to multi-wireless technologies
- Master and secondary nodes to relay information

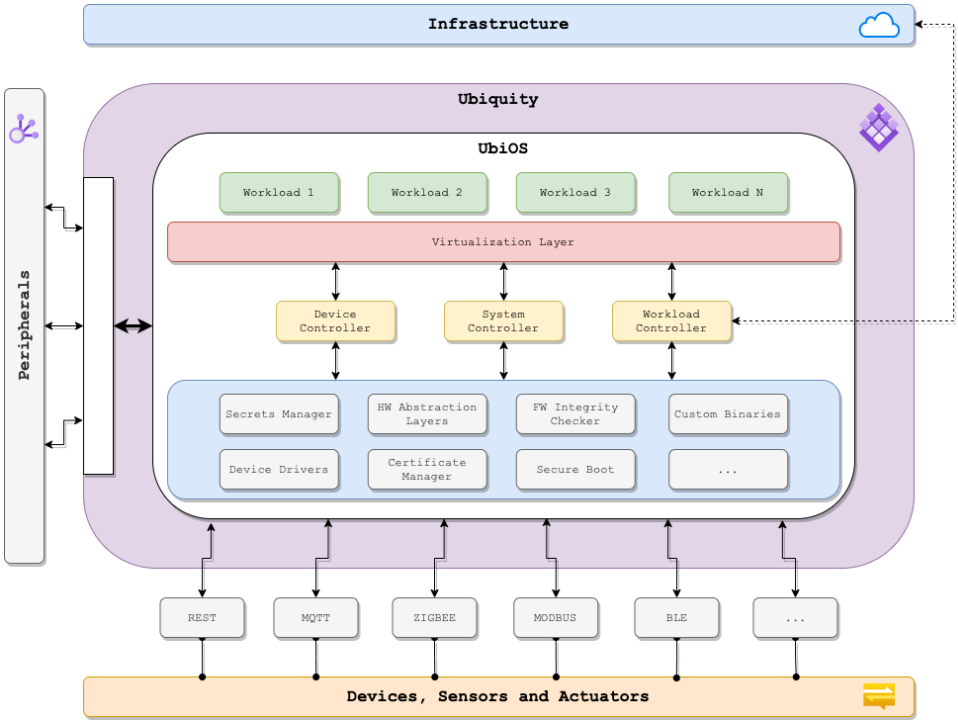

Figure 6-15 Customer Operating System (UbiOS) High-level Architecture.

We also intend to study how the Edge GWs/Nodes and Edge Computing can help in the deployment of sensors and cameras in ports. The collected data from the sensors/cameras or "data producers" is processed locally, enabling latency critical applications and maintaining data privacy. As exhibited in Figure 6-15, UbiOS also provides a virtualisation layer that enables containerised workloads to run locally.

To summarise, Context-aware connectivity involves deploying a range of technologies, as we can see in Figure 6-16, including sensors, wireless networks, and data analytics tools to gather and analyse data about ports operations. With this, the data can be optimized in various aspects of port operations, such as traffic flow, cargo handling and supply chain management. This is an important tool for enhancing the performance and competitiveness, allowing to better meet the needs of shippers, carriers, and other stakeholders in the global logistics industry.

Figure 6-16 Connectivity for maritime ports.

Connectivity solutions play a critical role in improving the performance and efficiency of maritime ports, enabling better communication and collaboration between stakeholders, and facilitating the use of advanced technologies to optimize port operations.

6.3.4 Context-aware and flexible RAN

The number of mobile robots using 5G-and-beyond cellular networks is expected to grow from 40,000 units in 2021 to 350,000 by 2030, according to ABI Research [ABI+22]. As mobile robots have proven to be effective in tasks such as material handling, transportation, and cleaning, industrial verticals are now interested in using them for outdoor applications due to their autonomous navigation, manipulation, and functional safety capabilities. To perform their mission-critical operations, mobile robots will continually execute complex object detection and autonomous navigation tasks, which require high-resolution video. Examples include multi-object detection for obstacle avoidance, people detection or human interaction.

However, continually sending this data to the edge of the network may eventually saturate the RAN, especially in robotic use cases where the RAN is shared with users. To this end, in recent years the concept of RAN slicing appeared as a technique that enables Virtual Network Operators (VNOs) to allocate and virtualize the computational and networking resources of the RAN based on their requirements. Notably, this technique is supported by the Open RAN (O-RAN) framework, which separates the hardware and software components of the NextG RAN to allow more precise, real-time control over the RAN components.

Current State-of-the-Art either does not consider RAN slicing or usually defines edge-based tasks as monolithic, which leads to sub-optimal performance. The scope of the context aware and flexible RAN that is shown on Figure 6-17 and is considering the contextual information from robot task to reduce the network overhead and allocate resources in a flexible way. Flexibility allows for the consideration of multiple computing allocations and RAN slices to the same task-related performance, ultimately improving the end-to-end system performance. Two main concepts are considered in this study. The fist main concept is that different robotic tasks have different tolerances to video compression. For example, a person can be more easily identified in a noisy video as opposed to a chair or a box. The second main concept is the flexibility in robot task deployment. Indeed, the robotic task can be executed locally on the robot, offloaded to the edge server or

pre-processed at the robot and then if needed offloaded to the edge. Therefore, the slicing algorithm can select the correct computation slicing, radio slicing and offloading policy while meeting the performance requirements. Having in mind these two concepts and aiming at better end-to-end system utilization of the existing radio and computation resources, this study tries to answer two questions: i) What portion of each robot task computation demand should be offloaded for remote execution in the edge servers to improve the number of allocated robot tasks and the lifetime of the robots? ii) How we can allocate slices having in consideration the relation between the video compression, classification accuracy, network latency and robot battery lifetime?

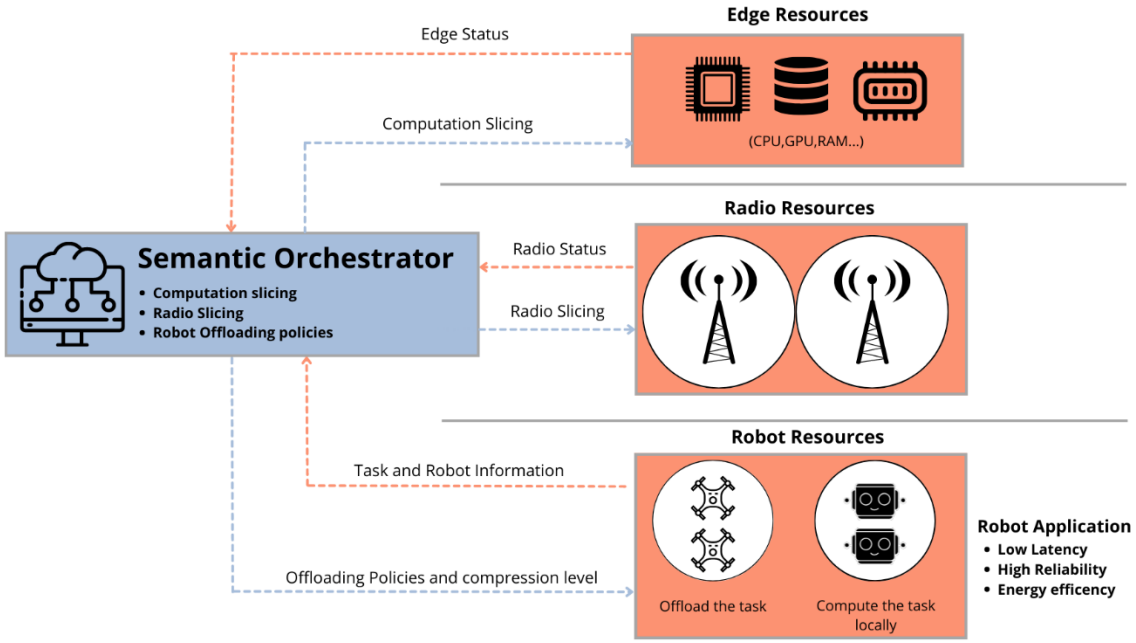

Figure 6-17 Context-aware and Flexible RAN in mobile robots.

6.3.5 User Plane supporting mobility of both ends of a path

One of the essential issues in mobile networks is the efficient data plane design that will avoid concentration points of traffic on its end-to-end route and simultaneously handle mobility and QoS of user sessions. The IP communication relies solely on locators (host interfaces' addresses) that are, unfortunately, also used as node/service identifiers at the network layer. The approach makes IP mobility management troublesome. As a result, traffic anchors and tunnels have been introduced to handle mobility while preserving the identifier exposed to the transport layer. The existing solutions use the GPRS Tunnelling Protocol (GTP), which involves an unavoidable overhead. Moreover, it enforces traffic aggregation at specific points. Using GTP leads to ineffective traffic steering because the payload data needs to leave the tunnel at its end and then can be redirected to the destination. It leads to the existence of long paths and contributes to the E2E delay. Using servers deployed at the edge (e.g., MEC hosts), integration of different connectivity domains (e.g., NTN) or context-aware operations require handling the mobility of both ends of the path. Moreover, an efficient traffic distribution, which would avoid traffic concentration and reduce overhead introduced by tunnels, is needed.

It is possible to solve the problems mentioned above with SDN. The most attractive feature of SDN in the context is an easy redirection of traffic flows, in which IP headers with source/destination addresses are used as labels only. Therefore, mobility can be natively supported with no anchors, no encapsulation overhead and no need to use tunnelling. However, SDN has its drawbacks, which lie in the lack of scalability and (so far) minimal QoS support.

In this section, a multi-domain SDN solution is proposed to solve the scalability problem. In the presented in Figure 6-18 approach, there are multiple, relatively small SDN domains, each with its own SDN Controller. The approach speeds up the setup of local paths that can be done in parallel in each of the domains that may

have a minimal number of nodes. The SDN domains are interconnected via dedicated Border Nodes. Each SDN controller can set up a path between Border Nodes or redirect traffic to a destination node if it is located in its domain.

The mobile networks from the beginning have required handling of the mobility of the end-user, so far supported by GTP. The use of servers deployed in edge (e.g., MEC hosts), integration of different connectivity domains (e.g., NTN) or context-aware operations require handling the mobility of both ends of the path. Moreover, an efficient traffic distribution, which would avoid traffic concentration and would reduce overhead introduced by tunnels, is needed. The SDN technology can be used for such purposes. However, it raises scalability issues as the solution is centralised (logically, only a single controller). A multi-domain SDN solution is proposed to solve the scalability problem, in which multiple SDN domains are interconnected via dedicated Border Nodes (see Figure 6-18). Each of the SDN controllers sets the path between Border Nodes up or redirects traffic to a destination node if it is located in its domain. The information about paths between Border Nodes of each domain is stored on-demand in the Global Connectivity Layer (GCL), which uses this information to create end-to-end paths. The GCL keeps the information about all available connections and allows for redirections of the end-to-end flows. The same mechanisms can be used by a Traffic Engineering Engine or a Mobility Management Engine deployed atop GCL. The traffic redirection can use source routing mechanisms (i.e., terminal-based path change).

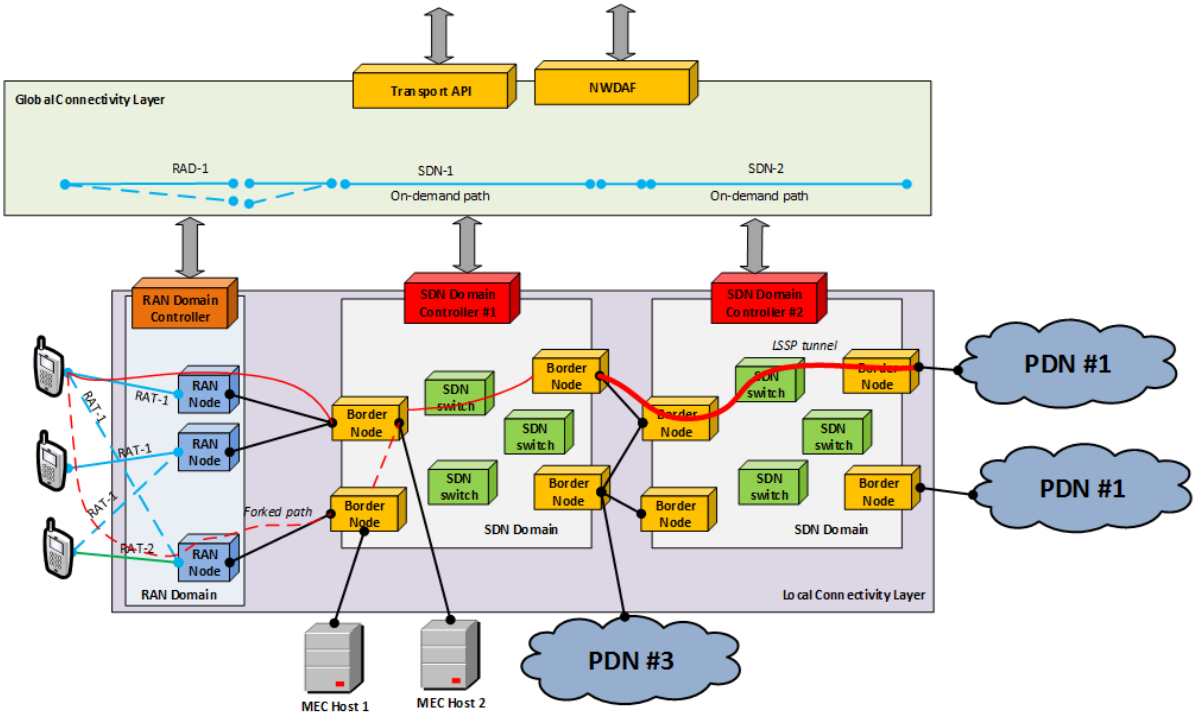

Figure 6-18 Tunnel-free User Plane architecture (SDN-based).

The information about paths between Border Nodes of each domain is stored on-demand in the Global Connectivity Layer (GCL) database, which uses this information to create the end-to-end paths. The GCL keeps the information about all available connections and their properties (delay, jitter, packet loss rate) and allows for redirections of the end-to-end flows, allowing the mobility of both ends of the path. The GCL has NWDAF-like functionality to predict the transport networks' status and expose information about the network (links load, topology, paths, jitter, flow to links assignment, etc.) to applications. Such applications may include traffic engineering or mobility management engines, which, in the concept, are deployed atop GCL. The GCL can be implemented as a distributed database to solve its scalability problem. One of the most challenging problems of the approach is traffic engineering, which should include optimising a path inside each domain and the overall end-to-end path optimisation. For that purpose, a cooperative agent approach can be used.

7 Network beyond communications

7.1 Introduction and overview

7.1.1 Network beyond communications enablers' overview

Communications has been the primary objective of wireless and mobile networks up to 5G. Some examples such as IoT and Edge Computing were pushed to the spotlight in the context of 5G networks, towards extending the network scope and capabilities beyond communications, nevertheless in those cases, connectivity was still the primary target. Towards 6G, the evolution of the network is pushing the boundaries, beyond conventional connectivity, into accommodating and supporting novel services, expanding the network's scope by processing data, generating insights, and delivering added value from societal, innovation, and business perspectives. Examples of new services comprise sensing, enhanced localization and tracking, compute-as-a-service, and AI-as-a-Service, as illustrated in Figure 7-1. These cutting-edge advancements will redefine the boundaries of industries, fostering seamless integration of sensors, data analytics, and computation, as well as unlock unparalleled levels of efficiency, productivity, and innovation. In the following we refer to the added network functionality, in the form of new services not primarily for communication, as Beyond Communication Services (BCS).

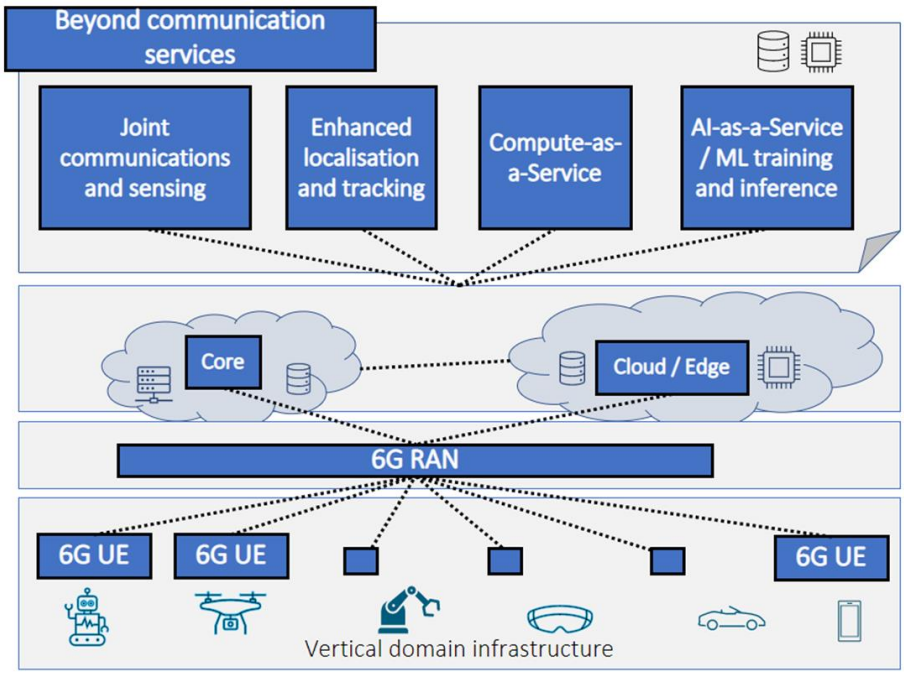

Figure 7-1 Beyond Communication Services overview.

The following sections provide a list of BCS enablers and respective study areas of focus that will be studied in the project. The enablers identified are grouped into four main categories, namely i) exposure of data and network BCS capabilities, ii) protocols, procedures and signalling optimisation aspects for supporting BCS, iii) application- and device-driven optimisation for BCS, and iv) enablers for enhancing JCAS capabilities.

7.1.2 Contributions to Hexa-X-II PoCs

To showcase its ambition and address project objectives, Hexa-X-II will develop three System-PoCs (namely System-PoC A, B and C); each System-PoC encompasses a set of Component-PoCs. As initially PoC A focused on integrating enablers related to management and orchestration aspects (WP6), PoC B will target to further evolve the System PoC and integrate enablers related to network architecture elements (WP3). More specifically, WP3 contributes to PoC #B.3 (Figure 7-2), leveraging the studies and enablers provided by the Trustworthy flexible topologies (Task 3.3, Section 6.1.5), and Network Beyond Communications.

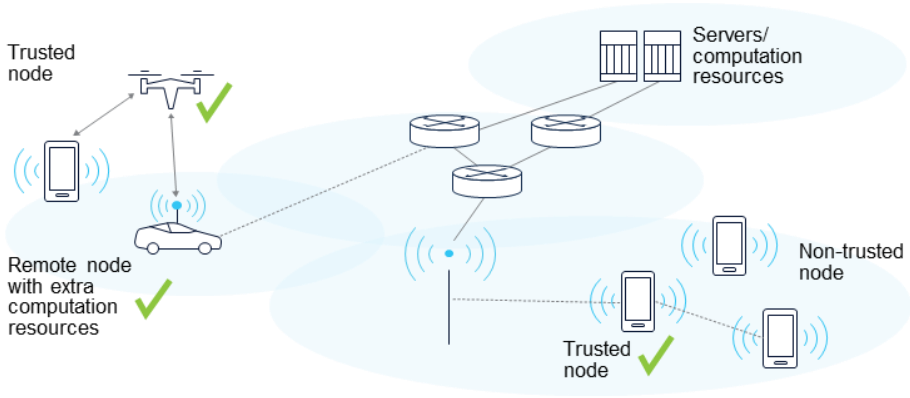

Figure 7-2 Component-PoC#B.3.

In the context of the latter, information regarding computing resources that can be used for certain workloads, such as computer vision-based AI/ML computing tasks will be exposed; enablers on exposure and data management (Enabler #2, Section 7.2), as well as related procedures (Enabler #3, Section 7.3) will require for the respective designation of the relevant communication resources; to this end, a trusted flexible topology, involving one or more flexible nodes (namely ground robot or UAV) will be established offering access to the available edge cloud computing resources.

7.2 Exposure and data management

With the further expected rise of IoT, to unprecedented levels, emergence of new applications, and an increasing number of connected devices supporting immersive and extreme 6G use cases, vast amounts of data will be transmitted and processed by various services and vertical applications in sectors like logistics, manufacturing, agriculture, and healthcare, leading to significant improvements in efficiency and productivity. The next generation of networks will need to support not only connectivity for broadband devices but also an improved support for sensors, like ambient and power limited devices, which may be static or mobile. The means via which the network can support the management of data communication for this wide range of devices will thus be highly important. The network shall, in an efficient way, handle small data packets with their required QoS, coming from more or less power limited sensors, and at the same time support streaming data to and from UEs and broadband services with their specific QoS requirements. Data management, as well as data and capabilities' exposure will be critical towards efficient interaction among involved data consumers, network functions, services, or 3rd party applications.

New functionalities such as sensing, advanced localization, and tracking, including JCAS, will inevitably generate large amounts of data with distinct characteristics from the existing user plane (UP) and control plane (CP) data. This includes the JCAS data, which may resemble UP data but won't be connected to a specific user, creating an entirely new stream of data. These expanded data volumes, or beyond-communication data, will have to be managed by the network or fused at various network locations, like access points, for efficient and coherent processing. Hence, appropriate design measures must be implemented to ensure realistic scaling that doesn't compromise either the delivery or the integrity of standard CP data. These measures might encompass interfaces for transferring the JCAS data to a new data plane and to external entities, as required. A thorough study of these new data volume requirements will facilitate the efficient design of envisaged functionalities and services associated with sensing and tracking in selected use cases.

In principle, the aggregation, processing, and exposure of the data through core-RAN continuum is an important attribute in 6G. A key challenge during the data aggregation rises from the trustworthiness of the data exposure dynamics. In 6G, the data that is collected from various sources will be cleaned and labelled at different locations. Minimizing the privacy risks and ensuring the trustworthiness during the complete data management cycle is a key challenge in enabling data-driven networks. As data processing and insight extraction become more complex, it is thus crucial to develop novel architectural enablers (Figure 7-3), which prioritize security and trust, ensuring data is protected and insights are generated safely and reliably. Those novel enablers should not only support secure and trustworthy management of data but also facilitate the allocation of functions and applications that interact with the data. By emphasizing security and trust aspects

in the design of these architectures, we can guarantee that data remains protected, and insights -as a second step- are derived in a manner that is safe, reliable, and trustworthy. Through intelligent (compute) node selection for application placement and processing, secure and efficient data management will be enabled.

Additionally, the expected increase in the data volume utilized in the network and by the applications can also cause latency challenges, which can drastically impact the performance. In addition to enabling trustworthiness and resource efficiency, sustainability is a major design metric for the data aggregation and exposure model.

Given this increase in data volume and the associated computational load, efficient data processing necessitates a change in the system's architecture. Rather than merely managing data, we must also consider the efficient allocation of the computational tasks related to the data. As such, when a device or network node decides to offload a computation, it will have to discover and select the candidate compute nodes, capable of performing the requested computation while satisfying the associated KPIs. To efficiently perform the processing (compute) node selection it is required to precisely define the parameters exchanged during the discovery and localization procedure. These parameters should include processing (computing) capabilities of network and/or device nodes and requirements, such as latency and computational load.

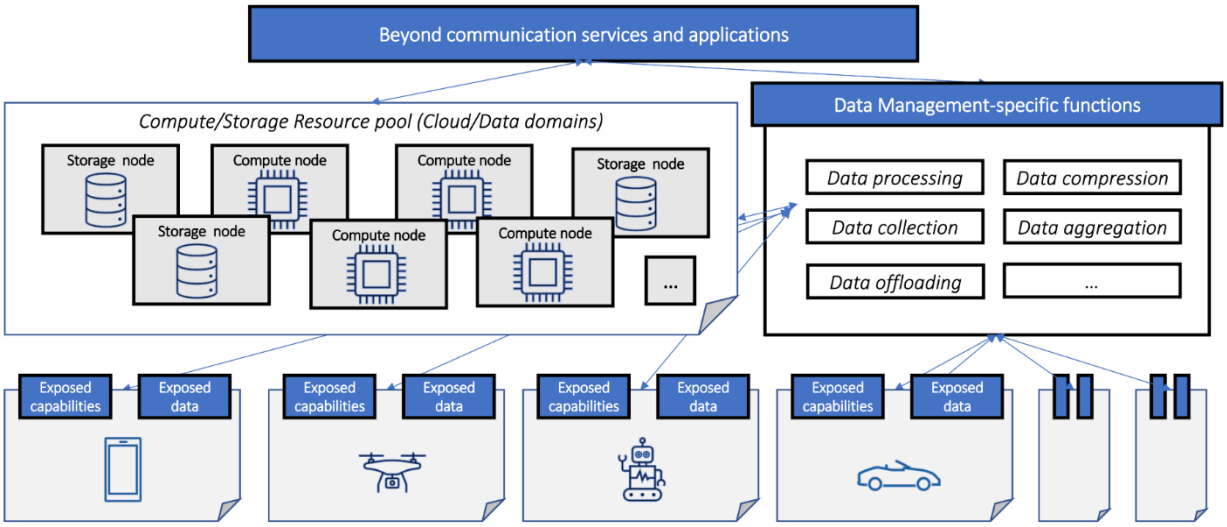

Figure 7-3 Exposure and data management enabler concept

Exposure and data management enablers, illustrated in Figure 7-3, are investigated in two study areas, structured in the following sections.

Section 7.2.1 will focus on selecting the use cases that are enabled by and/or benefit from JCAS service, on the allocation of applications that use sensing, as well as on development of end-to-end exposure framework enhancements considering privacy, security, sustainability, and performance challenges both for in-network and external exposure scenarios.

Section 7.2.2 discusses aspects related to investigating the data volumes in specific sensing scenarios and more specifically to assessing the data loads on the radio link (bandwidth) associated with a certain tracking (as a BCS) performance, and in relation to the number of users served. The KPIs and KVIs of the specific BCS will be also considered.

7.2.1 Data and functionality exposure for JCAS services

JCAS is currently associated to radio level functionality such as beam sensing, assisted beam training, tracking, prediction, power allocation, allocation of radio bandwidth between communication and sensing, etc. However, the potential of JCAS is much wider covering also the applications using similar sensing functionalities. Indeed, when considering JCAS use cases such as enhanced localization and tracking, monitoring and management of V2X and UAVs, or Smart Home/Factory, the sensing functionalities are being used/controlled by the underlaying applications associated to these use cases (cf. Figure 7-4).

The major challenges in defining enhanced functionality for JCAS are as follows.

- **Trust differentiation when exposing to 3rd party applications:** The interaction between network and applications as well as the availability and use of data from various data sources is expected to increase, which in turn may impact the exposure framework. Moreover, exposure of certain data to third parties facilitates the development of new applications and services that utilize the network data and vice versa. Considering the requirements of data exposure and different trust associations, proper mechanisms for exposure limitations should be in place and the level of exposure (i.e., the degree at which the third-party app can control the resources/data in the core) to 3rd party applications should be monitored.
- **Network overload on the exposed APIs**: Exposure of JCAS services and data is typically accomplished with application programming interfaces (APIs), use of message queues, or shared databases, which provide a standardized method for various network functions to communicate with one another. However, due to the anticipated large amount of data to be exposed by the network to several internal functions and 3rd party applications, problems related to potential rise in traffic propagated over the exposed APIs can arise. Performance efficiency of exposure needs to be enabled.
- **Location of applications using sensing functionalities**: depending on the considered use case, the location of the applications using sensing functionalities could become very crucial to meet the expected QoS. While the sensed objects in some use cases are moving in a relatively high speed and several transmission/reception nodes are contributing to JSAC activity, other use cases consider sensing low-speed objects but require strict delay performance to execute actions (e.g., stopping a robot machine after detecting a human). The placement of applications using sensing functionalities becomes very challenging to meet the target QoS.
- **Increase privacy risk:** With the vast amount of data that is expected to be generated and utilized for applications such as analytics generation and ML model training, privacy infraction risks exacerbate. While the goal is to make these data accessible to nodes wanting to utilize them, at the same time exposure of the collected and aggregated data needs a strong authentication and authorization mechanism to ensure that security and privacy will not be breached.
- **Latency challenge:** Due to additional steps necessary aside from data collection such as data aggregation, data cleaning, data labelling, etc., that needs to be done before data is exposed, the time from the data collection to the time where the data can be used may increase.

Some of the KPIs that can be considered related to data exposure includes:

- Data Privacy, Explainability, Security, Trustworthiness, Integrity
- Accuracy, latency, coverage
- Network load/Bandwidth Efficiency
- Sustainability / energy efficiency
- Number of services enabled in the network with optimized support.

In this study item, the research activities will first focus on selecting the use cases that are enabled by and/or benefit from JCAS services. We will further analyse how applications that use sensing functionality should be allocated. Finally, we will study the development of end-to-end exposure framework enhancements considering privacy, security, sustainability, and performance challenges both for in-network and external exposure scenarios.

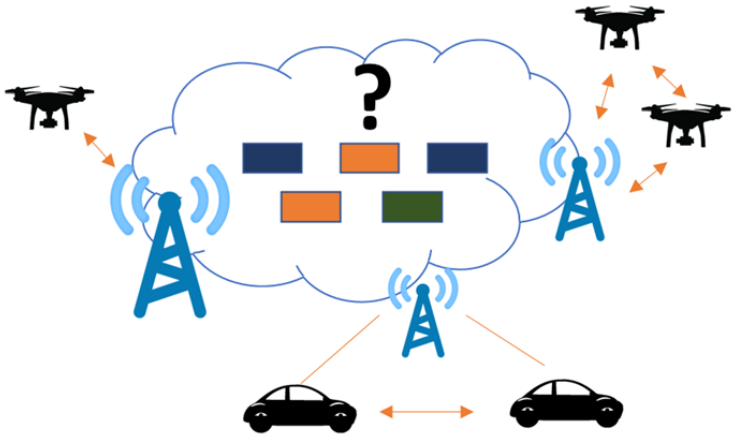

Figure 7-4 Extended JCAS functionality, e.g., V2X, enhanced localization and tracking etc.

7.2.2 Incorporation of L1 sensing functionality

Sensing in the context of Hexa-X-II, is the capability of the network to infer information about the physical situation and context surrounding network's radio nodes. For example, based on a proper use of the network's radio signals and protocols, knowledge about the presence, location, speed, etc of objects in the vicinity of the network's radio nodes may be obtained. The 6G architecture must be prepared to efficiently include the new sensing services and potentially sensing involves simultaneous inference of the physical parameters of a massive number of objects in the vicinity of a massive number of radio nodes. To assure sensing capabilities even in these scenarios, there is a need to study scaling issues with this new functionality.

Several questions will need to be answered. Firstly, based on some suitable and representative use cases, the potential volumes of new data that would be produced and processed needs to be assessed. Furthermore, and in particular, how do these data volumes trade off with the other performance metrics of the networks? Will bandwidth need to be sacrificed for sensing, and if so, in which scenarios? How much bandwidth would need to be sacrificed in these cases to meet sensing requirements and vice versa? An interesting question related to the scaling of the data relates to the identification of the very parameters that determine the data volumes (number of users, measurement rates, measurement compression ratios, etc). Finally, given the insights that follow from the first set of questions, how does this translate to requirements on the RAN interfaces?

This study will initially investigate the data volumes in a particular sensing scenario – the tracking of moving non-connected moving objects in the vicinity of a single radio transmitter/receiver where the objects are tagged with a reconfigurable intelligent surface. Figure 7-5 shows a scenario (left) reflecting this. An urban environment where non-connected moving objects (bicycles or other) are assumed to be tagged by a RIS. The right-hand figure shows the true trajectory along with an estimated.

This study will be concerned with the data loads on the radio link (bandwidth) associated with a certain tracking performance. Furthermore, we will investigate how the number of users that simultaneously be tracked plays a role. Finally, this study develops and proposes protocols for how to carry these data volumes in the radio link.

We will in other words assess the KPI's and KVI's associated with the tracking performance in terms of for instance squared position or velocity errors, but also in those associated with the price to pay in terms of communication capacity losses.

The resulting concept/solutions will involve understanding of the trade-offs, along with some initial signalling solutions.

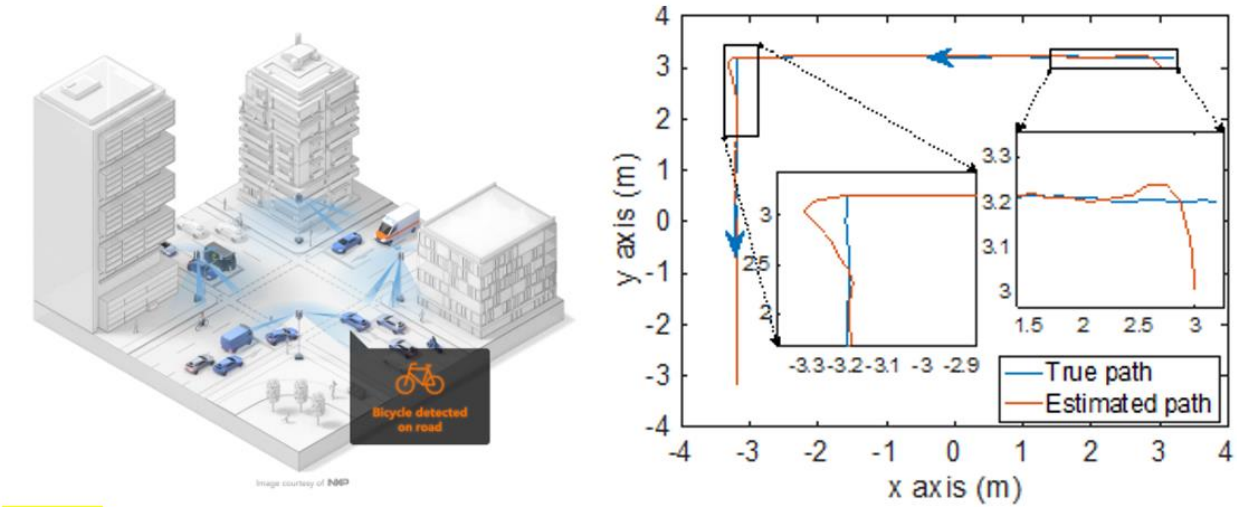

Figure 7-5 Tracking scenario where moving non-connected objects are tagged with a reconfigurable surface. Left: example scenario. Right: simulation example of a true trajectory (blue) along with an estimated trajectory (red).

7.3 Protocols, signalling and procedures

The emergence of new applications, and potentially new devices with diverse capabilities requires the next generation of networks to provide BCS (e.g., computing and sensing), in addition to the legacy communication ones. This would entail a tight integration of communication, computing and sensing as a result. However, the introduction of related services, such as Compute-as-a-Service (CaaS) and Sensing-as-a-Service (SaaS), should not increase the complexity of the communication protocol. The intelligent utilization of communication and computation resources should bring an optimized Quality of Experience (QoE) for the communication as well as the required resiliency and quality of computation and sensing.

Moreover, the true convergence of communication, computing and sensing will bring stringent requirements on latency, privacy/security, power consumption and data accuracy. Therefore, it is necessary to introduce the corresponding novel architectural enablers that include additions and/or modifications of network protocols and procedures. This imposes several challenges in connection establishment procedures that must be addressed, such as discovery (including the exchange of compute parameters discussed in Section 7.2) synchronization and coordination of computing nodes, as well as the impact of new sensing services on RAN interfaces and functionality.

The motivation for device computation offloading is likely due to save compute power and energy. Another motivation can be to offloading of a collaborative task shared by a local set of devices and due to devices with limited computational capabilities, e.g., IoT sensor devices. One possible solution may be to expand the application functionality dynamically from a mobile device to computing embedded into the network. The solution should offer offloading of critical tasks or functions to app developers, exposed as a network service. However, care must be taken to have an efficient network utilization and power consumption. For the network to handle a sensing request, several new functions are probably required such as management, configuration, authorization processing of the measurements.

Scalability is another important aspect of both the compute offloading and sensing functionality. This means that protocols must be efficient also for the case when these services are widely used in a wide area, both from a network resource point of view and energy efficiency. The efficiency of protocols is an aspect that is closely related to the study described in the previous Section 7.2.2

Protocols, signalling and procedures for BCS enablers are investigated in several study areas, structured in the following sections.
Section 7.3.1 focuses on the definition of generic properties of typical offloaded functions, such as degree of offloading, deployment options (pre-deploy/ad-hoc) and offloading initiator type (device/network). It further

proposes a dynamic device offloading solution, which aims to dynamically expand computation functionality from a mobile device to computation embedded into the network.

Section 7.3.2 addresses the optimization of the signalling and procedures for computation offloading. This includes the proposal of a general functional architecture for the distributed compute, where different functional nodes (i.e., Offloading, Computing, Controlling, Routing) and their functionalities are introduced.

Finally, Section 7.3.3 studies the impact of new sensing services, on RAN interfaces and functionality. This includes the investigation of the requirements, protocols, and solutions that enable the introduction of JCAS. This study will investigate new network functions that are required to execute sensing in cellular networks.

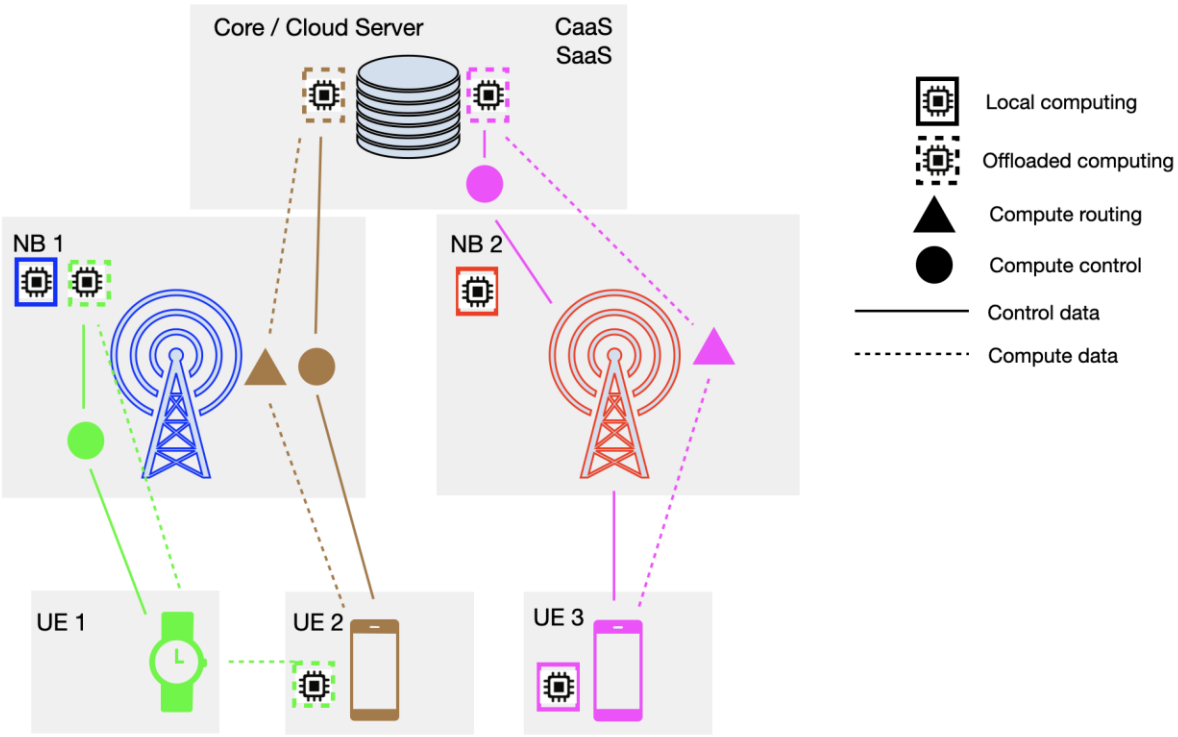

Figure 7-6: The general architecture and protocols for the converged communication and computing.

7.3.1 Distributed compute as a (beyond communication) service

Device offloading enabled via Compute-as-a-service (CaaS) provides *a mechanism to move computation from a mobile device to a connected site with more suitable compute and storage capabilities*, which is an aggregated definition fully in line with state-of-the-art definitions in literature.

The motivation of distributed compute is to trigger offloading of computational tasks from a device dynamically based on situational or environmental changes can include among other things reducing computational response times [GDT+22], [IDG+21]; balancing compute and energy trade-offs [GDT+22], [IDG+21], [MS22]; balancing performance and cost trade-offs [MS22]; reducing device heat; facilitating local synchronization and coordination; optimizing network utilization; or increasing application scalability and availability. Dynamic device offloading as a network service targets the long tail of enterprises/regional companies or even individual developers who want to extend their device applications with new functionality without going through elaborate onboarding steps as in ETSI MEC [KFF+18] or additional business relations beyond a cellular connectivity subscription. We identified a few generic properties of typical offloaded functions, and list an initial set of real-world use-cases that fall into these categories (see Figure 7-7)

- Temporary offload of critical functions from a device to balance computation power and energy consumption/battery drain (can be automated based on contextual triggers)
- Temporary offload of a collaborative task shared by a local set of devices
- Always offload from a device with limited computational capabilities, e.g., IoT sensor devices

Use case	Degree of offloading	Deployment	Scheduling	Initiator
Mine inspection through autonomous vehicles	Partial/full	Pre-deploy	Dynamic	Device
Mobile robots in factory facilities	Partial/full	Pre-deploy	Dynamic	Device
Service robots in public environments	Partial/full	Ad-hoc	Dynamic	Device/network
Remote rendering, e.g., XR	Partial	Ad-hoc/ Pre-deploy	Dynamic/ static	Device/network
Drones/Robots in disaster site inspection or search & rescue	Partial	Ad-hoc	Dynamic	Device/network

Figure 7-7 Use case examples for dynamic device offloading.

The KPIs for the offloading can for example be network utilization, computational response times, power consumption and device heat between critical application tasks run on a device vs offloaded into the network.

We propose a dynamic device offloading solution Figure 7-8, which aims to expand application functionality dynamically from a mobile device to computing embedded into the network. This is in contrast to existing Edge computing solutions, which rather expand the Cloud to the Edge or on-prem sites.

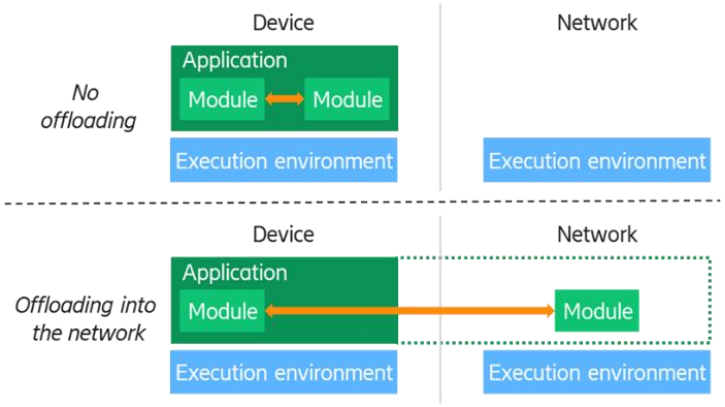

Figure 7-8 Dynamic device offloading as a network service.

The proposed dynamic device offloading solution will:

- offer offloading of critical tasks or functions to app developers, exposed as a network service
- support offloading of customized application modules
- offer arbitrary high application and user granularity through novel programming models and programmatic APIs
- initiate offloading dynamically based on contextual and situational changes
- take advantage of network resources, processes, and information, including piggybacking existing and emerging network procedures, sharing of (cloudified) network infrastructure, and offering in-network computation

7.3.2 Protocols and procedures for computational offloading

The emergence of new applications, heterogeneous and computationally heavy use cases, and potentially new devices in 6G brings a tight integration of computing and communication. Power constrained devices, or devices with limited computational resources should be able to offload some computations to an external (NW or device) and more capable node. However, the complexity of the protocol stack should not be increased by introducing the computing services. To achieve an optimized QoE for the communication and required resiliency and quality of computation, the utilization of communication and computation resources should be carefully designed. This imposes the following challenges in connection procedures that must be addressed:

- Discovery of the candidate compute nodes, including their configuration, their localization, selection criteria definition. These operations have to be performed with adapted/novel signalling and security procedures.
- Connection establishment between the offloading and compute nodes that comprises their synchronization, the exchange of the computing capabilities and requirements such as latency, computational load, and related parameters.
- Computation phase, which assumes joint compute and communication scheduling and management and transfer of the computation loads. New procedures must ensure compute service continuity and resiliency.

Immersive education / telepresence services and enhanced interactions are computationally heavy and may not be feasibly performed on a single device. Therefore, offloading to nearby devices or distributing the computation is not just beneficial but also necessary.

On-device machine learning training and inference is computationally demanding and can be partially offloaded in a distributed or collaborative fashion, while preserving user privacy requirements.
In a converged communications and computing system, to make computational offloading appealing for the 6G user device, some requirements on latency, power consumption, offloaded data accuracy and privacy will be imposed. The guarantees on these KPIs/KVIs can be achieved by carefully designing trade-offs between communication and computation requirements, e.g., communication vs. computation latency and power consumption.

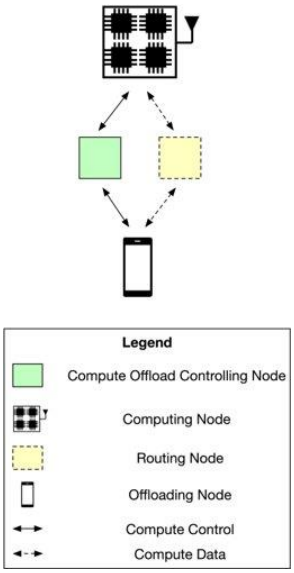

Figure 7-9 Distribute compute: General functional architecture.

The general functional architecture for computational offloading is shown in Figure 7-9. The Offloading Node is the node connected to a wireless network, having a compute task to be offloaded to one or more Computing Nodes, which are wireless network nodes with certain processing capabilities to perform an offloaded compute task and produce compute result. Compute Offload Controlling Node is a wireless network node that collects all compute capabilities from all available Computing Nodes and makes compute offload decision based on their current load. The Routing Node is an optional network node at which the compute task/compute result from Offload Node/Compute Node gets routed to one or more Computing Node(s)/Offload Node. The physical realization of the different logical entities (nodes) in the cellular NW architecture has different variants, e.g., Compute Node and Routing Node could reside in a single physical entity.
To improve the Computing Node discovery and selection, the exchanged parameters among the nodes should be investigated. Moreover, to address latency and power consumption requirements, the characterization of the offloading procedures and classification of compute workloads will be conducted. To efficiently utilize the communication and computation resources, the optimization of the signalling procedures for computation, such as computation node discovery and computation task/load transfer should be studied.

7.3.3 New network functions and procedures to support JCAS

The idea with JCAS is to use the radio resources for communication also to locate or trace objects in the cell. Different modes of sensing are anticipated, e.g., using characteristics of Sounding Reference Signals (SRS) to identify objects or introducing a radar-like procedure to measure the distance to objects. Several studies, especially on the physical layer procedures, exist.

In this study there will be a description of the different functions (that may represent a Network Function) that are needed to execute sensing in cellular networks and the signalling needed for how these functions interact. There will also be a description of how these fit in a future 6G architecture.

The Figure 7-10 shows a possible signalling diagram for a sensing request. In this diagram there are three new functions. One function to control the process, i.e., configure relevant parts of the RAN and CN, on function that checks if the requester is authorized and one function the processes the measurements.

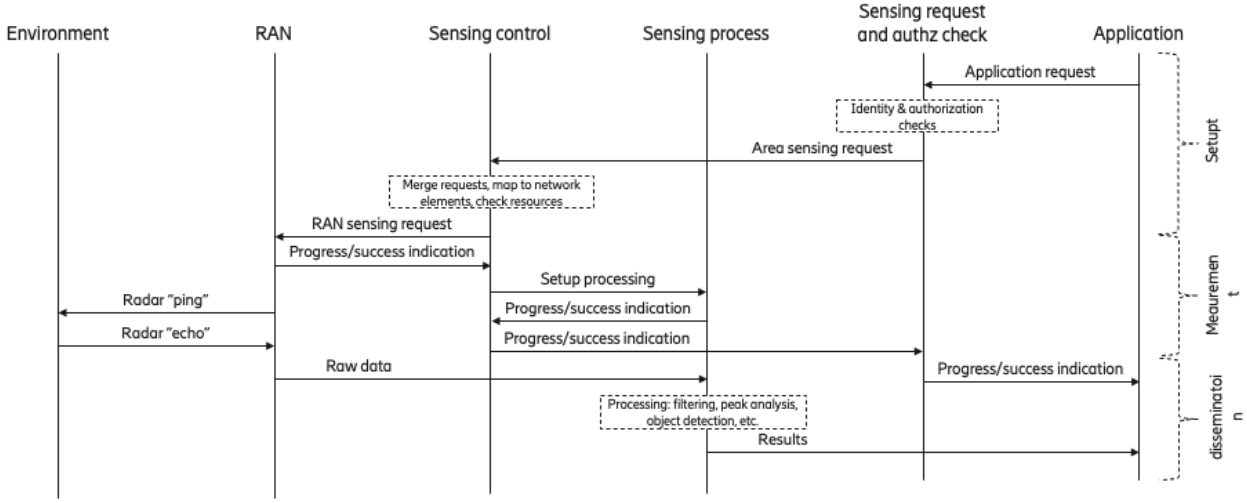

Figure 7-10 Signaling generated from sensing request hinting at new NFs.

In more detail, this is what happens in the figure:

- An application requests sensing output by sending a request comprising, e.g., area of interest, characteristics such as "is there anything in this location" or "how far away is this object", and some type of expected QoS.
- A new NF called, e.g., Sensing request and Authorization, handles the Sensing request and authorization. The NF both verifies if the client is genuine and identifies the area where sensing is to be performed.
- Once authorized, the request is forwarded to the NF Sensing Control, which configures RAN with the necessary settings and processing resources.
- The involved RAN and sensing processing nodes ack the request and configuration.
- The RAN node collects sensing measurements and propagate them to the specified sensing processing nodes while providing a status update to the sensing control
- The sensing processing NF interprets the measurements providing results that are disclosed to the requester. A status update is provided to the sensing control NF.
- Progress and success status are indicated to the application via sensing control in parallel to with sensing results.

This sequence is generic and represents sensing for both external and internal requests. An external request can be sent by a car trying to identify potential obstacles (from directions not covered by onboard lidar). An internal request can be sent by a device hoping to transmit with very high bitrate in the uplink and needs LOS for the transmission. In this study detailed functionality will be presented and how the functions fit in a future architecture.

For JCAS to work, according to the sequence diagram above, the below functions are needed:

1) Service exposure

There must exist a way for an application to request sensing, i.e., to initiate the procedure that determines where an object is located. This is probably an application programming interface (API) running on top of an existing exposure framework, existing data distribution network (e.g., Evolved Data Collection Architecture, EDCA) or a new function.

2) Handling of Sensing requests

To support various kinds of sensing there needs to be a function that handles requests from different applications. The function could be split into two logical, or physical, entities – control and processing. In addition to handling sensing scheduling and coordinates resources with communication scheduling.

3) Sensing data processing

This function interprets sensing measurements and converts them into a format meaningful to an external receiver. Further processing may involve, e.g., creation of maps. Available local data is handled by this function, e.g., base station ID and observations. Also, sensor fusion information could be included to enhance results.

4) Sensing service availability

A requesting application needs to understand what sensing services are available in the requested area and period. The API may include additional information regarding the sensing potential and availability of resources, e.g., active sensing might be limited in an area with high load, e.g., rush hour traffic. Further, the API should provide assurance for critical application coverage.

5) Authorization and data disclosure

The access to sensing data needs to be limited to satisfy privacy and security requirements. The access control applies to intermediated data and both internally and externally. Therefore, a sensing request must be verified, i.e., is the requester trustworthy.

6) Privacy preserving mechanisms

A core feature of the system should be privacy preservation. A privacy check should be performed with every sensing request and prior to the sensing result disclosure. A configurable privacy policy should be exposed to relevant actors, including a consent API for observed targets. Also, fallback mechanisms are needed, e.g., data anonymization, aggregation, deidentification, hashing etc.

7) Sensing service trustworthiness

Data use and retention assurance mechanisms and policies should extend to 3rd parties, e.g., impact of sensing results after delivery to its consumer. Impact should apply to combinations of data, e.g., although the sensing results does not reveal anything by itself, if combined with other data it may have an impact on a person's privacy. Data integrity and authenticity is core for critical applications and should be considered for both internal and external data as well as for raw measurements, semi-processed data and sensing results. Legally there will also be a need for data traceability; this is also useful to remediate accidental or intentional data leaks.

8) Application support

Sensing will support future applications. For example, with programmability new services can be created on top, and with applications controlling processing tasks there may be performance benefits, e.g., the application provides an AI model that is used by the network to detect desired conditions, or allowing an enterprise application determine data access control rules, etc.

7.4 Application- and Device-driven optimisation for BCS

The scope of the applications that will make use of beyond communication services will be broad and their components will span different domains. BCS can be leveraged to support various verticals, where specific QoS and QoE requirements must be met. However, these applications often face conflicting demands in terms of communication, sensing, computation, sustainability, and energy efficiency. Several applications leveraging

BCS, such as sensing, e.g., in the context of Industry 4.0 scenarios, such as Digital Twinning applications, will need to interact with the respective network components exposing such services and respective data. Also in some cases, the sensed objects are moving in a relatively high speed and several transmission/reception nodes are contributing to JCAS activity, while other use cases consider sensing low-speed objects but require strict delay performance to execute actions (e.g., stopping a robot machine after detecting a human). To meet the QoS of those applications (e.g., Digital Twin-related response times, sensing information latency, etc.), and considering the highly demanding computing requirements that may be involved, the placement of the respective application components is critical across the 6G compute continuum.

Considering the wide range of devices and applications that will utilize BCS such as JCAS, it becomes also imperative to establish a flexible approach for registration, as well as mobility management. Many devices and applications may have limitations that necessitate a dynamic and adaptable framework for seamless connectivity and mobility. By addressing these challenges and ensuring efficient registration and mobility handling, BCS such as JCAS can unlock its full potential and cater to diverse use cases in a wide array of industries.

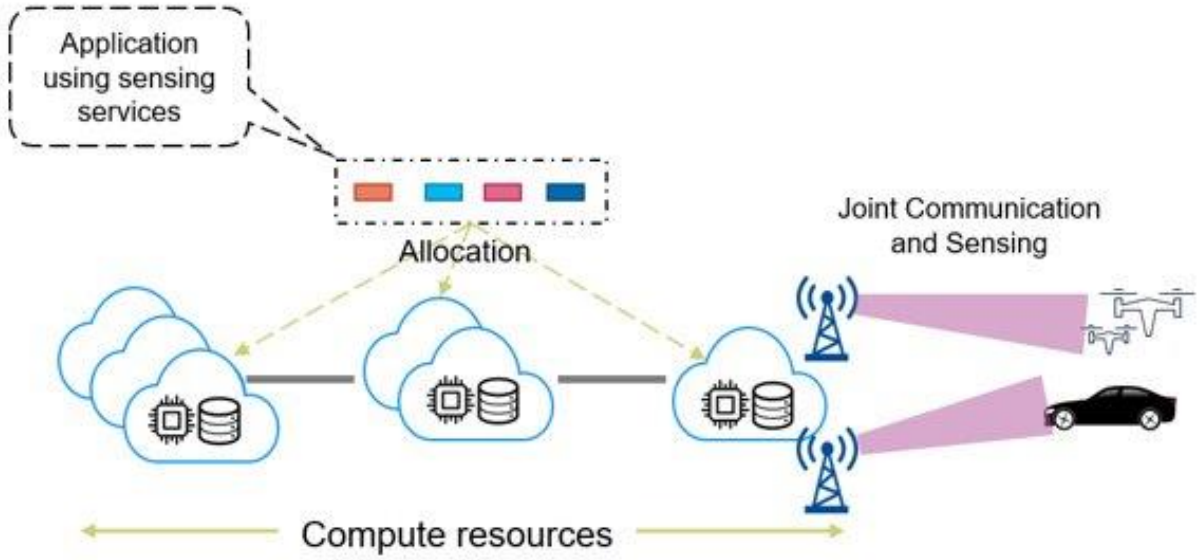

Figure 7-11 Overview of application- and device-driven optimisation for BCS.

The device needs to be registered in the network to communicate and use the applications. This registration is performed when entering the network e.g., when switching on the device, but also with periodic registrations and when the device has moved outside a given area. The initial registration and also periodic registrations are using many resources in both the radio, as well as on the end user side. Especially if sensing is performed by a sensor with very limited storage of energy this is not possible. Therefore, the network needs to be flexible to serve the wide range of devices supported by the network, both from connectivity as well as from a mobility point of view.

Application- and device-driven optimisation for BCS enablers, as illustrated in Figure 7-11 are investigated in two study areas, structured in the following sections.

Section 7.4.1 focuses on BCS information aspects, as well as functionality allocation and application placement challenges in Industry 4.0 scenarios, considering performance, privacy, and trust.
Section 7.4.2 discusses the requirements from the different type of devices that will be participating in the afore-discussed BCS-based use cases and describe how these devices will be managed in the network based on their needs and capabilities.

7.4.1 BCS information exposure and functionality allocation

As already discussed, towards the emergence of novel applications and services in 6G, beyond communications, data processing and insight extraction will become more complex; at the same time the authorized components and functions in the network to consume service data beyond communications will

need to be carefully assessed, prioritizing trust, privacy, and security. This will be important for ensuring data is protected and insights are generated safely and reliably. This study will focus on design considerations for efficient device information and network insights exposure, novel network functions to support data management in a trustworthy and energy-efficient manner, as well as challenges related to the functionality placement, related to both to application-agnostic processing components, as well as the various vertical application leveraging the respective BCS service and receive the BCS outputs/network insights, see Figure 7-12.

Efficient data management is crucial in the era of Industry 4.0, as vast amounts of data will be transmitted and processed by various services and vertical applications. The transmission frequency and data overhead must be optimized to support the ever-increasing connectivity demands and complex use cases, ensuring the stability and resilience of the system. By intelligently managing these factors, this solution will create a BCS data management scheme that balances the need for rapid and accurate information exchange with minimal overhead.

The challenge of the processing functionality placement by using intelligent algorithms to allocate processing capabilities across various components within the network is critical. These algorithms consider factors such as security, latency, and resource availability to determine the optimal locations for processing functions. Critical aspects also relate to the vertical applications that will make use of the BCS outputs/network insights. This approach ensures that applications can operate efficiently and effectively, benefiting from streamlined processing workflows and improved overall performance.

In the above, data privacy and security considerations will be key; to this end, the introduction of new network functions will be considered for addressing trust, security, and privacy constraints in data management and processing. Embedding these features within the proposed architectural enabler will help guarantee that data remains secure, and insights are derived in a manner that is not only efficient but also safe, reliable, and trustworthy.

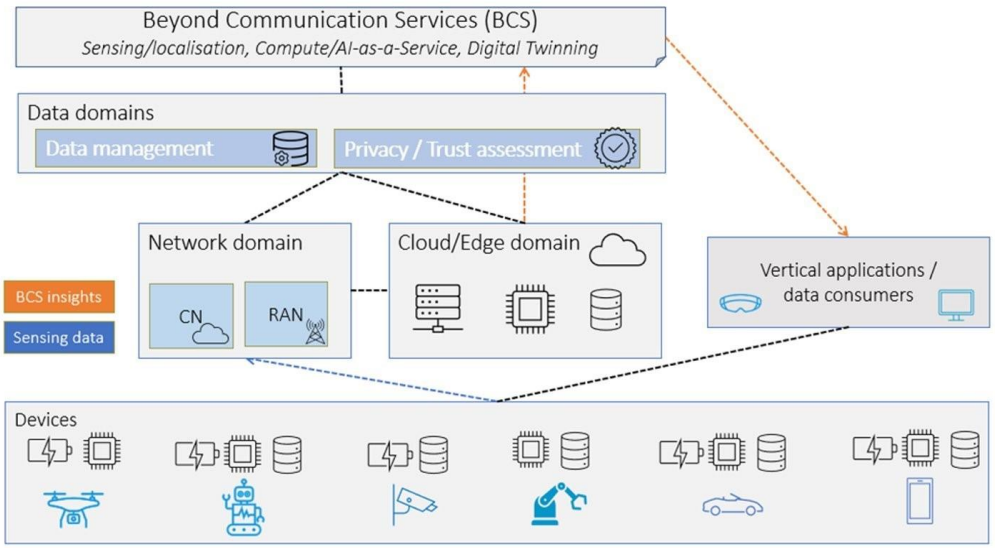

Figure 7-12 BCS data exposure and functionality allocation ensuring performance, privacy trust.

The proposed solution will focus on beyond communication service/application placement and optimization approaches driven by relevant 6G KPIs and KVIs, such as transmitted data overhead, processing time, trustworthiness, sustainability (energy/cost). By efficiently allocating computing resources and ensuring privacy, the BCS systems will offer advanced data management and processing capabilities in collaborative environments.

The scope of the proposed solution may include different BCSs such as sensing/localization, Compute/AI-as-a-service, and massive twinning. Vertical applications and interfaces with different data consumers are also in scope, resulting in a cohesive ecosystem.

7.4.2 New protocols supporting Ambient IoT devices

The next generation of networks needs to support not only connectivity for broadband devices but also an improved support for sensors like ambient and power limited devices etc. which may be static or mobile. Therefore, the architecture and the protocols need to be flexible for different usage of the connectivity provided by the 6G network.

The study will focus on how to solve the wide range of devices including applications using sensors supported by the network, both from connectivity point of view as well as from a mobility point of view.

One challenge is the connectivity ranging from broadband devices, using massive MIMO with ultra-low latency requirement to Edge servers, to small IoT devices with sensors without small or no battery. Any of these devices may be stationary and/or highly mobile. The network needs to be flexible in the management of these devices to support the requirements from each device.

In this study we aim to find the requirements of the different type of devices and describe how these devices will be managed in the network based on their needs and capabilities. This will require the network and the protocols to become much more flexible. The connection between the RAN service and the Core Network will be affected to simplify the communication between the sensing device and the application function. In Figure 7-13 the device periodically needs to update the status of the device to the network to the core network domain. This device needs to be connected during a relative long period for these updates and that is not possible for the power sensitive sensor device and it needs a more power efficient handling.

The radio resource management therefore needs to be flexible, handling both devices with no or low power storage, with a very low data rate which is sent infrequently, and the broadband devices with extremely high data rates. These different devices cannot be handled in the same way due to different service requirements but still work within the same network. Therefore, based on the requirements and capabilities this study aims to find solutions in the E2E system to handle this wide range of devices.

Based on these solutions the deployment of the network should also be easy to scale based on the supported services and handled UEs. A private network can e.g., be deployed to support sensors in a factory or in agriculture, this network should be possible to scale by e.g., not supporting some network functions which are only for high-speed broadband devices. The definitions of the network functions in Core Network as well as RAN needs to be evaluated to support the flexible management of the devices with different service requirements and capabilities in the network.

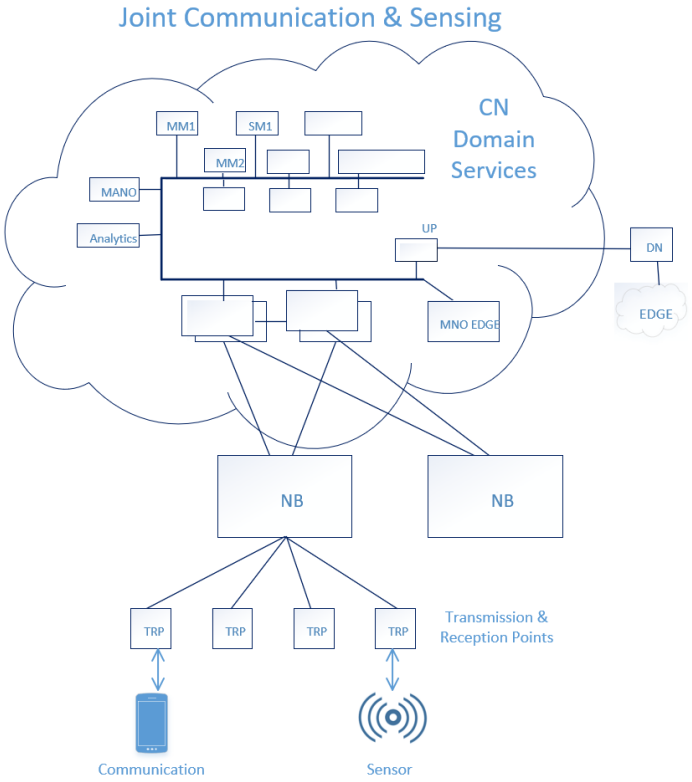

Figure 7-13 The E2E architecture to support a wide range of services.

7.5 Enhancing Joint Communication and Sensing Capabilities

The enhancement of communications networks with sensing capabilities is a very promising area that presents many opportunities and challenges. As shown on Figure 7-14, the main benefit of the communication networks in the context of sensing is that most of the infrastructure is already in place, which makes the sensing capabilities to be provided for free. The latest trends in wireless communication systems have been leaned towards providing more bandwidth, and with this, raised the need for higher carrier frequencies. This was noticeable in 5G and Wi-Fi, where mmWave bands around 24 - 28 GHz and 60 GHz respectively were considered to provide high bandwidths. The directional mmWave transmission and high temporal resolution from the multi-GHz bandwidth provide accurate localization (positioning). This localization potential stimulated the standardization bodies such as IEEE and 3GPP to propose standards improvements for the Next Generation of Positioning systems [CSL+23].

The Hexa-X project [HEXA] provided a conceptual overview of the novel use cases that will require this extreme localization performance together with several technical enablers and existing challenges [HEX-D31]. One of the identified use cases was precise and efficient simultaneous localization and mapping (SLAM) that helps to bring the digital world and the physical world together for quasi real-time interaction between the users that are located far apart. In addition, as one of the key technical enablers for fulfilling the requirements of this use case are the High-resolution Angle / Range processing at higher carrier frequencies that enables accurate directional sensing and imaging, while being less susceptible to ambient light and weather conditions. As Hexa-X, many other studies [WSL+21], [Hua22], identify SLAM as a mechanism enabling extreme localization performance that the 6G networks will bring, but none of them elaborates on functional solution.

The previous paragraphs support the increasing demand of sensing and the internal communication between sensors in the future communication network of 6G. The network communication overhead and ceaseless resource demand will significantly raise limitations for the classical technologies. For maintaining the QoS and QoE in future communication network, a shift in paradigm is necessary to limit the rising computational cost and energy requirements.

To this end in Section 7.5.1 we do a step forward studying the applicability of High-resolution Angle / Range processing at 60 GHz for performing SLAM with Commercial-off-the-shelf devices. The goal of this study will be to provide a functional solution that will use the sensing data obtained from the communication channel to build indoor maps. In addition, the massive, distributed communication and sensing can raise significant limitations for classical technologies because of communication overhead and usage of resources.

Integration of quantum technologies in the current communication network is the primary focus of the second study of the chapter (Section 7.5.2). Development of protocols to facilitate the necessary transition between quantum and classical communication is another essential aspect focused on the study. Enabling such a hybrid network will also enable inclusion of quantum sensing technologies in the network which can provide sensitivity that is infeasible classically.

Figure 7-14 Applicability of Joint Communication and Sensing in urban environments.

7.5.1 Indoor mapping using mmWave WiFi C&S

Simultaneous Localization and Mapping (SLAM) [DNC+01] is a technique that enables mobile robots to estimate their location in real-time and build a map of their environment while moving. The global SLAM technology market is expected to experience substantial growth from USD 226.7 million in 2021 to USD 9425.7 million by 2030, at a compound annual growth rate of 49.41%, as per a report by Straits Research [STR21] This growth is fuelled by the industrial sector's rising demand for autonomous mobile robots that can enhance productivity, logistics, and decrease production costs [FIP+20]. Currently, creating indoor maps through human-driven methods such as Google Maps Indoor, HERE Indoor maps, and Apple indoor maps is complicated and not suitable for industrial scenarios. An alternative approach is to utilize autonomous robots to create indoor maps, which incurs zero additional cost and holds great promise.

State-of-the-Art (SotA) mobile robot solutions currently rely on optical sensors like Lidars [SNH03] RGB cameras [DXN+15] or stereo cameras [HKH+14] to generate highly precise indoor maps [CTJ+18]. However, these optical sensors are generally not energy-efficient and require costly integration into mobile robots. Furthermore, optical sensors' performance is significantly hindered by environmental factors such as dust, fog, or smoke, and they may be less effective in indoor spaces with glass/mirror walls or inadequate lighting. Although manufacturing situations may involve airborne obscurants like dust or insufficient light, sight glass is commonly used in manufacturing equipment's panels, lenses, and covers. Some recent attempts [LRZ+20] to address Lidar's limitations have explored radar-based systems operating at millimeter-wave (mmWave) frequencies. However, these solutions are complicated to integrate and consume a significant amount of energy.

In recent years, mmWave technology has become the technology to cope with the increasingly growing demand for higher data rates. Mobile robots are also part of this trend, where recent applications such as inspection, and computer vison (CV)-based navigation requiring the delivery of high-quality videos at data rates (i.e., 400 Mbps-1.8 Gbps) [LCG+21], [MBF+22] are now unattainable through sub-6 GHz communication technology. The unique characteristics of mmWave frequency communication include high bandwidth, low latency and high directionality, which are necessary to manage propagation loss and unfavourable atmospheric absorption. These features demand a meticulous network design to ensure that mmWave communication technology is deployable on a large scale [FAC+19]. A key benefit of mmWave is the capability to JCAS into a single unified system that not only provides high-capacity wireless connectivity but also achieves high accuracy for localization [BMG+22] and sensing [PLR+22]. The high temporal resolution from the multi-GHz bandwidth is suitable for accurate distance estimations and the large number of elements in a directional mmWave antenna can enable very accurate signal angle estimations. These accurate estimates of the angle and the distance are the two main inputs that are required for building a discrete set of data points in space (Point Cloud) needed for indoor mapping.

Despite these advantages, mmWave-based indoor mapping with Commercial-Off-The-Shelf (COTS) devices is still unexplored. The map generation process (see Figure 7-15) requires high-resolution Point Cloud estimations (i.e., a huge collection of individual points) from the surrounding environment.

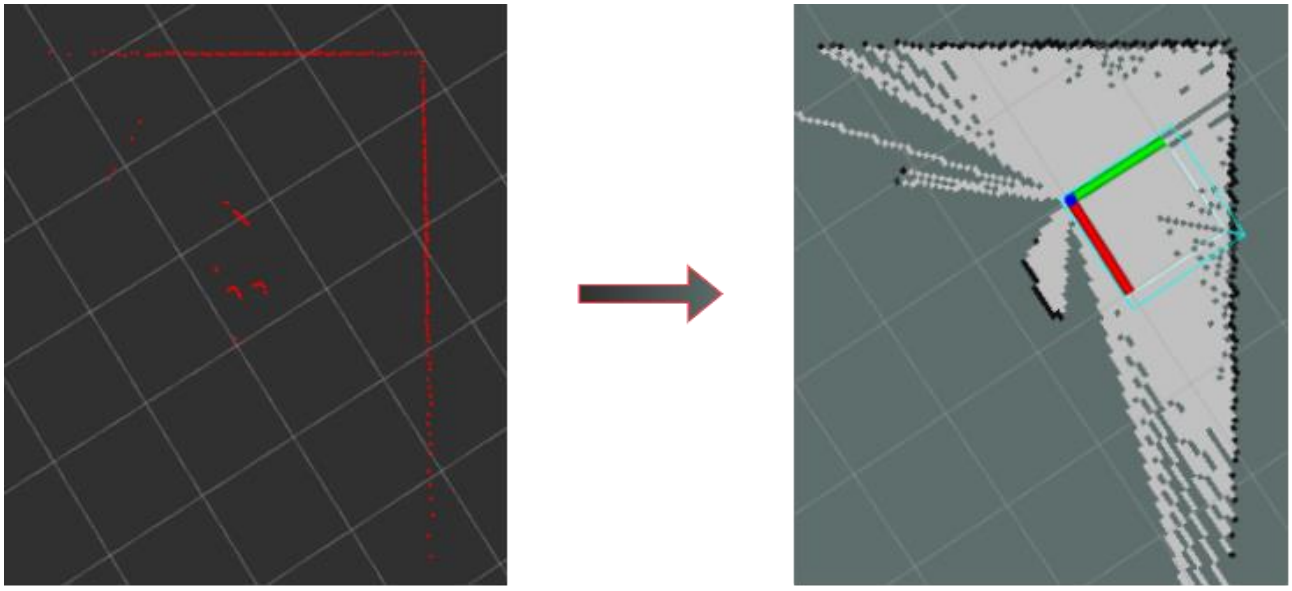

Figure 7-15: The obtained indoor map (right) from Lidar individual points (left).

The main challenge lies in the fact that for communication in mmWave WLAN, we need directional connectivity between the pair of communicating devices. The angle and distance estimations that can be obtained from the mmWave communication channel can be useful for the real-time localization and navigation of a mobile robot but cannot be used for building the Point Cloud estimations needed for indoor mapping. To this extent, this study will investigate the feasibility of achieving indoor mapping with COTS devices for NextG network-assisted mobile robots.

7.5.2 Quantum-enhanced 6G Communication and Sensing

Massive, distributed communication and sensing can raise significant limitations for classical technologies because of communication overhead and computational complexity. As classical technologies have their infeasibilities, integration of quantum communication can enhance the capabilities of 6G beyond communication networks.

SDN and NFV have provided flexibility and efficiency and have opened pathways for accommodating many other technological advances. The paradigm has changed from store-and-forward process to compute-and-forward which enables usage of general-purpose hardware instead of dedicated ones. Though, due the classical limits and increasing computational complexity demanded by future generation networks, a radical change is necessary to alleviate from this situation. Even though softwarisation of network has its own setbacks, it

enables the accommodation of newer technologies rather conveniently because of its versatility. In that context, integration of quantum mechanical principles in the current classical architecture can enhance the issues of computing and security, harnessing quantum entanglement.

The envisioned 6G network thereby would be a hybrid classical-quantum network, where quantum virtual machines will hold entanglements and qubits for its usage in the network (see Figure 7-16). Since qubits themselves do not possess any header, separate control signals via classical channels must be sent so that the repeater can correlate the qubits stored in their memory to the qubit that is required. This way, both classical and quantum aspects can be fused together for future generation of networks to efficiently use the unique quantum properties.

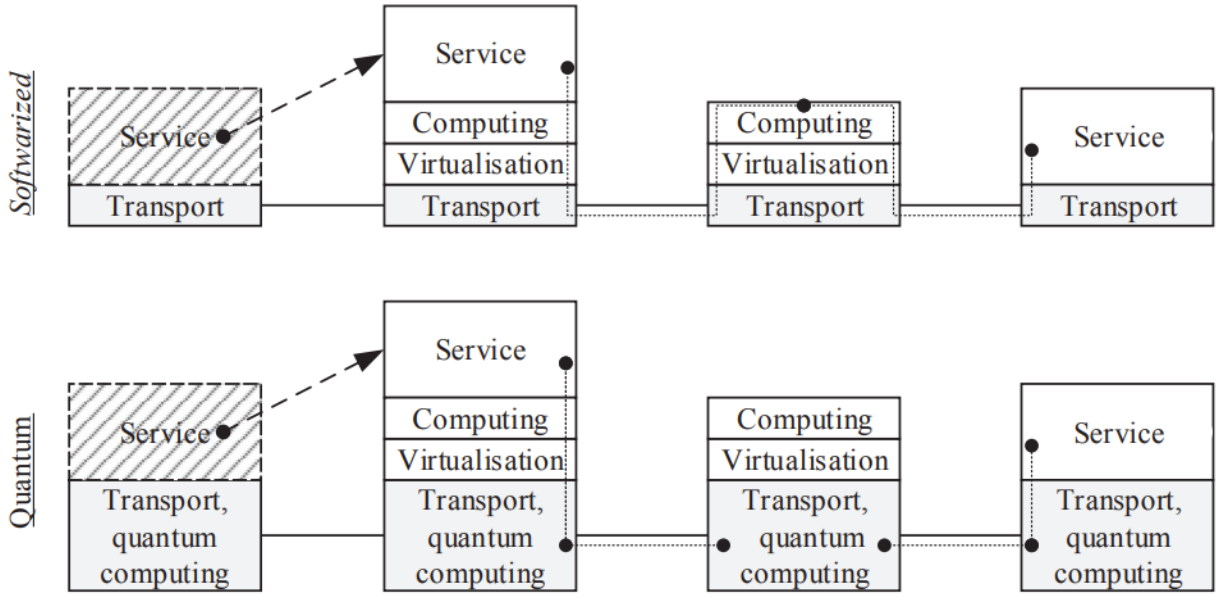

Figure 7-16: Different paradigms of softwarized versus Quantum communication networks.

As displayed in Figure 7-16, the quantum part resides in the physical layer of the protocol stack which would be controlled by a classical interface. Furthermore, the actual IoT industry has evolved to optimize the operation of many productive processes based on the use of massive networks of sensors. Quantum technologies may also find their use as quantum sensors offering a higher measurement sensitivity and precision than their classical counterparts, because quantum states are inherently susceptible to environmental changes which can be further recorded and analysed upon.

For this, the quantum phenomena such as entanglement is exploited, syncing several quantum sensors as one device. Thus, quantum-enhanced sensing will be explored to define their integration into IoT architectures, potentially achieving an overall optimization in many industries [DFP17].

It has already been shown that quantum sensing can boost the measurement sensitivity by exploiting the high sensitivity that quantum systems inherently offer [HST+22].

With this objective, the aim is to tackle the issues of latency and security by integrating quantum principles in the current classical communication architecture. In this study we will attempt to realize such a hybrid quantum-classical system and will analyse its characteristics using simulation-emulation platforms. Enabling various quantum protocols for the hybrid system is another major focus of our study to improve latency and security aspects.

8 Virtualisation and cloud continuum transformation

In recent years, cloud computing became the de-facto standard for managing web-based and web-scale applications. While this architectural paradigm is suitable for a big subset of multimedia human-scale applications, it shows its limitations when it comes down to supporting the upcoming latency sensitive 6G use cases (e.g., industrial automation, holographic telepresence, eHealth). More specifically, the cloud-based architectures have very well-known limitations when it comes down to latency, throughput, connectivity and security and interoperability [HW10]. To address these limitations, in the past years Edge and Internet of Things (IoT)[2] computing has arisen as a paradigm that aims to provide compute, storage and networking capabilities in near proximity of the end-users, while providing the same pay-as-you-go model of Cloud computing. While edge computing enables application developers and content providers to leverage Cloud computing capabilities and an IT service environment at the edge of the network, IoT computing distributes resources and services across Cloud, Edge, and devices on the field to create the so called IoT-Edge-Cloud continuum. With these initiatives the Cloud evolution towards the edge of the network begin and in this section the main enablers with their representative studies are identified that will facilitate the Cloud transformation. In particular, Section 8.1 elaborates on the integration and orchestration of IoT-Edge-Cloud continuum resources into a single 6G architecture. Section 8.1.3 addresses the multi-cloud federation challenges aiming at addressing the interoperability constrains of cloud computing. In Section 8.3 the network function placement in the end-to-end resource continuum is discussed. Finally, Section 8.4 tackles quantum computing and how it will influence the ongoing cloud transformation.

8.1 Integration and orchestration of cloud continuum resources

Cloud computing has been proven to be effective in managing latency tolerant applications and network functions, but it falls short in latency sensitive operations. This limitation is mainly due to the lack of control over the connectivity between the Cloud and the end-users, which spans across different service providers (SPs). Such kind of latency sensitive applications and underlying network functions requires responsiveness at very short time-scales while todays Cloud computing functions mostly require human-scale responsiveness. To address this connectivity limitations, in the past years, Edge computing tried to use the same elasticity and pricing model of Cloud computing (pay-as-you-go) and apply it in the border of the network. Compute, network and storage capabilities now are available closer to the user. This trend led to the birth of different initiatives such as:

- From the Telco industry: Multi-access Edge Computing (MEC) [MEC003] enables cloud-native functions to be executed at the network's edge.
- From the IT industry: hybrid-cloud that is envisioned as a way to manage on-premises infrastructure and public clouds allowing data and functions to be shared between them.
- From the manufacturing industry: Fog computing [FOG+18] that is envisioned as an extension of the Industrial IoT (IIoT) distributes resources and services across Cloud, Edge, and extreme Edge (xEdge) resources to create a cloud-to-thing continuum.

Although these approaches scope the Edge computing concept differently, their ultimate goals are the same, to provide reduced latency, bounded jitter and improved overall Quality-of-Service (QoS) by bringing computing, storage, and networking closer to the end-users. To accommodate this trend, the computing continuum is moving towards a more decentralized infrastructure in which computing resources with different capabilities and characteristics can host functions, applications and services. As shown in Figure 8-1 this will ultimately result in a distributed system that needs to be properly integrated, orchestrated and managed so it can become like the familiar cloud model.

[2] Note that in the literature exists terms such as Internet of Things (IoT), Fog or Extreme Edge (xEdge) computing that refer to the same concept of including the end devices as part of the computing infrastructure.

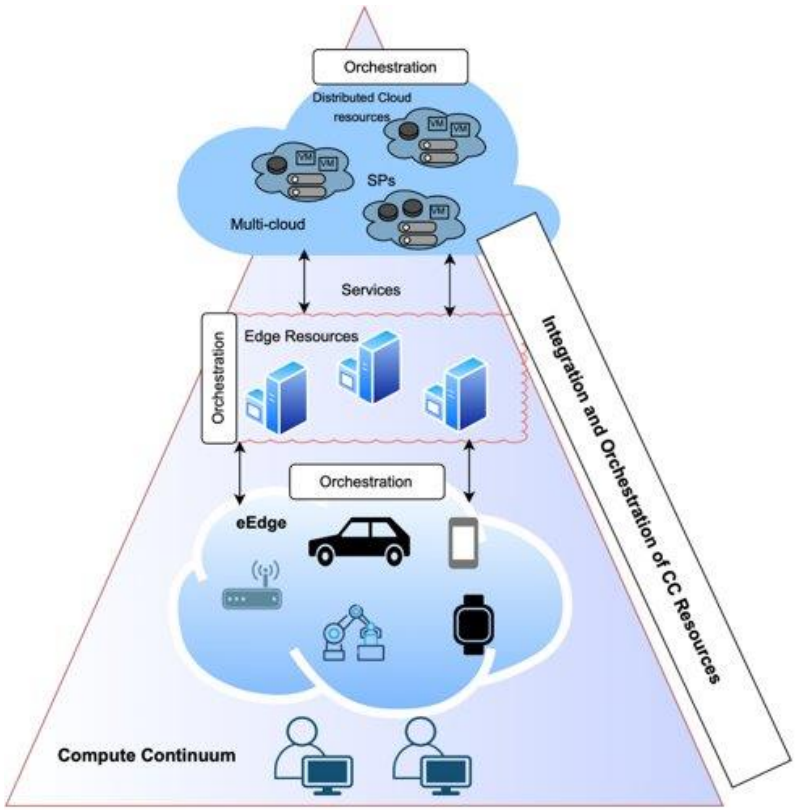

Figure 8-1 The 3-layer compute continuum: xEdge, Edge and Cloud resources that are part of the Compute Continuum (CC).

In Section 8.1.1, the study area "ETSI MEC deployment in constrained devices" aims to address some of the main existing challenges of the ETSI MEC framework by proposing the constrained MEC (cMEC) version that integrates constrained end-user devices into the computing continuum.

The study area "continuum management and orchestration", detailed in Section 8.1.2, aims to propose a Multi-Technology Resource Orchestrator to manage and orchestrate over the resources in the continuum (i.e., Extreme Edge, Edge and Cloud resources). Ad-hoc strategies for allocation and migration of resources will be addressed and particular focus will be given to the development of new mechanisms to handle the volatility and dynamicity of Extreme Edge resources (e.g., prediction algorithms to foresee the evolution in time of dynamic constraints).

In Section 8.1.3, the study area "Compute continuum Smart Management" aims to cope with extreme-edge management requirements in an end-to-end approach from cloud to edge to extreme edge, with special focus on potentially new mechanisms to predict the asynchronous and volatile nature of the end devices.

8.1.1 Extensions of ETSI MEC framework in constrained devices

Multi-access Edge Computing (MEC) is currently the primary standardized framework in the field of edge computing, which has introduced a whole new set of network services and applications. With its undeniable benefits, MEC is being developed by the European Telecommunications Standards Institute (ETSI) and is being developed as a technology that has the potential to meet the core Key Performance Indicators (KPIs) of 5G [KFF+18] and beyond. Together with other edge computing paradigms such as fog computing [IFB+18] and cloudlet computing [BKA+21] MEC aims to reduce latency and the workload on cloud infrastructure, ultimately improving communication latency and bandwidth utilization. This technology provides significant benefits for the families of use cases targeted by 6G technologies, including Joint Sensing and Communication (JSAC), Energy-harvesting and low-power operations, and Ultra-Reliable Low-Latency Communication (URLLC).

Upcoming applications, such as the next-generation of highly-distributed applications (e.g., edge robotics, augmented environments, or smart agriculture) have even more stringent requirements. Therefore, relying

solely on deploying MEC servers at the telecommunication network edge may not be sufficient. There are already scenarios where the MEC framework showed limitations:

- **Loss of connectivity**. While on-the-move, devices might temporally lose their connectivity. Consequently, applications supported by a MEC server cannot guarantee service continuity. Although application relocation mechanisms exist, they either assume that the MEC infrastructure is deployed everywhere or that there are deployments in aggregation points of the infrastructure, making delays so large that the edge benefits are minimized.
- **Near-zero latency applications**. Computation offloading to an edge server might also be inadequate whenever applications require extremely low latency (i.e., sub-1ms robotics control loop). In addition, fluctuations in the communication would likely introduce undesirable jitter.
- **Privacy and security**. MEC is part of a multi-domain ecosystem composed by several stakeholders (e.g., infrastructure owners, service providers, system integrators and application developers) [SRN+21] thus placing generated data outside of the owner's domain. Although data privacy and security can be enforced by its owner, offloading functions to a MEC server increases the risk of a data leak or unauthorized access by a third-party [ZCZ+18].

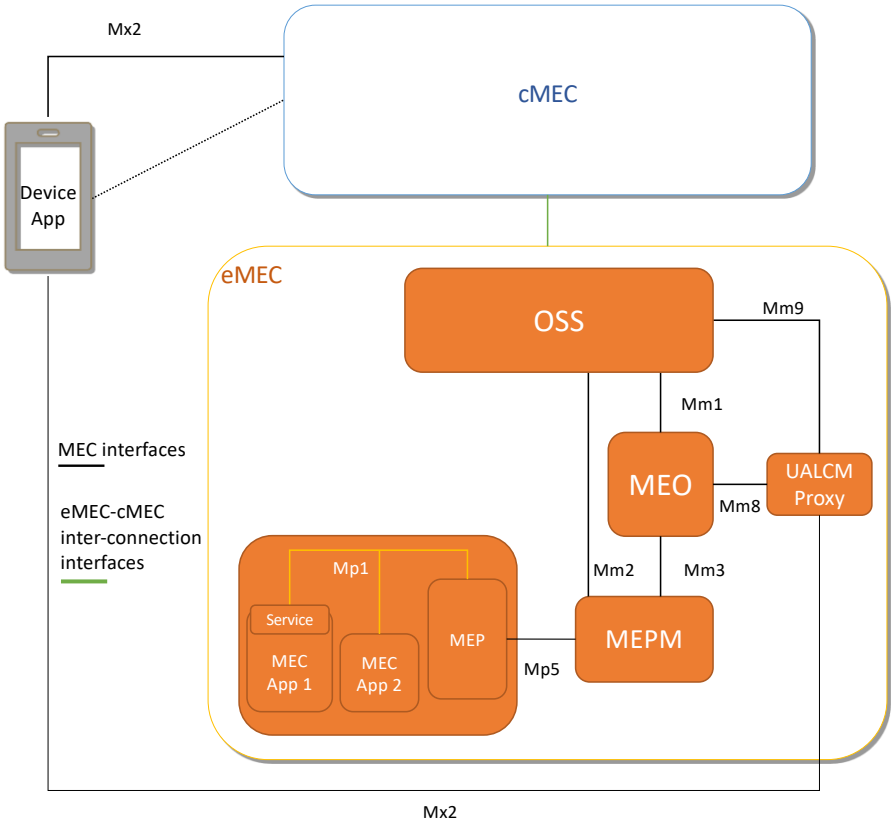

Figure 8-2 Architectural scheme of constrained Multi-access Edge Computing.

The aforementioned challenges can be mitigated by exploiting dynamic computational offloading techniques. Additionally, integrating MEC platforms towards end-devices or constrained devices in the close vicinity of end-users is currently the subject of study in ETSI GR MEC 036 [CONS21] also devised by other Standards Development Organizations (SDOs), such as IETF [IETF23]. A standardized method for integrating computation at constrained devices and traditional MEC servers, where the former preserves only subset of MEC capabilities, enables a holistic computational offloading while allowing resource orchestration at a finer granularity and exploitation of MEC services.

This study will contribute towards such a vison of integrated end-to-end cloud-to-thing continuum by proposing the constrained MEC (cMEC) architecture (see Figure 8-2), as a lightweight design of the MEC framework. By constrained device, this study refers to mobile end-devices or computational constrained mobile devices in the close locality of the end-users. cMEC considers that constrained devices can on-board

and support a subset of MEC functional elements to expand the computational reach of current MEC framework. MEC applications can then run locally and/or in a remote telco MEC system. In doing so, cMEC can take over on the applications execution whenever the connectivity to the network cannot be sustained, whether due to outage, mobility, or to incomplete coverage, and when the latency towards the edge MEC system is unreliable. In contrast, eMEC in Figure 8-2, corresponds to the standard telco MEC.

8.1.2 Management of continuum resources for E2E service orchestration

Continuum Management & Continuum Orchestration capabilities, provided through a dedicated Multi-Technology Resource Orchestrator (Continuum Multi-Technology Management and Orchestration Platform – Continuum-MT-M&O Platform), can be leveraged by Verticals and Service providers to manage and discover different kind of resources placed in the continuum, i.e., Extreme Edge, Edge and Cloud, and perform orchestration operations over them. Moreover, the proliferation of Extreme Edge nodes with embedded computing capacity and programmability can be exploited to run distributed applications closer to the users and data sources with potential gains in terms of energy efficiency by the means of ad-hoc developed resource allocation and migration strategies.

Figure 8-3 depicts the high-level functional components of the Continuum-MT-M&O Platform highlighting the components, with numbered blue circles, that provide the innovative management and orchestration functionalities over the continuum and represent an evolution in respect to the Resource Orchestration Platform that has been designed and developed in Hexa-X [HEX-D63].

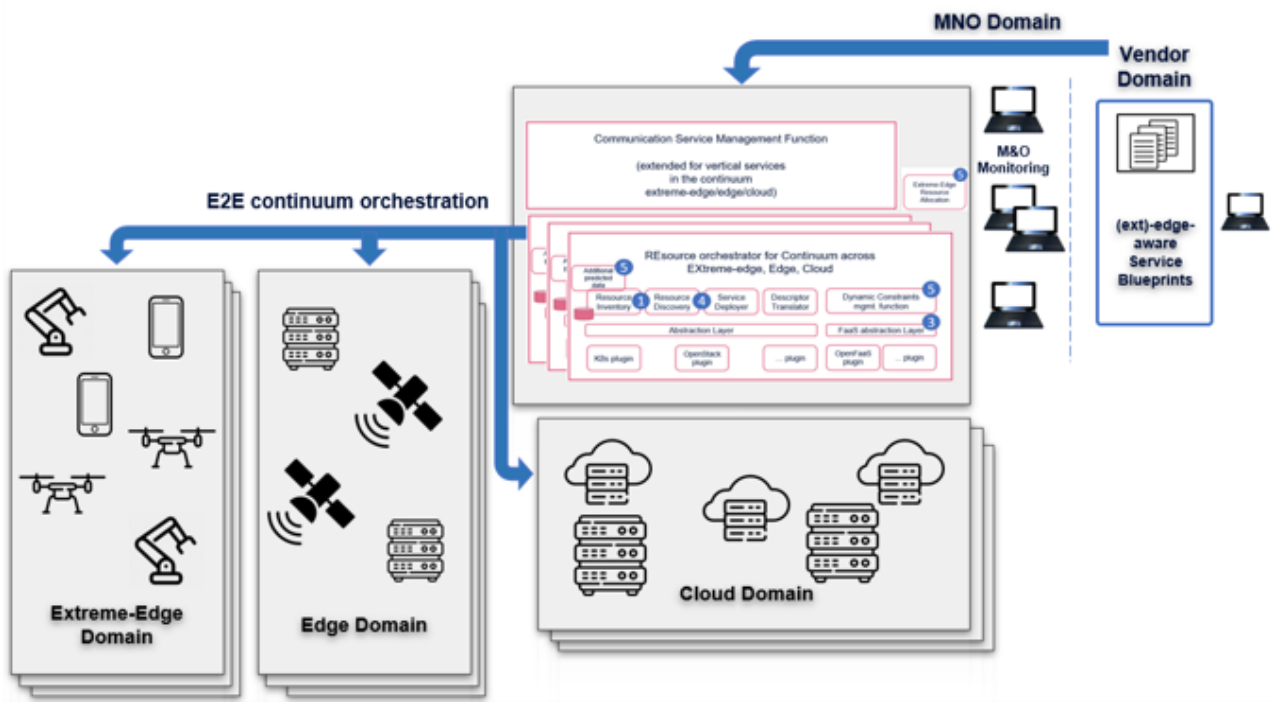

Figure 8-3 Continuum Multi-Technology Management and Orchestration Platform: Continuum Management & Orchestration through Communication Service Management Function (CSMF) and REsource orchestrator for Continuum across EXtreme-edge, Edge, Cloud (REC-EXEC).

For sake of visualization, the components of the Continuum-MT-M&O, labelled with blue circles, are detailed in Figure 8-4. The proposed Continuum-MT-M&O Platform Architecture can result in a consistent way of managing and inventorizing Continuum Resources from multiple administrative domains as well as enable the possibility of seamless orchestration operations with a special focus on the extreme edge where ad-hoc strategies for allocation and migration of resources will be developed.

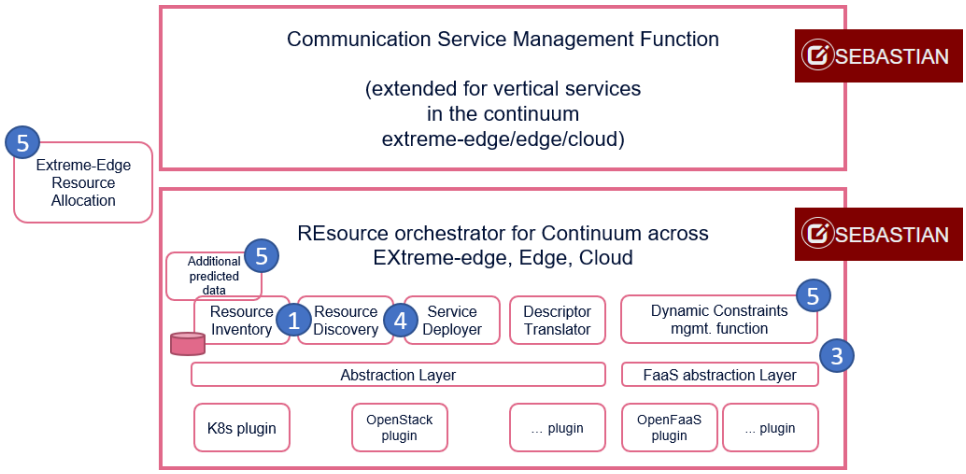

Figure 8-4 Communication Service Management Function (CSMF) and REsource orchestrator for Continuum across EXtreme-edge, Edge, Cloud (REC-EXEC) details.

The development of a Continuum-MT-M&O Platform for Resource Continuum allows the dynamic discovery and continuous monitoring of different kinds of extreme edge, edge and cloud nodes at the resource orchestration level. Thus, the high volatility of extreme edge resources can be handled and monitored continuously. These functionalities can be covered by the Resource Inventory and Resource Discovery functionalities (1) of the Continuum-MT-M&O Platform.

Federation mechanisms and abstraction models for pooling resources from multiple cloud platforms, potentially owned and controlled by different stakeholders, leveraging different virtualization technologies can be introduced through functionalities offered by the Continuum-MT-M&O Platform along with access control and disclosure policies, exposure of APIs and control delegation, etc.

Integration of serverless computing in end-to-end service orchestration across the continuum can be achieved through the introduction of a Function as a Service (FaaS) abstraction layer (3) over the service deployment functional components of the Continuum-MT-M&O Platform. The Abstraction Layers will enforce the creation of an abstract view of the continuum resources.

The definition of strategies to address mobile network connectivity implications on the orchestration of resources can be tackled by the Continuum-MT-M&O Platform by the introduction of dedicated Resource Discovery and Service Deployment mechanisms (4).

Resource allocation and migration strategies based on extreme edge nodes' characteristics constraints will be developed through the definition of prediction algorithms to foresee the evolution in time of dynamic constraints (e.g., battery level, energy consumption, availability and quality of the connectivity of the node, computing load from concurrent –user- applications, etc.); the allocation and migration strategies could be introduced through dedicated resource allocation and management functionalities (5) of the Continuum-MT-M&O Platform.

The relevant use case for this study is the Orchestration of the extreme edge across the compute continuum.

8.1.3 Decentralised compute-continuum smart management

Services orchestration systems in the future 6G networks will have to face a new challenge in what regards the integration of those resources in the extreme-edge domain. As already introduced in [HEX-D62], those extreme-edge resources are envisaged to be, not only small-scale input/output devices receiving information or sending certain data (such as the legacy end-user mobile devices, or the IoT sensors in previous mobile generations), but a wide range of devices with a considerable capacity for information processing and data storage. Indeed, small-scale devices like those in previous generations may still be available, but there will be a variety of devices with valuable computing, storage, and communications capacity (e.g., robots, on-boarded infrastructure in vehicles, domestic appliances, industrial equipment, smart-city devices, etc.). As anticipated in [HEX-D62], that additional pool of resources could be used to deploy certain network service components on them, which would be orchestrated together with other components belonging to the same service chain in

the core and the edge network domains. This would be the so-called "device-edge-cloud" continuum orchestration concept, already introduced in the previous Hexa-X (I) project [HEXA].

However, the integration of this new extreme-edge domain brings some significant challenges in what regards the services orchestration on it, e.g., regarding the scale of this domain (the extreme-edge is much larger in scale than the MNO domain itself), the high diversity of devices on it, which additionally, won't be deployed in well-controlled premises (like those in the facilities of a typical MNO), the volatility of those devices, which could move, and that could be unexpectedly switched on/off, as well as the ownership of the devices, which could belong to different and diverse stakeholders (e.g., vertical industries, different MNOs, hyperscalers, public institutions, or even end-users). These and other factors could lead to a very high level of complexity in the administration, monitoring, and configuration of that large number of resources, as well as the network services running on them, contributing also to increase the operational costs for the MNOs.
The common approach to network services orchestration proposed for the previous fifth generation has typically consisted of MNO-centric M&O frameworks (typically based on the ETSI NFV MANO specification [MAN004]), meaning that the services M&O problem was typically addressed from the perspective of a single MNO, assuming that the ownership of the network was completely in the hands of the mobile operator. It is true that, although multi-domain or federation models involving different MNOs or other different stakeholders have been proposed during the development of the 5G technology [MEC003] [TAS+19], the problem was still usually approached from the perspective of an MNO, i.e., trying to solve the problem of integrating different domains (within or without the MNO scope) for a given operator [SLR+19] [BCC19] [CAV22].

Towards 6G, an initial approach to address the services orchestration problem has been to still rely on that MNO-centric approach in the M&O architecture, but extending it with a hierarchy of orchestrators in a multi-domain approach, similar to the one already in [KNE16], but considering the extreme-edge just as an additional domain (although still within the MNO scope), and having several specific orchestrators for the different domains (e.g., extreme-edge, edge, and cloud), while another orchestrator on top of them provides the E2E functionalities throughout the entire network continuum, as well as interfaces to other external network domains [HEX-D63][MAT+23].

However, here we would like to propose a different approach, which is considered to be more flexible and practical in accordance with the high heterogeneity, size, and dynamicity of the extreme-edge resources. This new approach consists in delegating most of the services M&O mechanisms on the network services themselves, thus providing a more autonomic and decentralised approach. This approach would provide the following broad benefits:

- Each service could adapt its M&O resources locally depending on its specific context and needs. Certain services may not require the same level of complexity in the orchestration than others (e.g., in what regards data monitoring, use of AI/ML techniques, etc.), being the case, that certain services may require even very simple orchestration primitives. Adapting the M&O mechanisms to the specific service needs can simplify orchestration processes, and help reducing complexity and operational costs for the MNOs.
- Being more decentralised this approach can be more resilient to failures, as a problem in one specific service would not necessarily affect other services. On the other hand, the typical MNO-centric orchestrator could become a bottleneck, or even a single point of failure.
- This distributed approach can be easier to scale as the network grows including new services, the associated M&O mechanisms can come hand in hand with those new services, and tailored to the specific needs that these new services may require. The need to align these new services with the specific requirements and functionalities of an MNO-centric orchestrator (which may become obsolete) is minor, since the decentralised approach might be more effective in fostering ongoing innovations, as concentrating control within a single entity could hinder new concepts and developments.
- Higher adaptability: Distributed orchestration can more easily adapt to changes in network topology or unforeseen conditions, which is especially relevant regarding the extreme-edge, as services can make autonomous decisions in real time without relying heavily on a central point.
- Simplified integration of software components from different stakeholders to compose network services: stakeholders might define specific usage policies for their components, which, for example, could be

materialized through exposed APIs. SLAs would affect only those stakeholders needing to connect their components each other, and only for those components needed to be deployed for a particular service. The MNO-centric approach, however, typically assumes multi-domain M&O mechanisms at a general level, i.e., for the entire M&O framework itself, which may not always be feasible to all stakeholders. This is particularly relevant given that services composition on the device-edge-cloud continuum would be done with components not always owned by a single MNO, but from different stakeholders.

- Increased Security: Distributed services M&O can be more secure, as there is no single central point or rules to orchestrate the services. In an MNO-centric orchestration approach, if the central M&O framework were compromised, that could put the entire services orchestration system at risk. Also, the distribution and customisation of decision making can help minimising risks that could generally affect the entire ecosystem.

As a whole, the decentralised services M&O approach presented here relies on four main principles: (i) to rely on a fully cloud-native microservices-based approach, based on services composed of micro-services (this is aligned with the main architectural design principle in Hexa-X-I [HEX-D14][HEX-D62]); (ii) a specific information model for the resources orchestration, targeting not only the cloud and edge resources, but also the special features of the extreme-edge domain; (iii) as mentioned, to delegate the M&O of the services business logic, as well as the service components life-cycle management, to the network services themselves; and (iv), a set with four new network elements for orchestrating the network resources and to deploy the network services on them. They are the following:

- The Deployment Node (DN) component. This would be the entry point to the network for deploying service components, which would be deployed using a declarative intent-based approach, i.e., by specifying just the desired final result regarding the deployment, but without needing to specify how to achieve that result. Multiple DNs would be distributed through the entire network, hosted by MNOs, Hyperscalers, Vertical Industries, or other stakeholders. In turn, different approved stakeholders could access these nodes to request the deployment of "their" network services. If the service were successfully deployed, the DN would return a "handler" to the stakeholder that requested the deployment. That handler would later be used to perform M&O operations on the deployed service, using the service-specific M&O mechanisms, which would be deployed together with the network service itself.
- The Infrastructure Registry Service (IRS). This would be a distributed database containing updated information about the available infrastructure components (e.g., device type, IP addresses, owner, network domain on which it is deployed, reachable networks, available computing and storage resources, etc.). Registered infrastructure components could be both: physical and/or virtual components. The IRS would be accessed by the DN to select the specific infrastructure components (devices) on which any service could be deployed. Initially, the information for each device would be provisioned upon the attachment of the devices to the network (this device attachment process could be manual or automatic). However, due the intrinsic volatility of the extreme-edge domain, this information would be updated on a near real-time basis (specific closed control loop processes could be used for that).
- The Services Registry Service (SRS). This would be another distributed database containing updated information on the current execution environment for the deployed services. This service would provide the translation between the handler provided to the service owners during its deployment, and the actual endpoints available to access the service to perform management operations on it (it should be considered that service components could be re-located during their lifecycle, due to the extreme-edge volatility, or because of the service business logic itself).
- The Infrastructure Status Prediction Module (ISPM). This would be an optional component intended to help dealing with the extreme-edge dynamicity and volatility. It would provide information on which devices might be available/unavailable in the near future, based on predictions. This information could be of use during the initial deployment of the services (e.g., communicating to the DN the forecast of the change of state of a device close in time), but also, when network services were already running (e.g., based on the device state predictions, alerts could be generated towards the services to enable proactive adaptive behaviours). The devices forecasted information could be generated based on data analytics, which could be based on AI/ML techniques. ISPMs could be deployed in a per-service basis, but also as an overall network element able to combine information from multiple network devices and domains in order to make more accurate predictions.

Figure 8-5 shows a sequence diagram illustrating a simplified service deployment process involving the above-mentioned components.

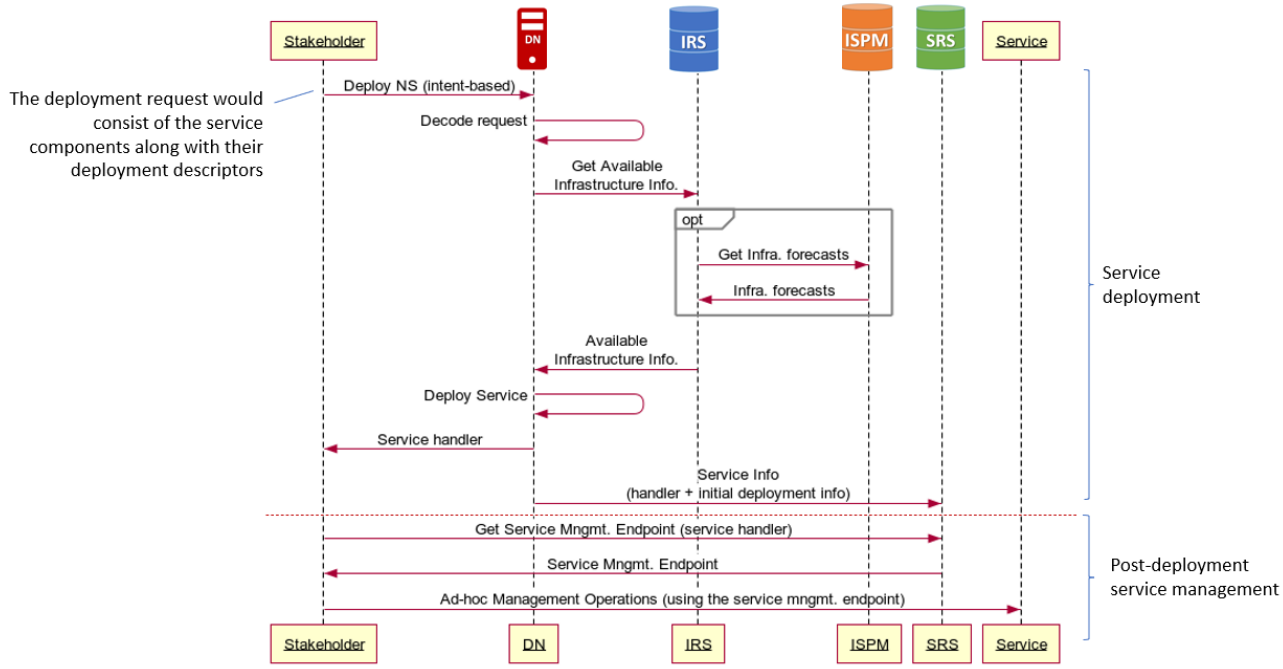

Figure 8-5 Simplified distributed compute-continuum smart management process.

As it can be seen, once deployed, the service orchestration mechanisms would be delegated on each deployed service itself, and according to its specific needs and M&O resources. Those service orchestration mechanisms could include the service components life-cycle management operations, as well as the service business logic orchestration mechanisms (e.g., the migration of the service components among the different infrastructure nodes, the implementation of scaling and elasticity mechanisms, high availability mechanisms, etc). All this could be implemented in very different ways, tailored to the requirements of each service and according to the needs and capabilities of the service stakeholders. E.g., specific implementations could be based on containers orchestrators already in the SotA (e.g., [K8S], [SWARM], [NOMAD]), ad-hoc orchestration systems specifically developed for the services (e.g., in case a certain stakeholder already could provide that), or even through microservices choreographies (i.e., without relying on any specific orchestration component or framework) [CDT18]. Regarding the management responsibility, for those services that may require it, they could be managed by one or multiple stakeholders independently, and with or without the need of the MNO to be involved (which could help to reduce operational costs to the MNO). Nevertheless, in this open ecosystem, MNOs, like any other stakeholder, would also participate in different ways, e.g., by providing access to certain network functions within its own domain, by deploying its own services for third parties, or by implementing specific network management services for itself or other parties.

8.2 Multi-domain/Multi-cloud federation

Multi-cloud/Multi-domain refers to the use of multiple cloud computing platforms that can be provided by multiple providers to meet an organization's needs. Rather than relying on a single cloud provider, organizations adopt a multi-cloud strategy to leverage the strengths and offerings of different cloud platforms, such as public clouds, private clouds, or hybrid clouds. As illustrated in Figure 8-6, organizations distribute their workloads, applications, and data across multiple cloud providers, enabling them to take advantage of diverse services, pricing models, geographical locations, and specialized capabilities. To effectively use and manage a multi-cloud environment, organizations utilize cloud management tools, orchestration frameworks, and automation solutions that provide centralized control and visibility across different cloud platforms. These tools help with workload deployment, monitoring, performance optimization, cost management, security, and governance.

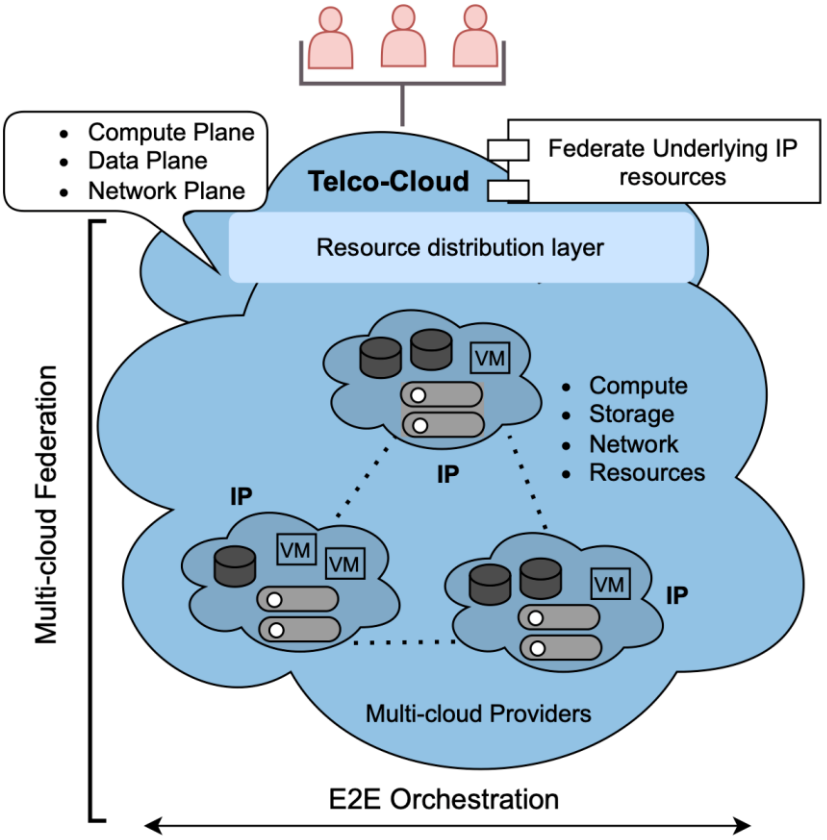

Figure 8-6 Multi-domain/Multi-cloud federation and orchestration.

However, managing such a multi-cloud environment today also introduces several challenges. Organizations, including telcos, must carefully plan their multi-cloud strategy, consider factors like workload placement, data transfer costs, and data interoperability, and establish appropriate governance and security measures to ensure a successful multi-cloud implementation. Complexity in network connectivity, data integration, security across different cloud providers (trust), and skill requirements for managing diverse cloud platforms can introduce issues that must be properly addressed.

In addition, the multi-cloud requires the integration of Infrastructure data centres of multiple providers and their exposure in a uniform manner to orchestrators. Such an approach needs the definition of business interfaces towards Infrastructure Providers, as ETSI NFV assumes the use of private IaaS only.

Finally, as organizations continue to adopt hybrid cloud architectures, the integration of multiple public and private clouds will become increasingly important. Federating these public and private clouds resources, also between different operators of different nations, will enable efficient data processing, reduce latency and enhance overall performance, allowing federated operators to create a coherent cloud layer that spans over national boundaries and can be offered to third parties. Moreover, such uniform cloud capabilities can be enhanced by the network enablers provided by the operators. A resource distribution layer is needed to manage and coordinate the resources, across federated environments. The orchestrator acts as a centralized control entity that enables seamless integration, interoperability, and efficient utilization of resources.

To address some of the above challenges, Section 8.2.2 elaborates on how we can use multi-domain federation at data centres and how it will work and the benefits of using it.

In Section 8.2.3 we present a study that addresses the integration of multi-providers data centers infrastructure and its exposure by introducing a Resource Layer which consists of many functions solely focused on resources. The presented approach is in line with the Cloud Continuum concept and exposes integrated resources to applications uniformly.

Section 8.2.3 will analyse the challenges that arise in a multi-cloud environment, highlighting the need of an orchestrator layer that hides complexities linked to the operator infrastructure from the high-level requirements. The study will mostly consider federation scenarios among different operators separated by

national boundaries extending the concept of cloud continuum and evaluating how productions needs can be achieved in those scenarios

8.2.1 Multi-domain federation in data centres

Requirements for latency critical or data privacy/sovereignty can require applications and services to run closer to the data producers, while still accessing more central services that do not have such strict latency or privacy requirements. While the cloud environment is relatively stable and resourceful, the near edge and particular the far edge pose challenges. At the edge and far-edge, infrastructure is much more heterogenous, and resources constrained when compared to central data centres, see Figure 8-7.

By virtualising and federating resources at the far-edge, managing these as a pool, it is possible to replicate (to a much lesser degree) some of the elasticity that cloud environments provide and have edge nodes running neighbouring workloads. Workloads that are not time critical can be offloaded to the cloud, assuring an efficient resource management at the edge.

The distributed nature of the edge nodes at the far-edge, makes these very reliant on network conditions which tend to be more unstable and unreliable, particularly when compared to cloud environments. Self-healing capabilities and workload handovers are necessary to keep services and applications running at the edge in cases where network connections severely degrade or fail.

Last but not the least, typically edge nodes have limited storage capacity and to surpass this challenge, a Distributed Storage Layer (Smart Storage) will be designed with the purpose of supporting edge native applications.

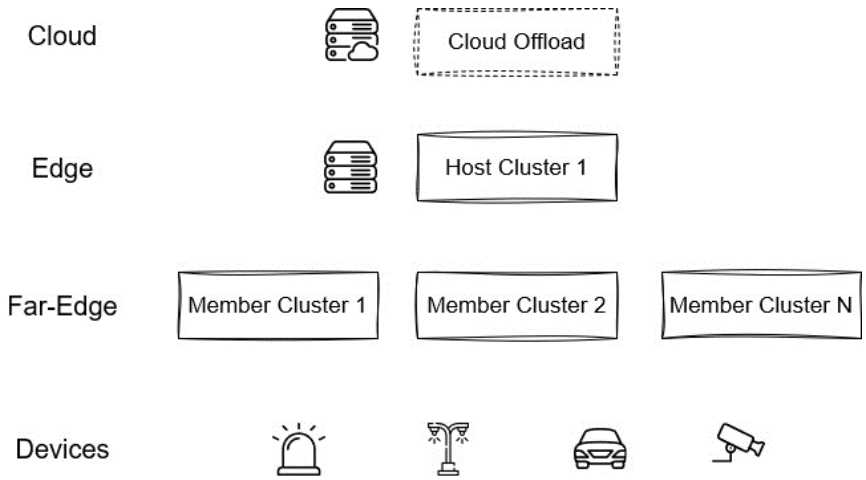

Figure 8-7 Far edge architecture.

In that way we have Custom OS for Edge Nodes, that will allow us work with federation of Edge Node (Member Clusters). Each edge node is a k8s cluster, host cluster propagates the target configuration for the member clusters and cluster-wide configurations are handled through a single API so will be a Cross-cluster discovery. Resource syncing across clusters supports to Edge Native Apps in which neighbour nodes can run workloads for other neighbouring edge nodes, quotas for each edge node, cloud offload capabilities and smart storage layer (distributed storage).

8.2.2 Multiple providers in the cloud continuum concept

The ETSI MANO framework, so far, allows only the Infrastructure as a Service (IaaS model**)** of the Private Cloud model, so it defines no business interfaces**.** It has, however, some mechanisms that can be exploited in a multi-provider scenario. For example, NFVI charging, NFVI performance evaluation, and roles of framework components in energy consumption have been outlined in ETSI NFV specifications, but they are not fully supported yet. The definition of the ETSI NFV approach to multi-domain lies in the use of domain-dedicated orchestrators. Such an approach provides benefits in terms of scalability and heterogeneity; however, the separation of infrastructure domains leads to inefficient use of resources.

This section describes ETSI NFV MANO framework modification by introducing a Resource Layer (RL) to support the Cloud Continuum concept, i.e., integration and exposure of cloud resources of multiple providers to orchestrators in a uniform way. Such an approach needs business interfaces with cloud providers. The business information should be secure and enable the exchange of information concerning data centre description (amount of resources, localisation, resource cost, etc.). The Cloud Continuum approach allows for the uniform use of resources. Still, it comes with a new set of problems related to business interfaces, the dynamicity of the cloud infrastructure and the efficient allocation of resources from a big resource pool. To solve the problems, the concept of the Resource Layer (RL) has been proposed and described in this section. The RL handles all Infrastructure-related operations and acts as a proxy between modified orchestrators that in the concept are service orchestrators only – the orchestration of resources is a part of RL. The modification has to include business-oriented interfaces and a mechanism that can be used for the selection of the partition of resources to be used by the orchestrator for the deployment of a specific Network Service. For this purpose, each data centre is described by the Data Centre Features (DCF) metric, which consists of data centres' geographic location, total capacity, delay of links between a specific data centre and other data centres, and the resource cost. Some of the data centre parameters, such as resource consumption, energy consumption, power status, and estimated reliability, are updated by data centres or RL in real time. In the case of mobile data centres (LEOs, UAVs, cars, user terminals), details concerning data centre mobility patterns are provided externally or calculated by the RL.

The main features of the proposed approach are the following:

- The RL has mechanisms for dynamic adding and removing data centres of multiple Infrastructure Providers (IPs) using secure interfaces. All RL databases are updated accordingly.
- The RL supports multiple orchestrators interacting with them by secure business interfaces. The RL exposes to orchestrators only a partition of RL resources with their topology - this is no longer the role of the orchestrator (ETSI MANO case). The partition is created using Network Service requirements that include data centre location, cost, energy efficiency, reliability, inter-data centre delay, etc. The RL provides to orchestrators not only information about resource consumption but also resource consumption predictions, reliability estimation of a data centre, etc., preferably using AI-driven algorithms. The service orchestrators can be data-driven therefore the resource scaling can be triggered by them, not only by the resource consumption level.

The RL internal components, presented in Figure 8-8, use a message bus, and new RL components can be added (orchestrated). The RL consists of the following components: Resource Database (updated in real-time) that keeps information about available and allocated resources, Resource Partitions that contains of map of resources allocated to partitions Resource Orchestrators that, in cooperation with Service Orchestrators (not described in this section) allocates/re-allocates resources , set of function responsible for autonomic operations of RL (these include self-management of RL), security functions responsible for authentication of data centres, service orchestrators and data centre site providers.

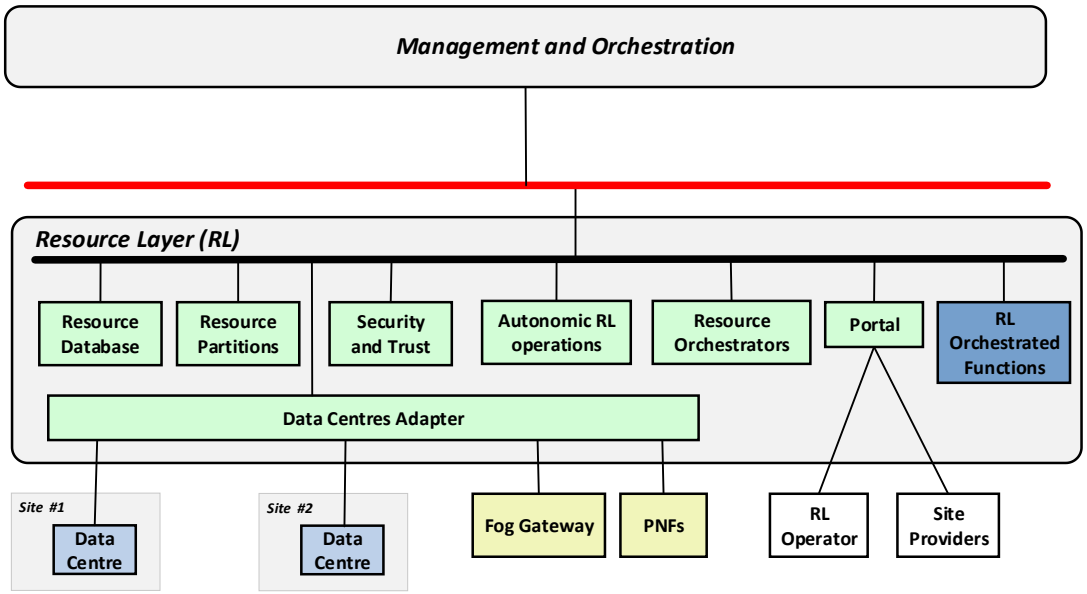

Figure 8-8 Integration of infrastructure resources using the Resource Layer concept.

Finally, RL consists of portal that is used by all business entities to exchange and negotiate resource related operations. The RL autonomic functions, mostly linked with resource allocations and predictions of KPIs and resource status, can be AI-driven. Please note that, in opposite to ETSI NFV MANO, in the approach the resource oriented operations are many but due to separation of concerns the complexity of RL is not visible externally.

8.2.3 Multi-cloud orchestration in federation scenarios

The new challenges in today networks include managing a heterogeneous environment composed of elements very different from one another (public cloud, private cloud, core/central cloud, edge clouds). Network complexity, SLAs and regulatory compliance are just a few of the challenges that arise when talking about multi-cloud management in a telco network.

In such heterogeneous environment, there is the need for a multi-cloud coordination and capability discovery (public cloud, private cloud, core/central cloud, edge clouds) while maintaining telco grade reliability, function discovery and load balancing.

As shown in Figure 8-9, this multi-cloud orchestration layer shall offer a unified and simplified view of the cloud resources managed by the distinct providers at the different hierarchical levels. It shall expose such cloud resources to network orchestration layers for deployment and management of network modules. This new layer should also take into consideration that the underlying cloud continuum can extend across operator boundary and across national geographical borders. Federation among MNOs shall rely on the multi-cloud orchestration layer to seamlessly give to the end user the same quality of service for a given application.

This study primarily focuses on analysing the requirements for the federation interfaces between different operators or, in general, cloud domains. Such interface shall be able to convey all the information required between different management and orchestration layers, to allow the creation of a single and coherent cloud continuum that spans throughout the separated domains.

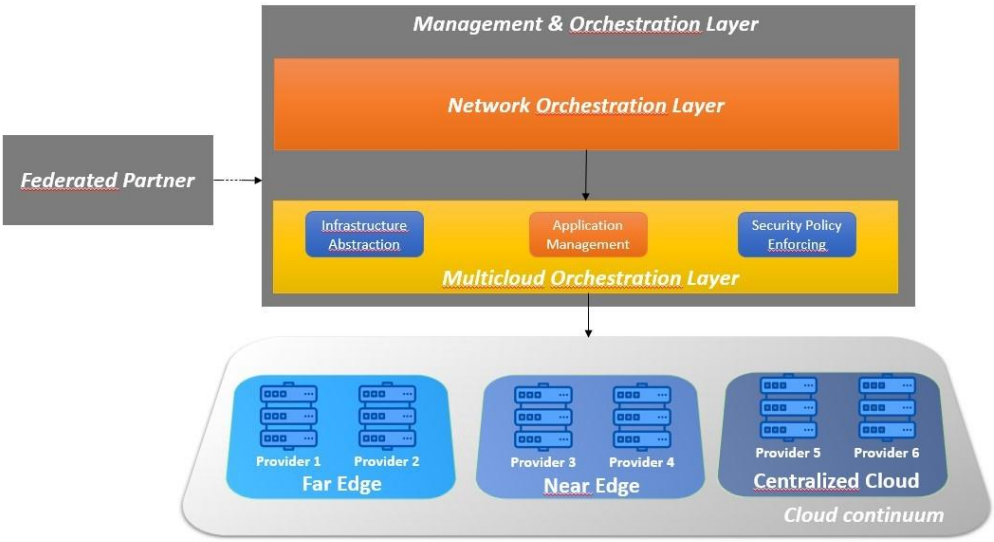

Figure 8-9 Integration of the multi-cloud orchestration layer with the cloud continuum.

Specifically, the emphasis will be on understanding how cloud resources would be shared between different operators leveraging the different levels of the cloud continuum, while ensuring a carrier-grade experience to the end-user. The study will propose solutions on how to manage the integration of various cloud service providers with different technologies considering the implications of data residency, data sovereignty, and cross-border data transfer in federation scenarios.

8.3 Network modules placement

Edge computing is deemed as a key enabling technology of the current and next generation of mobile networks. It allows overcoming cloud limitations associated with latency and enables applications that require a short range of delay. The next paradigm for delivering computing resources closer to the users is commonly known as extreme edge cloud, where compute resources beyond RAN are also considered. This comes with the promise of extremely reduced latency, which is a key requirement for different 6G use cases. Stitching together all the different compute resources would therefore form a cloud continuum, as illustrated in Figure 8-10.

Cloud environments are commonly known for managing vertical applications. However, beside vertical applications, a cloud environment is also expected to host network modules. The latter are responsible for managing data and control plane operations, making them also a key aspect in delivering the expected QoS to the end users. The upcoming generation of mobile networks has already initiated consideration to re-design network modules.

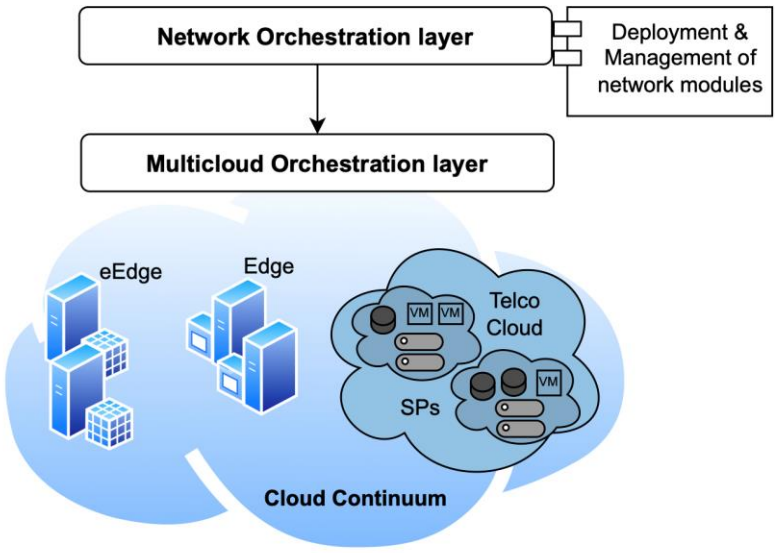

Figure 8-10 Network module placement in the resource continuum.

This enabler addresses the placement of network modules across the cloud continuum. More precisely, two dimensions are considered: while Section 8.3.1 targets the deployment of ETSI MEC components (cloud provider perspective), Section 8.3.2 focuses on the deployment of network operator components (telco perspective).

8.3.1 ETSI MEC placement in constrained devices

This enabler takes as input the same study areas as defined in Section 8.1.1, although with a different point of view. As indicated in Section 8.1.1 one of the elements, which may be streamlined and placed across the Compute Continuum (CC), is the ETSI MEC component. In fact, once the CC is defined and deployed, the different architectural enhancements for the deployment of ETSI MEC functions in constrained devices as defined in Section 8.1.1 can be applied, therefore re-structuring the network.

8.3.2 Network module placement across cloud continuum

Network modularization has already initiated considerations to re-design modules for 6G. Section 5 emphasizes with different studies, such as slice as a meta module, flexible UPF design and CN-RAN refactoring. The re-design of network module is mainly motivated by the new requirements of 6G (e.g., in terms of deployment and execution time) which are challenging to meet with the current modules. In order to operate, network modules will be deployed/hosted in a cloud environment that provides the computation resources.

Nowadays, cloud platforms can provide computation resources at different locations of the network. This ranges from a central cloud that is far from the user to an extreme edge cloud which is closer to the end devices (after the Access Network). Cloud platforms are also associated with different capabilities and can be used by different tenants.

As a hosting environment for network modules, cloud platforms need to provide the adequate capabilities to meet the expected requirements. Figure 8-11 illustrates a cloud continuum environment, where network modules are deployed at some target locations to meet the expected QoS. More precisely, in order to address the above, this study targets the following:

- Definition of APIs to expose the underlying capabilities of cloud platforms. This covers different compute levels, ranging from a central cloud to an extreme edge cloud which is located near to the end user.
- Development of techniques to perform network module placement in a cloud continuum environment (leveraging the exposed capabilities) that meet module requirements.

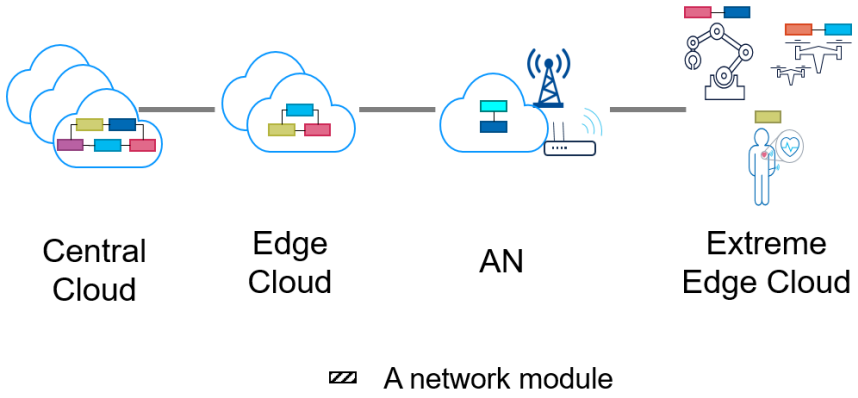

Figure 8-11 Placement of network modules in a cloud continuum.

As illustrated in Figure 8-11, the target use case is a cloud continuum environment, where network modules with expected requirements need to be deployed. The main target KPI are data latency and energy/performance trade-off.

8.4 Cloud transformation in 6G-quantum architecture

The softwarized continuum of cloud network of future will result in an explosion of network control traffic. Specific cases of unsupervised machine learning and optimization problems are computationally expensive. This can be exponentially reduced by encoding large amounts of data into the coefficients of a quantum state. The quantum states here are the counterpart of classical bits also known as quantum bits or qubits. The innate property of qubits of having a superposition of the possibility of being either '0' or '1' at the same time makes it resourceful in terms of computation. Some general class of optimization problems – principal component analysis (PCA), support vector machines (SVM), and any semidefinite program (SDP) can be efficiently run in polynomial time by a quantum computer if the data is encoded onto quantum states.

By introducing an intelligent cloud hosting control plane operations and integrating quantum technologies, we can incur a reduction of the traffic load. The proposed model of the quantum enabled cloud hosting control plane [RBD+21] are depicted in Figure 8-12. It can be seen in Figure 8-12; the collection of classical big data is being encoded into quantum states in a distributed manner. Each data source node collects the data based on few qubits that it possesses. The 6G network data-plane node will be controlled by up to 2^11 parameters which can be encoded in 11 qubits [IZA14]. This way only log(N) qubits are required for a N number of classical bits, that is to be sent to the quantum data centre hosting the hypervisor for processing.

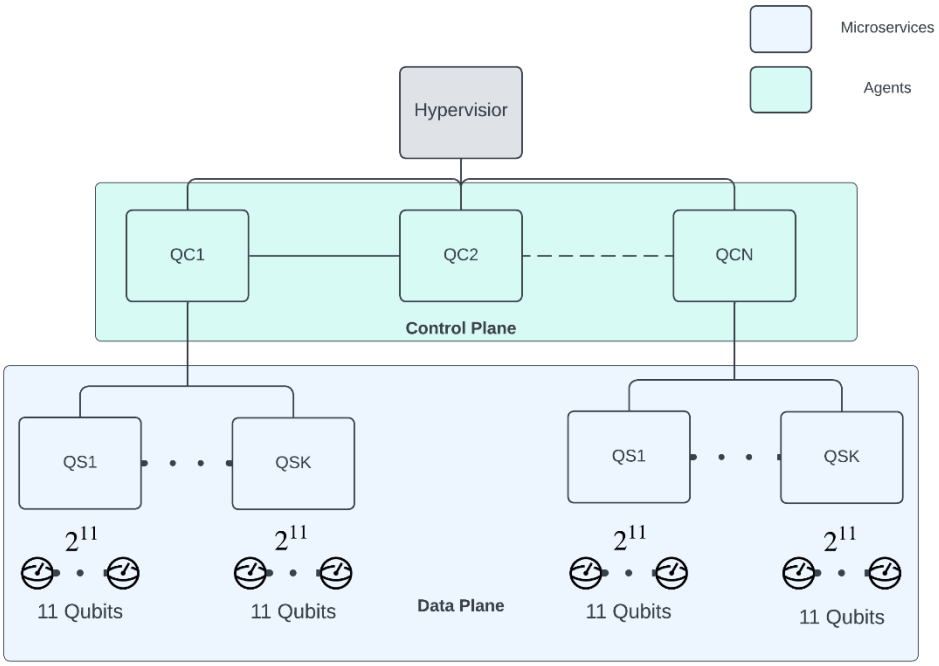

Figure 8-12 Example cloud hosting control plane functionalities and network intelligent hypervisor to configure network routing node.

The SDN controllers in control plane depicted in Figure 8-12 collects quantum states from K different switches or microservices. Then log K qubits are required by the SDN node to join the qubits from all the node into 11 + log K qubits. The 11 qubits of all the microservices are joined into the same 11 qubits via entanglement with the log K qubits the SDN node (or agent) possess. The agents will then further send the newly encoded quantum states with minimum overhead to the primary hypervisor, which in turn will join the states received by N quantum agents into a state of 11+log K + log N qubits. The dramatic effect has been shown in the plot in Figure 8-13.

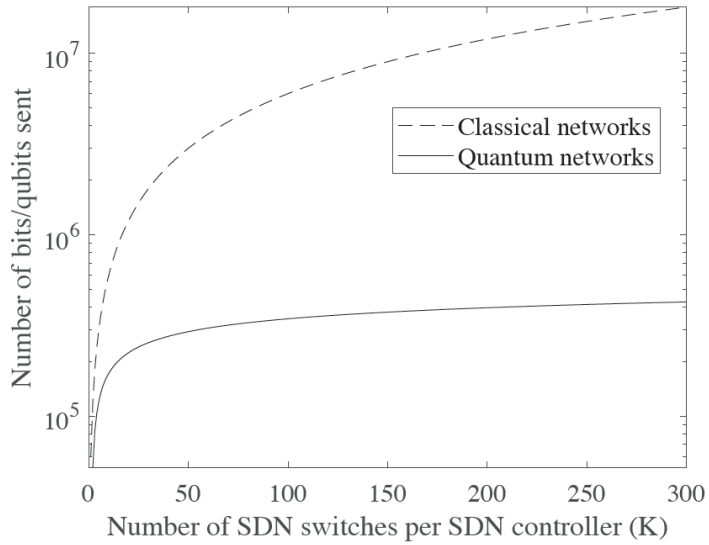

Figure 8-13 Scaling of the number of bits sent to the hypervisor, compared to the number of qubits sent [RBD+21].

This reduction of packets sent, stored, and processed within the control plane will reduce congestion and can target several sources of delays such as queueing, transmission, or processing. Furthermore, shared entanglement between nodes can even distribute load across nodes and reduce communication complexity by sending data that is more correlated than the classical ones.

9 Summary and Conclusions

This document gives an initial description of identified enablers for the 6G architecture. The document further gives an overview of previous work in Chapter 2 and describes the use cases that are applicable to the enablers in Chapter 3. The bulk of the document is spent on describing the enablers, including the reasoning and motivation why the enabler is important for the 6G architecture. Also, there is an introduction to the studies planned for each enabler area.

The main areas of the 6G architecture described in this document are the data-driven architecture, modular network, new access and flexible topologies, beyond communication and finally, the cloud transformation.

For the 6G **data-driven architecture**, several AI enablers are described. These enablers comprise architectural means and protocols, Machine Learning Operations (MLOps), Data Operations (DataOps), AI as a Service (AIaaS), and Intent-based management. The AI enablers form a robust framework for seamlessly integrating AI into the fabric of 6G networks.

To enable flexibility without increasing complexity, 6G needs an easily deployable architecture of modules (e.g., network functions) that can grow and adapt to the current needs. The document describes an initial concept on how the **network modularity** can decompose the 6GS into orthogonal building blocks (i.e., network functions, services and interfaces) with the right level of granularity. Modularisation of the network functions needs to be performed with an E2E vision, considering not only the network function granularity but also the necessary interfaces and deployment options to incorporate existing and new use cases such as NTN, programmability and XaaS.

New access and flexible topologies consist of the "network of networks" enabler. In this document, initial concepts on how to how to integrate subnetworks and Non-Terrestrial Networks (NTN) to the 6G architecture are given. To support new accesses, new 6G multi-connectivity innovations are proposed, both for the terrestrial network but also between the Terrestrial Network and NTN.

The **beyond conventional connectivity** is expanding the network's scope by processing data, generating insights, and delivering added value from societal, innovation, and business perspectives. This document describes several of the resulting new services such as sensing and compute-as-a-service.

The off-the-shelf cloud from e.g., Amazon and Microsoft are suitable for a big subset of multimedia human-scale applications, but it has its limitations when it comes down to supporting the upcoming latency sensitive 6G use cases. This document describes an initial concept on how to **transform the cloud** so it fits applicable 6G requirements such as management, latency, security and connection reliability, etc. Further on, future 6G networks should consider **cloud computing** capabilities across the entire network, from the extreme edge (including the UE) to Telco grade clouds (CC, Compute Continuum).

The enablers are summarized in Table 9-1, including a short background and the current understanding of the benefits of each enabler. Further on, the last column gives an initial understanding of the implications of the enabler, i.e., what is needed to implement the enabler in the 6G system.

The work with the enablers and their benefits and implications will continue and be further expanded in next deliverable D3.3, to be released in April 2024.

Table 9-1 Summary of the architecture enablers

Enabler	Background	Benefits	Implications
Architectural means and protocols	6G data-driven architecture will require architectural support that enables communication for AI	Can help define the inter-layer APIs and the protocols used to connect the layers of an E2E system design	Define internal and external APIs that realize the inter-layer interaction

MLOps	AI execution environments will be everywhere (e.g., UE, RAN and Core) and require tools for managing the lifecycle of these AI models	Improve operation, management, and maintenance of the E2E system design. Reduces the hardware requirements (compute/memory) in the distributed nodes. Enable cross-layer training in decentralized way (when datasets cannot be moved). Customizing models without additional labelled data in inference.	Privacy-aware data collection and AI management are needed. There is a communication and synchronization overhead between the compute nodes. Also, there is a trade-off between computation vs data collection.
AIaaS	AIaaS is a framework that offers a wide range of AI services as well as personalized inference capabilities to the AI service itself	Improve M&O and FCAPS of the network as well as impact on design of the E2E system	Impact on the E2E system design, AIaaS needs DataOps, MLOps and protocols
DataOps	Data shall be delivered, pre-processed, and stored where and when required. This imposes requirements on a flexible data ingestion architecture.	Efficiently collect and process data, as well as provide inferences to the data consumers within the E2E system design	Impact on the RAN and CN architecture; functions, protocols and interfaces may be needed.
Optimized network function composition	Modular design minimizes the dependencies between different modules while the relevance of the NF functionalities within a module is maximized	Increased flexibility, optimized signaling, and efficient resource usage	Impact on the CN NF design in 6G since the design need to be different from 5G
Streamlined network function interfaces & interaction	The network modules and their interfaces need to support the coexistence of these use cases as well as the related services.	Extend the support for new and existing use cases as they could be optimized based on the NF (or Network module) placement choices (e.g., centralized and distributed cloud deployments).	Impact on NF design and 5G procedures. Need for new interfaces and interaction
Flexible feature development and run-time scalability with modular network functionality	Exploring the possible enhancements to the E2E modularization (e.g., network slicing in 5G) to optimize the functionality	Enhanced network slicing & performance, flexibility via modularization, customization of E2E functionality	E2E impacts as the design and placement of network modules through the cloud continuum (e.g., cloud, edge, access, extreme edge etc.) would be revisited.
Network autonomy & Multi-X orchestration	In 5G, network slicing was a key enabler to facilitate the co-existence of various use cases with demanding and often conflicting requirements. The management and orchestration are built upon open loop slice configurations and semi-static parameters from	Improved data-based slice management. With a more autonomic and closed-loop based slice orchestration mechanisms it will be possible to address the orchestration of the network services including the extreme-edge domain, which is highly dynamic, heterogeneous, and volatile.	E2E impacts as the NF placement decisions through cloud continuum would be optimized with a higher time granularity based on the network dynamics. It requires closed-loop control and more flexible orchestration mechanisms as well as an enhanced exposure process. It will require also to define a comprehensive information model capturing the peculiarities of

	SLAs which often result in low resource utilization		those devices in extreme-edge domain.
Network migration	To perform this transition as efficiently as possible 6G should take this into consideration from the beginning: how to migrate from 5G to 6G	Critical, will determine how the E2E system will look like	It has fundamental impact to the 6G RAN & CN as it will outline the evolution path from 5G to 6G
Network of networks	Integration of multiple subnetworks, including terrestrial and non-terrestrial networks in order to create a seamless and ubiquitous communication system	Improved coverage, reduced complexity, increased reliability and more efficient management of network resources	New UE roles and responsibilities in a subnetwork, communication between non-terrestrial nodes, trust of diverse network nodes, communication and computation resource management
Multi-connectivity	Multi-connectivity enables multiple frequency ranges by different physically separated nodes, the aggregation of different radio access technologies, carriers, and access networks	Robustness and reliability, increased throughput and efficiency of the resource usage	Depending on the solution, new interfaces and protocols between nodes may be needed, which may lead to an increased complexity in coordinating different NW nodes. New mechanisms and procedures for integration of multiple RANs should be defined
E2E context awareness management	Mechanisms to allow each network component to dynamically adapt to the context to ensure the expected E2E QoS	Mission-critical operations to reduce the network overhead and to allocate edge resources flexibly, ultimately improving the system performance by allowing multiple edge allocations and RAN slices.	Different network components e.g., RAN/CN, transport, applications, should become aware of the context and need to interact, implying the need for signalling and synchronisation. Effective resource allocation and orchestration mechanisms that operate even when incomplete or partial context awareness is available should be designed
Exposure and data management	Functions to process the data collected and how to expose the (managed) data to external an internal usage	Expose data that may enable new 6G services	Impact mainly on the CN architecture and the Network-centric application layer; new functions, protocols and interfaces may be needed
Protocols, signalling and procedures	Discovery of compute nodes and impact of new sensing services on RAN interfaces and functionality	Critical to implement the Beyond Comm. Functionalities (BCFs)	New radio measurements needed; protocols needed to collect data to the data management.
Application- and Device-driven optimisation for Beyond Communication Services	Defining the requirements associated to applications using JCAS, Digital Twinning, etc	Improved QoS/QoE through efficient placement of BCF/BCS data/inference consumers (application, devices) within the E2E system design	Enhanced orchestration mechanisms across the continuum; efficient network – application component communication and service exposure.
Enhancing Joint Communication and Sensing Capabilities	Several 6G use cases require extreme localization performance, such as highly precise SLAM. Furthermore, massive distributed JCAS can raise	Accurate indoor mapping in challenging scenarios with the help of sensing information from COTS devices. Overcoming limitations for classical technologies on	Impacts the sensing design and data collection from COTS devices for SLAM; Hybrid classical-quantum network, where quantum virtual machines will

	significant limitations for classical technologies; quantum technologies is thus needed.	massive communication overhead and computational complexity	hold entanglements and qubits for its usage in the network
Integration and orchestration of computing continuum resources into the 6G architecture	Future 6G networks will consider computing capabilities across the full network, from the extreme edge (including the UE) to Telco grade clouds (CC, Compute Continuum	Better management of the resources and services in the CC	Impacts the extension of the CC, where strong emphasis is given on the extreme-edge integration, management and usability
Multi-domain/Multi-cloud federation	Different Telco-Cloud Providers are offering Compute, Storage and Network resources as a service on different platforms complicates management	Aggregation of resources, unification of existing domain-specific orchestration frameworks into an e2e federated architecture	It has fundamental impact in addressing the exiting challenges in multi domain federation as it will define important design, integration and orchestration principles for federation of services or/and resources.
Network modules placement in the resource continuum	The function (module) placement in the CC is challenging considering the heterogeneity and volatility of the Edge and Extrema Edge resources	Improve flexibility and critical services of the network by real-time function placement	Definition of APIs to expose capabilities of heterogeneous computing resources in the CC
Cloud Transformation in 6G-quantum architecture	The 6G edge supporting the softwarized network continuum will imply the explosion of network control traffic. Quantum technologies and a 6G-quantum network architecture can improve the optimal use of resources.	By using lossy encoding of data in quantum bits and adapting the cloud with algorithms to process them, it is possible to reduce the load of data mining procedures.	Will impact the reduction of traffic load in the CC by integrating quantum technologies

10 References

[22.856] 3GPP TR 22.856, “Study on localized mobile metaverse services”, Mar. 2023

[22.916] 3GPP TR 22.916 “Study on Network of Service Robots with Ambient Intelligence”, version 0.4.0 Release 19, June 2023.

[23.222] 3GPP TS 23.222, “Functional architecture and information flows to support Common API Framework for 3GPP Northbound APIs; Stage 2”, Version 18.2.0, June 2023

[23.288] 3GPP TS 23.288, “Architecture enhancements for 5G System (5GS) to support network data analytics services”, version 18.1.0, March 2023.

[23.501] 3GPP TS23.501 “System architecture for the 5G System (5GS)”, version 16.6.0 Release 16, October 2020.

[23.502] 3GPP TS23.502 “Procedures for the 5G System (5GS)”, version 17.5.0 Release 17, July 2022.

[26.928] 3GPP TR 26.928, “Extended Reality (XR) in 5G”, V18.0.0, Mar. 2023

[26.998] 3GPP TR 26.998, “Support of 5G glass-type Augmented Reality / Mixed Reality (AR/MR) devices”, Sep. 2022

[32.500] 3GPP TS 32.500 “Self-Organizing Networks (SON): Concepts and requirements (Release 8)”, version 1.0.1, December 2008.

[38.300] 3GPP TS 38.300 “NR and NG-RAN Overall Description; Stage 2”, V17.5.0, June 2023.

[38.401] 3GPP TS 38.401, “TSG RAN; NG-RAN; Architecture Description,” Release 16, v16.3.0, November 2020

[38.413] 3GPP TS 38.413, “NG-RAN; NG Application Protocol (NGAP) (Release 17)”, V17.6.0, Sep 2023

[38.420] 3GPP TS 38.420: "NG-RAN; Xn general aspects and principles", V17.2.0, September 2022.

[38.821] 3GPP TR 38.821 “Solutions for NR to support non-terrestrial networks (NTN)”, V16.1.0 (2021-05).

[37.340] 3GPP TS 37.340 “Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2”, V17.6.0, September 2023

[5GIA21] The 5G Infrastructure Association (5GIA). European Vision for the 6G Network Ecosystem. https://5g-ppp.eu/wp-content/uploads/2021/06/ WhitePaper-6G-Europe.pdf. Accessed: 2022-09-08.

[5GP22] 5G PPP Architecture Working Group, “The 6G Architecture Landscape: European Perspective”, Version 1.0, December 2022.

[6G-ANNA] 6G-ANNA, “Holistic approaches for 6th generation mobile networks,” August 2023. [Online]. Available: https://6g-anna.de/.

[6GFlagship] 6G Flagship, “6G-Enabled Wireless Smart Society & Ecosystem,” May 2022. [Online]. Available: https://www.6gflagship.com.

[6GPG] IMT-2030(6G) Promotion Group. https://www.imt2030.org.cn

[ABI+22] ABI Research, “Labor Shortages and Workplace Safety Concerns Propel Shipments of Outdoor Mobile Robots to 350,000 by 2030,” November 2022. [Online]. Available: https://www.abiresearch.com/press/labor-shortages-and-workplace-safety-concerns-propel-shipments-of-outdoor-mobile-robots-to-350000-by-2030/.

[AKP+21] M. Agiwal, H. Kwon, S. Park, and H. Jin, “A survey on 4G-5G dual connectivity: Road to 5G implementation,” IEEE Access, vol. 9, pp. 16193–16210, 2021.

[AMC19] Ö. U. Akgül, I. Malanchini and A. Capone, "Dynamic Resource Trading in Sliced Mobile Networks," in IEEE Transactions on Network and Service Management, vol. 16, no. 1, pp. 220-233, March 2019, doi: 10.1109/TNSM.2019.2893126.

[APIG] Apigee [Online]. Available at: https://cloud.google.com/apigee (Accessed: 04 July 2023).

[ASC+22] M. M. Azari S. Solanki, S. Chatzinotas, et. al, "Evolution of Non-Terrestrial Networks From 5G to 6G: A Survey," in IEEE Communications Surveys & Tutorials, vol. 24, no. 4, pp. 2633-2672, Fourth quarter 2022, doi: 10.1109/COMST.2022.3199901.

[AWSAG] AWS API Gateway [Online]. Available at: https://aws.amazon.com/api-gateway (Accessed: 04 July 2023).

[AWSCF] AWS CloudFormation, 2023. [Online] Available at: https://aws.amazon.com/cloudformation (Accessed: 04 July 2023).

[AzPol] Microsoft, “Azure Policy, Real-time cloud compliance at scale with consistent resource governance,” 2023. [Online] Available at: https://azure.microsoft.com/en-us/products/azure-policy (Accessed: 04 July 2023)

[AzRM] Microsoft, “Azure Resource Manager,” 2023 [Online] Available at: https://azure.microsoft.com/en-us/get-started/azure-portal/resource-manager/ (Accessed: 04 July 2023).

[B5G6G] Beyond 5G R&D Promotion Project. National Institute of Information and Communications Technology. https://b5g-rd.nict.go.jp/en/program

[BCC19] E. Borcoci, C. Contu, A. Ciobanu, “5G Slicing Management and Orchestration Architectures - Any Convergence?”, AFIN 2019 : The Eleventh International Conference on Advances in Future Internet, Available at: http://personales.upv.es/thinkmind/dl/conferences/afin/afin_2019/afin_2019_1_20_40013.pdf (Accessed: 27 September 2023).

[Ber94] H. R. Berenji, "Fuzzy Q-learning: a new approach for fuzzy dynamic programming," Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference, Orlando, FL, USA, 1994, pp. 486-491 vol.1, doi: 10.1109/FUZZY.1994.343737.

[Bha6G] Bharat 6G Vision https://bharat6galliance.com/img/Bharat-6G-Vision-Statement-copy%202_1.pdf

[BKA+21] M. Babar, M. S. Khan, F. Ali, M. Imran, and M. Shoaib, “Cloudlet Computing: Recent Advances, Taxonomy, and Challenges,” IEEE Access, vol. 9, pp. 29 609–29 622, 2021

[BLG+23] Ö. Bulakçı, X. Li, M. Gramaglia, A. Gavras, M. Uusitalo, P. Rugeland, M. Boldi, "Towards Sustainable and Trustworthy 6G: Challenges, Enablers, and Architectural Design", Boston-Delft: now publishers, 2023. http://dx.doi.org/10.1561/9781638282396

[BMG+22] A. Blanco, P. J. Mateo, F. Gringoli, and J. Widmer, “Augmenting MmWave localization accuracy through sub-6 GHz on off-the-shelf devices,” in Proc. of ACM MobiSys, 2022, p. 477–490.

[CAV22] P. Cruz, N. Achir, A.C. Viana. On the Edge of the Deployment: A Survey on Multi Access Edge Computing. ACM Computing Surveys, In press, 55 (5), pp.1-34. ff10.1145/3529758. hal-03637105.

[CB17] H. Corrigan-Gibbs and D. Boneh Prio, “Private, Robust, and Scalable Computation of Aggregate Statistics,” Stanford University, March 14, 2017.

[CDT18] T. Cerny, M.J. Donahoo, and M. Trnka, "Contextual understanding of microservice architecture: current and future directions," ACM SIGAPP Applied Computing Review, vol. 17, no. 4, pp. 29-45, 2018.

[CONS21] ETSI GR MEC 036 “Multi-access Edge Computing (MEC); MEC in resource constrained terminals, fixed or mobile” DGR/MEC-0036ConstrainedDevice, ETSI Std. v3.0.4 Draft, 2021.

[FOG+18] IEEE Standard for Adoption of OpenFog Reference Architecture for Fog Computing, IEEE 1934-2018, Jul. 2018.

[CSL+23] C. Chen, H. Song, Q. Li, F. Meneghello, F. Restuccia and C. Cordeiro, "Wi-Fi Sensing Based on IEEE 802.11bf," in IEEE Communications Magazine, vol. 61, no. 1, pp. 121-127, January 2023, doi: 10.1109/MCOM.007.2200347.

[CTJ+18] Y. Chen, J. Tang, C. Jiang, L. Zhu et al., “The accuracy comparison of three simultaneous localization and mapping (slam)-based indoor mapping technologies,” Sensors, vol. 18, no. 10, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/10/3228

[CTM+22] M. Corici, E. Troudt, T. Magedanz and H. Schotten, "Organic 6G Networks: Decomplexification of Software-based Core Networks," 2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Grenoble, France, 2022, pp. 541-546.

[DCT+22a] M. Diamanti, P. Charatsaris, E. E. Tsiropoulou and S. Papavassiliou, "The Prospect of Reconfigurable Intelligent Surfaces in Integrated Access and Backhaul Networks," in IEEE Transactions on Green Communications and Networking, vol. 6, no. 2, pp. 859-872, June 2022, doi: 10.1109/TGCN.2021.3126784.

[DCT+22b] M. Diamanti, P. Charatsaris, E. E. Tsiropoulou and S. Papavassiliou, "Incentive Mechanism and Resource Allocation for Edge-Fog Networks Driven by Multi-Dimensional Contract and

Game Theories," in IEEE Open Journal of the Communications Society, vol. 3, pp. 435-452, 2022, doi: 10.1109/OJCOMS.2022.3154536.

[DFP17] C. L. Degen, R. Friedemann, and P. Cappellaro. "Quantum sensing." Reviews of modern physics vol. 89, no. 3, pp. 035002, 2017.

[DNC+01] M. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte et al., "A solution to the simultaneous localization and map building (slam) problem," IEEE Transactions on Robotics and Automation, vol. 17, no. 3, pp. 229–241, 2001.

[DXN+15] J. Dong, Y. Xiao, M. Noreikis, Z. Ou et al., "IMoon: Using smartphones for image-based indoor navigation," in Proc. of ACM SenSys, pp.85–97, 2015.

[Eri22] Ericsson Research Blogpost. "Decentralized learning and intelligent automation: the key to zero-touch networks?", March 2022.

[FAC+19] C. Fiandrino, H. Assasa, P. Casari, and J. Widmer, "Scaling millimeterwave networks to dense deployments and dynamic environments," Proceedings of the IEEE, vol. 107, no. 4, pp. 732–745, 2019.

[FAS+21] F. Saeik, M. Avgeris, D. Spatharakis, N. Santi, D. Dechouniotis, J. Violos, A. Leivadeas, N. Athanasopoulos, N. Mitton, and S. Papavassiliou, "Task offloading in Edge and Cloud Computing: A survey on mathematical, artificial intelligence and control theory solutions," in Computer Networks 195, 2021, doi: 10.1016/J.COMNET.2021.108177.

[FBD+21] R. Ferrara, R. Bassoli, C. Deppe, F. H. Fitzek, and H. Boche, "The computational and latency advantage of quantum communication networks," IEEE Communications Magazine, vol. 59, no. 6, pp. 132-137, 2021.

[FIP+20] G. I. Fragapane, D. A. Ivanov, M. Peron, F. Sgarbossa et al., "Increasing flexibility and productivity in industry 4.0 production networks with autonomous mobile robots and smart intralogistics," Annals of Operations Research, vol. 308, pp. 125–143, 2020.

[FV07] N. Fulda and D. Ventura, "Predicting and preventing coordination problems in cooperative Q-learning systems", IJCAI'07: Proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007.

[GC21] A. Garcia-Saavedra and X. Costa-Pérez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," IEEE Communications Standards Magazine, vol. 5, Art. no. 4, 2021.

[GD22] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," Artif. Intell. Rev. vol. 55, pp. 895–943, 2022. [Online]. Available: https://doi.org/10.1007/s10462-021-09996-w

[GDT+22] K. Gasmi, S. Dilek, S. Tosun and S. Ozdemir, "A Survey on Computation Offloading and Service Placement in Fog Computing-Based IoT," The Journal of Supercomputing, vol. 78, no. 2, p. 1983–2014, 2022.

[GKM+22] M. Gramaglia, M. Kajo, C. Mannweiler, O. Bulakci and Q. Wei, "A unified service-based capability exposure framework for closed-loop network automation", Transactions on Emerging Telecommunications Technologies, vol. 33, no 11, p.e4598, July 2022.

[GPR+23] T. Geoghegan, C. Patton, E. Rescorla, and C. A. Wood "Distributed Aggregation Protocol for Privacy Preserving Measurement," IETF. July 10, 2023.

[GSH+22] E. Goshi, R. Stahl, H. Harkous, M. He, R. Pries and W. Kellerer, "PP5GS -An Efficient Procedure-Based and Stateless Architecture for Next Generation Core Networks," in IEEE Transactions on Network and Service Management, 2022. doi: 10.1109/TNSM.2022.3230206.

[GSMA] GSMA Open Gateway [Online]. Available at: https://www.gsma.com/futurenetworks/gsma-open-gateway (Accessed: 04 July 2023).

[GUA+16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," The journal of machine learning research, vol. 17, no. 59, pp. 1-35, 2016. [Online]. Available: https://jmlr.org/papers/volume17/15-239/15-239.pdf

[HEX2-D11] Hexa-X-II Deliverable D1.1, "Environmental, societal and economical drivers and goals for 6G", Jun. 2023, Online: https://hexa-x-ii.eu/wp-content/uploads/2023/07/Hexa-X-II_D1.1_final-website.pdf

[HEX2-D21] Hexa-X-II Deliverable D2.1, “Draft foundation for 6G system design”, June., 2023, Online: https://hexa-x-ii.eu/wp-content/uploads/2023/07/Hexa-X-II_D2.1_web.pdf[HEXA] Hexa-X website, https://hexa-x.eu

[HEX-D12] Hexa-X Deliverable D1.2, “Expanded 6G vision, use cases and societal values”, Apr., 2021, Online: https://hexa-x.eu/wp-content/uploads/2022/04/Hexa-X_D1.2_Edited.pdf

[HEX-D13] Hexa-X Deliverable D1.3, “Targets and requirements for 6G – initial E2E architecture”, Mar., 2022, Online: https://hexa-x.eu/wp-content/uploads/2022/03/Hexa-X_D1.3.pdf

[HEX-D14] Hexa-X Deliverable D1.4, “Hexa-X architecture for B5G/6G networks – final release”, June, 2023, Online: https://hexa-x.eu/wp-content/uploads/2023/07/Hexa-X-D1.4-Final.pdf

[HEX-D31] Hexa-X Deliverable D3.1, “Localisation and sensing use cases and gap analysis”, Dec. 2021, Online: https://hexa-x.eu/wp-content/uploads/2022/02/Hexa-X_D3.1_v1.4.pdf

[HEX-D51] Hexa-X Deliverable D5.1, “Initial 6G Architectural Components and Enablers”, Dec. 2021, Online: https://hexa-x.eu/wp-content/uploads/2022/03/Hexa-X_D5.1_full_version_v1.1.pdf

[HEX-D52] Hexa-X Deliverable D5.2, “Analysis of 6G architectural enablers’ applicability and initial technological solutions”, Oct. 2022, Online:Hexa-X_D5.2_v1.0.pdf

[HEX-D53] Hexa-X Deliverable D5.3, “Final 6G architectural enablers and technological solutions”, April 30, 2023

[HEX-D61] Hexa-X Deliverable 6.1, “Gaps, features and enablers for B5G/6G service management and orchestration”, June 30, 2021

[HEX-D62] Hexa-X Deliverable D6.2, “Design of service management and orchestration functionalities”, April 29, 2022, Online: https://hexa-x.eu/wp-content/uploads/2022/05/Hexa-X_D6.2_V1.1.pdf

[HEX-D63] Hexa-X Deliverable D6.3, “Final evaluation of service management and orchestration mechanisms”, April 30, 2023, Online: https://hexa-x.eu/wp-content/uploads/2023/05/Hexa-X_D6.3_v.1.1.pdf

[HEX-D71] Hexa-X Deliverable D7.1, “Gap analysis and technical work plan for special-purpose functionality”, June., 2021, Online: https://hexa-x.eu/wp-content/uploads/2021/06/Hexa-X_D7.1.pdf

[HKH+14] P. Henry, M. Krainin, E. Herbst, X. Ren et al., RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. Springer Berlin Heidelberg, 2014, pp. 477–491.

[HST+22] Y. Hatano, J. Shin, J. Tanigawa, Y. Shigenobu, A. Nakazono, T. Sekiguchi, S. Onoda, T. Ohshima, K. Arai, T. Iwasaki, and M. Hatano, “High-precision robust monitoring of charge/discharge current over a wide dynamic range for electric vehicle batteries using diamond quantum sensors,” Sci Rep vol. 12, pp. 13991, 2022.

[Hua22] Huawei Technologies, “Integrated sensing and communication: Concept and practice,” 2022, [Online]. Access: https://www.huawei.com/en/huaweitech/future-technologies/integrated-sensing-communication-concept-practice

[HW10] P. Hofmann and D. Woods, "Cloud Computing: The Limits of Public Clouds for Business Applications," in IEEE Internet Computing, vol. 14, no. 6, pp. 90-93, Nov.-Dec. 2010, doi: 10.1109/MIC.2010.136.

[IBN+19] G. Interdonato, E. Björnson, H. Quoc Ngo, P. Frenger, and E. G. Larsson, “Ubiquitous cell-free Massive MIMO communications,” EURASIP Journal on Wireless Communications and Networking, vol. 2019, Art. no. 1, 2019.

[IDG+21] A. Islam, A. Debnath, M. Ghose and S. Chakraborty, "A Survey on Task Offloading in Multi-Access Edge Computing," Journal of Systems Architecture, vol. 118, p. 102225, 2021.

[IETF23] IoT Edge Challenges and Functions, IETF Std. draft-irtf-t2trg-iot-edge-08, 2023.

[IFB+18] M. Iorga, L. Feldman, R. Barton, M. J. Martin, N. S. Goren, C. Mahmoudi et al., “Fog Computing Conceptual Model,” 2018.

[IFV21] S. Ickin, M. Fiedler, and K. Vandikas, “QoE Modeling on Split Features with Distributed Deep Learning,” Network, vol. 1, pp. 165-190. 2021. [Online]. Available: https://doi.org/10.3390/network1020011.

[ILR+22] S. Ickin, H. Larsson, H. Riaz, X. Lan, and C. Kilinc, “Decentralized learning and intelligent automation: the key to zero-touch networks?” – Ericsson Blogpost. March, 2022. [Online]. Available: https://www.ericsson.com/en/blog/2022/3/decentralized-learning-for-zero-touch-networks.

[ITU3172] ITU-T Rec. Y.3172, “Architectural Framework for Machine Learning in Future Networks Including IMT-2020,” 2019

[IZA14] A. Imran, A. Zoha, and A. Abu-Dayya, “Challenges in 5G: how to empower SON with big data for enabling 5G,” IEEE Network, vol. 28, no. 6, pp. 27–33, 2014.

[JHH+21] W. Jiang, B. Han, M. A. Habibi and H. D. Schotten, "The Road Towards 6G: A Comprehensive Survey," in IEEE Open Journal of the Communications Society, vol. 2, pp. 334-366, 2021, doi: 10.1109/OJCOMS.2021.3057679.

[JPQ+22] V. Jain, S. Panda, S. Qi and K. K. Ramakrishnan, "Evolving to 6G: Improving the Cellular Core to lower control and data plane latency," 2022 1st International Conference on 6G Networking (6GNet), Paris, France, 2022, pp. 1-8, doi: 10.1109/6GNet54646.2022.9830519.

[K3s23] (2023) K3s. [Online] Available at: https://k3s.io/ (Accessed: 04 July 2023).

[KFF+18] S. Kekki, W. Featherstone, Y. Fang, P. Kuure, A. Li, A. Ranjan, D. Purkayastha, F. Jiangping, D. Frydman, G. Verin, K. W. Wen, K. Kim, R. Arora, A. Odgers, L. M. Contreras, and S. Scarpina, “MEC in 5G networks,” ETSI Whitepaper #28, 2018, [Online] https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf

[KNE16] K. Katsalis, N. Nikaein and A. Edmonds, "Multi-Domain Orchestration for NFV: Challenges and Research Directions," 2016 15th International Conference on Ubiquitous Computing and Communications and 2016 International Symposium on Cyberspace and Security (IUCC-CSS), Granada, Spain, 2016, pp. 189-195, doi: 10.1109/IUCC-CSS.2016.034.

[KONG] Kong [Online]. Available at: https://konghq.com (Accessed: 04 July 2023).

[K8S] Kubernetes [Online]. Available at: https://kubernetes.io (Accessed: 28 September 2023).

[LBZ+21] L. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, and P. Hui, (2021). All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda. ArXiv, abs/2110.05352.

[LCG+21] M. C. Lucas-Estañ, B. Coll-Perales, and J. Gozalvez, “Redundancy and diversity in wireless networks to support mobile industrial applications in industry 4.0,” IEEE Transactions on Industrial Informatics, vol. 17, no. 1, pp. 311–320, 2021.

[LCW+20] S. Luo, X. Chen, Q. Wu, Z. Zhou and S. Yu, "HFEL: Joint Edge Association and Resource Allocation for Cost-Efficient Hierarchical Federated Edge Learning," in IEEE Transactions on Wireless Communications, vol. 19, no. 10, pp. 6535-6548, Oct. 2020, doi: 10.1109/TWC.2020.3003744.

[LFT+20] L. Li, Y. Fan, Y., M. Tse et. al.,” A review of applications in federated learning,” Computers & Industrial Engineering, November 2020, Vol. 149, https://doi.org/10.1016/j.cie.2020.106854

[LHA13] A. Lozano, R. W. Heath, and J. G. Andrews, “Fundamental Limits of Cooperation,” IEEE Transactions on Information Theory, vol. 59, Art. no. 9, 2013.

[LNXF] Linux Foundation [Online]. Available at: https://www.linuxfoundation.org (Accessed: 04 July 2023).

[LQW+20] Y. L. Lee, D. Qin, L. -C. Wang and G. H. Sim, "6G Massive Radio Access Networks: Key Applications, Requirements and Challenges," in IEEE Open Journal of Vehicular Technology, vol. 2, pp. 54-66, 2021, doi: 10.1109/OJVT.2020.3044569.

[LRZ+20] C. X. Lu, S. Rosa, P. Zhao, B. Wang et al., “See through smoke: Robust indoor mapping with low-cost MmWave radar,” in Proc. of ACM MobiSys, p. 14–27, 2020.

[LT11] S. Lasaulce and H. Tembine, "Game Theory and Learning for Wireless Networks," Academic Press, 2011.

[LWW+23] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 3347-3366, 1 April 2023, doi: 10.1109/TKDE.2021.3124599.

[LYC+22] S. Liu, G. Yu, X. Chen, and M. Bennis, "Joint User Association and Resource Allocation for Wireless Hierarchical Federated Learning With IID and Non-IID Data," in IEEE Transactions on Wireless Communications, vol. 21, no. 10, pp. 7852-7866, Oct. 2022, doi: 10.1109/TWC.2022.3162595.

[LZS+20] L. Liu, J. Zhang, S. H. Song, and K. B. Letaief, "Client-Edge-Cloud Hierarchical Federated Learning," ICC 2020 - 2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/ICC40277.2020.9148862.

[M.2516-0] Future technology trends of terrestrial International Mobile Telecommunications systems towards 2030 and beyond. Report ITU-R M.2516-0, November 2022. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2516-2022-PDF-E.pdf

[MAT+23] Militano, L.; Arteaga, A.; Toffetti, G.; Mitton, N. "The Cloud-to-Edge-to-IoT Continuum as an Enabler for Search and Rescue Operations." Future Internet, vol. 15, no. 2, 2023. https://doi.org 10.3390/fi15020055 Available at: https://www.mdpi.com/1999-5903/15/2/55 (Accessed: 27 sept. 2023).

[MAN004] ETSI GS NFV 006, "Network Functions Virtualisation (NFV) Release 4; Management and Orchestration; Architectural Framework Specification", version 4.4.1 (2022-12)

[MBF+22] A. Mahmood, L. Beltramelli, S. Fakhrul Abedin, S. Zeb et al., "Industrial IoT in 5g-and-beyond networks: Vision, architecture, and design trends," IEEE Transactions on Industrial Informatics, vol. 18, no. 6, pp. 4122– 4137, 2022.

[MEC003] ETSI GS MEC 003, "Multi-access Edge Computing (MEC); Framework and Reference Architecture", version 3.1.1, (2022-03)

[MET17-D24] METIS-II Deliverable D2.4, "Final Overall 5G RAN Design", June 2017, online: https://metis-ii.5g-ppp.eu/wp-content/uploads/deliverables/METIS-II_D2.4_V1.0.pdf

[MinKB] Minikube, 2023. [Online] Available at: https://minikube.sigs.k8s.io/docs/ (Accessed: 04 July 2023).

[MKs23] MicroK8s: Low-ops, minimal production Kubernetes, for devs, cloud, clusters, workstations, Edge and IoT (2023). [Online] Available at: https://microk8s.io

[MS22] M. Maray and J. Shuja, "Computation Offloading in Mobile Cloud Computing and Mobile Edge Computing: Survey, Taxonomy, and Open Issues," Mobile Information Systems, vol. 202, 2022.

[MSIT22] Ministry of Science and ICT, "6G, Korea takes the lead once again 6G R&D implementation plan established," June 2022 , https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=18&bbsSeqNo=42&nttSeqNo=517&searchOpt=ALL&searchTxt=

[NAY+17] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," IEEE Transactions on Wireless Communications, vol. 16, Art. no. 3, 2017.

[NextG] NextG: 5G-and-Beyond Technology. https://www.nist.gov/advanced-communications/nextg-5g-and-beyond-technology

[NGA] Next G Alliance (NGA). 6G Library. https://www.nextgalliance.org/ 6g-library/. Accessed: 2022-09-08.

[NLB21] D. T. Nguyen, L. B. Le and V. Bhargava, "Price-Based Resource Allocation for Edge Computing: A Market Equilibrium Approach," in IEEE Transactions on Cloud Computing, vol. 9, no. 1, pp. 302-317, 1 Jan.-March 2021, doi: 10.1109/TCC.2018.2844379.

[NOMAD] Nomad [Online]. Available at: https://www.nomadproject.io (Accessed: 28 September 2023).

[NPS+23] S. S. Nande, M. Paul, S. Senk, M. Ulbricht, R. Bassoli, F. H.P. Fitzek, and H. Boche. 2023. Quantum enhanced time synchronisation for communication network. Comput. Netw. 229, Computer Networks (Jun 2023). https://doi.org/10.1016/j.comnet.2023.109772

[OAPI] Open API [Online]. Available at: https://swagger.io/specification (Accessed: 04 July 2023).

[OD22] J. Ordonez-Lucena and F. Dsouza, "Pathways towards network-as-a-service: the CAMARA project," In Proceedings of the ACM SIGCOMM Workshop on Network-Application Integration, pp. 53-59, August 2022.

[ORAN] O-RAN Alliance, "O-RAN: Towards an open and smart RAN," White paper, vol. 19, 2018.

[PBD+23] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," IEEE Communications Surveys & Tutorials, vol. 25, Art. no. 2, 2023.

[PLR+22] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "SPARCS: A sparse recovery approach for integrated communication and human sensing in mmwave systems," in Proc. of ACM/IEEE IPSN, 2022, pp. 79–91.

[Pos07] R. A. Posner, “Economic Analysis of Law (Seventh ed.),” Austin, TX: Wolters Kluwer, ISBN 978-0-7355-6354-4, 2007.

[PTL+12] S. M. Perlaza, H. Tembine, S. Lasaulce, and M. Debbah, "Quality-Of-Service Provisioning in Decentralized Networks: A Satisfaction Equilibrium Approach," in IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 2, pp. 104-116, April 2012, doi: 10.1109/JSTSP.2011.2180507.

[PVC+19] M. G. Poirot, P. Vepakomma, K. Chang, J. Kalpathy-Cramer, R. Gupta, and R. Raskar, “Split Learning for collaborative deep learning in healthcare,“ 2019. [Online]. Available: https://arxiv.org/abs/1912.12115.

[RBD+21] F., Roberto, R. Bassoli, C. Deppe, F. H. P. Fitzek, and H. Boche. "The computational and latency advantage of quantum communication networks," IEEE Communications Magazine vol. 59, no.6, pp. 132-137, 2021.

[Res22] B. Reselman, “5 design principles for microservices,” 2022. [Online] Available at: https://developers.redhat.com/articles/2022/01/11/5-design-principles-microservices (Accessed: 22 August 2023).

[RESTR] Reactive Streams [Online]. Available at: https://www.reactive-streams.org/ (Accessed: 04 July 2023).

[RINGS] Resilient & Intelligent NextG Systems (RINGS). https://new.nsf.gov/funding/opportunities/resilient-intelligent-nextg-systems-rings

[RP-223534] 3GPP, “NR NTN (Non-Terrestrial Networks) enhancements, Rel-18 Work Item Description,” 2022. [Online]. Available: https://www.3gpp.org/ftp/TSG_RAN/TSG_RAN/TSGR_98e/Docs/RP-223534.zip

[SAE+11] L. C. Schmelz, M. Amirijoo, A. Eisenblaetter, R. Litjens, M. Neuland and J. Turk, "A coordination framework for self-organisation in LTE networks," 12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops, Dublin, Ireland, 2011, pp. 193-200, doi: 10.1109/INM.2011.5990691.

[SCT+22] A. M. Sanchez, A. S. Charismiadis, D. Tsolkas, D.A. Guillen, and J.G. Rodrigo, "Offering the 3GPP Common API Framework as Microservice to Vertical Industries", 2022 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), pp. 363-368, IEEE, June 2022.

[Sha48] C.E. Shannon, “A Mathematical Theory of Communication”, Bell System tech. Journal”, Vol. 27, pp. 379-423 and pp. 623-656, 1948.

[SIM+22] N. F. S. de Sousa, M. T. Islam, R. U. Mustafa, et al., ”Machine Learning-Assisted Closed-Control Loops for Beyond 5G Multi-Domain Zero-Touch Networks,” J Netw Syst Manage, vol. 30, no. 46, 2022. https://doi.org/10.1007/s10922-022-09651-x

[SLR+19] N.F. Saraiva de Sousa, D. A. Lachos Perez, R. V. Rosa, et al, "Network Service Orchestration: A Survey", arXiv:1803.06596v4 [cs.NI] 17 May 2019, Available at: https://arxiv.org/pdf/1803.06596.pdf

[SNH03] H. Surmann, A. Nüchter, and J. Hertzberg, “An autonomous mobile robot with a 3D laser range finder for 3D exploration and digitalization of indoor environments,” Robotics and Autonomous Systems, vol. 45, no. 3, pp. 181–198, 2003.

[SRH17] D. Shadija, M. Rezai, and R. Hill. "Microservices: granularity vs. performance." Companion Proceedings of the10th International Conference on Utility and Cloud Computing. 2017.

[SRN+21] D. Sabella, A. Reznik, K. R. Nayak, D. Lopez, F. Li, U. Kleber, A. Leadbeater, K. Maloor, S. B. Mary Baskaran, L. Cominardi, C. Costa, F. Granelli, V. Gazis, F. Ennesser, and X. Gu, “MEC Security: Status of Standard Supports and Future Evolutions,” ETSI white paper, vol. 46, pp. 1–26, 2021.

[STR21] StraitsResearch, “Slam technology market size, share, growth, trends, analysis, industry report 2030,” https://straitsresearch.com/report/slam-technology-market, 2021, [Accessed on 31 March 2023].

[SWARM] Docker Swarm [Online]. Available at: https://docs.docker.com/engine/swarm (Accessed: 28 September 2023).

[SYN+21] O. Serhane, K. Yahyaoui, B. Nour, and H. Moungla, “A Survey of ICN Content Naming and In-Network Caching in 5G and Beyond Networks,” IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4081–4104, 2021.

[TAS+19] Taleb, T., Afolabi, I., Samdanis, K., & Yousaf, F. Z. (2019). On multi-domain network slicing orchestration architecture and federated resource control. IEEE Network, 33(5), 242-252.

[TCZ+22] F. Tang, X. Chen, M. Zhao and N. Kato, "The Roadmap of Communication and Networking in 6G for the Metaverse," in IEEE Wireless Communications, 2022, doi: 10.1109/MWC.019.2100721.

[TrrFM] Terraform "Infrastructure automation to provision and manage resources in any cloud or data center," 2023. [Online] Available at: https://www.terraform.io (Accessed: 04 July 2023).

[VK21] H. Vural and M. Koyuncu, "Does Domain-Driven Design Lead to Finding the Optimal Modularity of a Microservice?," in IEEE Access, vol. 9, pp. 32721-32733, 2021.

[WDY+23] J. Wang, L. Dai, L. Yang, and B. Bai, "Clustered Cell-Free Networking: A Graph Partitioning Approach," IEEE Transactions on Wireless Communications, p. 1, 2023.

[WHM+22] N. Waqar, S. A. Hassan, A. Mahmood, K. Dev, D. -T. Do and M. Gidlund, "Computation Offloading and Resource Allocation in MEC-Enabled Integrated Aerial-Terrestrial Vehicular Networks: A Reinforcement Learning Approach," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 21478-21491, Nov. 2022, doi: 10.1109/TITS.2022.3179987.

[WSH+21] R. Wang, M. Shen, Y. He, and X. Liu, "Performance of Cell-Free Massive MIMO With Joint User Clustering and Access Point Selection," IEEE Access, vol. 9, pp. 40860–40870, 2021.

[WSL+21] H. Wymeersch, D. Shrestha, C. M. de Lima, et. al, "Integration of Communication and Sensing in 6G: a Joint Industrial and Academic Perspective," 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Helsinki, Finland, 2021, pp. 1-7, doi: 10.1109/PIMRC50174.2021.9569364.

[YTA+20] A. Yazar, S. D. Tusha, H. Arslan, "6G Vision: An ultra-flexible perspective", ITU Journal on Future and Evolving Technologies, Vol. 1, No. 1, December 2020.

[ZCZ+18] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data Security and Privacy-preserving in Edge Computing paradigm: Survey and Open Issues," IEEE access, vol. 6, pp. 18 209–18 237, 2018.

[ZSM002] ETSI GS ZSM 002, "Zero-touch network and Service Management; Reference Architecture", August 2019.

[ZXM+19] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 28–41, 2019.

[ZZH+22] Y. Zhan, J. Zhang, Z. Hong, L. Wu, P. Li and S. Guo, "A Survey of Incentive Mechanism Design for Federated Learning," in IEEE Transactions on Emerging Topics in Computing, vol. 10, no. 2, pp. 1035-1044, 1 April-June 2022, doi: 10.1109/TETC.2021.3063517.