Call: H2020-ICT-2020-2

Project reference: 101015956

Project Name:

A flagship for B5G/6G vision and intelligent fabric of technology enablers connecting human, physical, and digital worlds

Hexa-X

Deliverable D5.1
Initial 6G Architectural Components and Enablers

Date of delivery: 31/12/2021 Version: 1.0

Start date of project: 01/01/2021 Duration: 30 months

Document properties:

Document Number:	D5.1
Document Title:	Initial B5G/6G Architectural Components and Enablers
Editor(s):	Mårten Ericson (ERI), Hannu Flinck (NOK), Panagiotis Vlacheas (WINGS), Stefan Wänstedt (ERI)
Authors:	Riccardo Bassoli (TUD), Mårten Ericson (ERI), Frank H.P. Fitzek (TUD), Hannu Flinck (NOK), Mu He (NOK), Bahare Masood Khorsandi (NOK), Gerald Kunzmann (NOK), Petteri Pöyhönen (NOK), Janne Tuononen (NOK), Panagiotis Vlacheas (WINGS), Stefan Wänstedt (ERI), Gunnar Mildh (ERI), Merve Saimler (ERI), Mehdi S. H. Abad (ERI), Damiano Rapone (TIM), Miltiadis Filippou (INT), Markus Dominik Mueck (INT), Thomas Luetzenkirchen (INT), Giacomo Bernini (NXW), Elena Bucchianeri (NXW), Pekka Pirinen (OUL), Giovanni Nardini (UPI), Giovanni Stea (UPI), Amina Boubendir (ORA), Ignacio Labrador (ATO), Ricardo Marco (ATO)
Contractual Date of Delivery:	31/12/2021
Dissemination level:	PU[1]/
Status:	Final
Version:	1.0
File Name:	Hexa-X D5.1_full_version_v1.0

Revision History

Revision	Date	Issued by	Description
v0.1	2021-10-11	Hexa-X WP5	First full version
v0.2	2021-11-01	Hexa-X WP5	Updated version after internal review
v0.3	2021-11-01	Hexa-X WP5	External review version
V0.4	2021-11-15	Hexa-x WP5	Updated version
V0.5	2021-11-26	Hexa-x WP5	PMT/GA review
V1.0	2021-12-23	Hexa-x WP5	Final version

[1] CO = Confidential, only members of the consortium (including the Commission Services)

PU = Public

Abstract

This document provides the results of the gap analysis of Hexa-X work package 5: “Architectural enablers for 6G”. In addition, the documents provide the general architectural direction for a possible 6G architecture. This includes identifying the necessary enablers (e.g., functions, algorithms, and enhancements) to support the new 6G architecture. Thereafter, the initial scope on the novel architectural enablers is described. These enablers allow an intelligent distributed network, new network topologies in a flexible way and enable an efficient deployment of future networks. The document also defines the research scope and the problems of the novel architectural enablers, to be further evaluated in the coming deliverables.

Keywords

6G architecture, Intelligent Networks, Flexible Networks, Efficient Networks.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101015956.

Executive Summary

This is the first deliverable of work package 5 (WP5), “Architectural enablers for 6G”, denoted D5.1. The main objectives of WP5 are to develop architectural components for 6G that support a new flexible network design, full AI integration and network programmability while, at the same time, streamline and redesign the architecture for a network of networks.

The work in WP5 started with Task 5.1 and the so called “Architecture transformation”. This included a gap analysis of the current mobile network architecture, and transformation necessary in the 6G timeframe. After completion, the other tasks in WP5, namely Task 5.2 “Intelligent Networks”, Task 5.3 “Flexible Networks” and Task 5.4 “Efficient Networks”, started the work of defining the enablers to fulfil the architecture transformation.

The deliverable D5.1 has two objectives. The first objective is to perform a gap analysis of the current architecture and thereafter establish the general architectural direction for a possible 6G architecture. This includes identifying the necessary enablers (e.g., functions, algorithms, and enhancements) to support the new 6G architecture. The second objective is related to the novel architectural enablers and how these enablers shall support the future use cases and the requirements. Thereafter, the initial scope of the novel architectural enablers is described. These enablers allow an intelligent distributed network, new network topologies in a flexible way and enable an efficient deployment of future networks.

The work with the architectural enablers is refined in the three tasks, i.e., Task 5.2 Intelligent Networks, Task 5.3 Flexible Networks, and Task 5.4 Efficient Networks. The proposed enablers for Intelligent Networks of Hexa-X are meant to facilitate dynamic adaptability of the network architecture to accommodate new use cases and deployment scenarios beyond what the current cellular networks could offer, while keeping the infrastructure and energy costs at acceptable and sustainable levels. Flexible Networks intend to enable extreme performance and global service coverage. This is achieved by developing solutions that are capable of managing local ad hoc networks, in coordination with the infrastructure, as well as distributing their functionalities between them and at the edge. The Efficient Networks will streamline the interfaces assuming a cloud-native RAN and CN, by separation of concerns i.e., clarifying responsibility and functionality of each network function.

The document also defines the research scope and the problems of the novel architectural enablers, which will be further investigated in the coming deliverables.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2G	2nd Generation mobile wireless communication system
3D	Three-dimensional
3GPP	3rd Generation Partnership Project
4D	Four-dimensional
4G	4th Generation mobile wireless communication system
5G	5th Generation mobile wireless communication system
5GC	5G Core
5G-PPP	The 5G Infrastructure Public Private Partnership
AP	Abstract Processing
ACL	Agent Communication Language
ADL	Architecture Description Language
AF	Application Function
AI	Artificial Intelligence
AIaaS	AI as a Service
AIS	AI Information Service
AI/ML	Artificial Intelligence / Machine Learning
AMC	Autonomic Management and Control
AMF	Access and Mobility management Function
API	Application Programming Interface
AUSF	AUthentication Server Function
B5G	Beyond 5G
BPEL	Business Process Execution Language
CAPEX	Capital Expenditures
CA	Carrier Aggregation
CBSE	Component-Based Software Engineering
CI/CD/CT	Continuous Integration/Continuous Delivery/Continuous Testing
CN	Core Network
CNCF	Cloud Native Computing Foundation
COMP	Coordinated Multipoint Transmission
CP	Control Plane
C-RAN	Centralized RAN
CRAS	Connected Robotics and Autonomous Systems
CS	Circuit Switching
CU	Central Unit
CRAN	Centralized RAN
D2D	Device-to-Device
DAPS	Dual Active Protocol Stack
DC	Dual Connectivity
DFP	Dynamic Function Placement
DL	Downlink
DLT	Distributed Ledger Technology
DMO	Direct Mode Operation
D-MIMO	Distributed Multiple-Input and Multiple-Output
DN	Data Network
DO	Data Objects
DRAN	Distributed RAN
DT	Digital Twin
DU	Distributed Units
E2E	End-to-End
EC	European Commission
eMBB	Enhanced Mobile Broadband
EMF	Electromagnetic Field
EN-DC	Enhanced Dual Connectivity
EPC	Evolved Packet Core
EU	European Union
eURLLC	Enhanced Ultra-Reliable Low-Latency Communication

FL	Federated Learning
FPGA	Field Programmable Gate Array
FSO	Free-space Optics
FWA	Fixed Wireless Access
CGM	Grid Component Model
CVM	Compute Virtual Machine
CPU	Central Processing Unit
CU-CP	Centralized Unit – Control Plane
CU-UP	Centralized Unit – User Plane
GDP	Gross Domestic Product
GGSN	Gateway GRPS Support Node
GNSS	Global Navigation Satellite System
GPRS	General Packet Radio Service
H2020	Horizon 2020
HAPS	High-Altitude Platform Station
HARQ	Hybrid Automatic Repeat reQuest
HLR	Home Location Register
HLS	High Layer Split
HSPA	High Speed Packet Access
IAB	Integrated Access/Backhaul
ICT	Information and Communication Technology
IoRT	Internet of Remote Things
IoST	Internet of Space Things
IoT	Internet of Things
ITU	International Telecommunication Union
KPI	Key Performance Indicator
KVI	Key Value Indicator
LCM	Life-Cycle Management
LED	Light Emitting Diodes
LEO	Low Earth Orbit
LiDAR	Light Detection and Ranging
LTE	Long Term Evolution
LWA	LTE WLAN Aggregation
MAC	Medium Access Control
MANA	Multi-agent-based network automation
MAPE	Monitoring-Analysis-Planning-Execution
MBB	Mobile Broadband
MC	Multi-connectivity
MDAF	Management Data Analytics Function
MDD	Medical Devices Directive
MEC	Multi-Access Edge Computing
MFAF	Messaging Framework Adaptor Function
MGW	Media Gateway
MIMO	Multiple-Input and Multiple-Output
ML	Machine Learning
MME	Mobility Management Entity
mMTC	Massive Machine Type Communications
MNO	Mobile Network Operator
MR	Machine Reasoning/Mixed Reality
MSC/VLR	Mobile Service switching Centre/Visitor Location Register
Multi-TRP	Multi Transmission and Reception Points Network as a Service
NaaS	Non-Access Stratum
NAS	NR-E-UTRA DC
NE-DC	Network Exposure Function
NEF	Network Function
NF	Network Function Virtualization
NFV	Network Exposure Function
NEF	Next Generation Mobile Networks
NGMN	Next Generation RAN
NG-RAN	NG-RAN EUTRA-NR DC

NGEN-DC	Non-Line of Sight
NLOS	Non-Terrestrial Network
NTN	Non-standalone NR (5G) network
NSA	New Radio
NR	NR-NR Dual Connectivity
NR-DC	Network Repository Function
NRF	Non-Public Networks
NPN	Network Slice
NS	Network Service Mesh
NSM	Network Slice Selection Function
NSSF	Network Data Analytics Function
NWDAF	Operations, Administration and Maintenance
OAM	Open Network Operating System
ONOS	Operating Expenditures
OPEX	Open Service Mesh
OSM	Open Radio Access Network
O-RAN	Operational Technology
OT	Optical Wireless Communication
OWC	Programming Protocol-independent Packet Processors
P4	Policy Control Function
PCF	Packet Data Convergence Protocol
PDCP	Packet Data Network Gateway
P-GW	Programme Making and Special Events
PMSE	Public Protection and Disaster Relief
PPDR	Programmable Protocol Stack
PPS	Proximity Services
ProSe	Packet Switching
PS	Public Switched Telephone Network
PSTN	Quality of Experience
QoE	Quality of Immersion
QoI	Quality of Service
QoS	Research and Innovation
R&I	Radio Access Network
RAN	Radio Access Technology
RAT	Representational State Transfer
REST	Radio Frequency
RF	Radio Network Controller
RNC	Radio Resource Control
RRC	Remote Radio Head
RRH	Radio Resource Management
RRM	Round-Trip Time
RTT	Radio Virtual Machine
RVM	Standalone NR
SA	System Architecture Evolution
SAE	Service-based Architecture
SBA	Service Based Interface
SBI	Service Based Management Architecture
SBMA	System Architecture Evolution
SEA	Service Component Architecture
SCA	Self-Controlled service Component
SCC	Stream Control Transmission Protocol
SCTP	Serving GPRS Support Node
SGSN	Serving Gateway
S-GW	Service Communication Proxy
SCP	Sustainable Development Goal
SDG	Software Defined Networking
SDN	Standards Developing Organization
SDO	Software Defined Radio
SDR	Simultaneous Localization and Mapping
SLAM	Session Management Function

SMF	Service Management and Orchestration
SMO	Service-Oriented Architecture
SOA	Simple Object Access Protocol
SOAP	Self-organized Network
SON	Source RAN
S-RAN	Supplementary Uplink
SUL	Smart Networks and Services
SNS	Total Cost of Ownership
TCO	Terrestrial Networks
TN	Target RAN
T-RAN	Time Sensitive Networking
TSN	Time to Market
TTM	Unmanned Aerial Vehicle
UAV	Universal Description Discovery and Integration
UDDI	Unified Data Management
UDM	User Equipment
UE	United Nations
UN	User Plane
UP	User Plane Function
UPF	Universal Resource Locator
URL	Ultra-Reliable Low-Latency Communication
URLLC	UMTS Terrestrial Radio Access Network
UTRAN	Vehicle-to-Network
V2N	Vehicle-to-Everything
V2X	Virtualized Infrastructure Manager
VIM	Visible Light Communication
VLC	Virtual Machine
VM	Virtual Network Function
VNF	Vehicle-to-Everything
V2X	Virtual Reality
VR	Work Package
WP	Web Services Description Language
WSDL	Wireless Termination
WY	Anything as a service
XaaS	Explainable AI
XAI	Extended Reality
XR	Extensible Markup Language
XML	Zero touch network & Service Management
ZSM	

1 Introduction

Hexa-X is one of the 5G-PPP projects under the EU Horizon 2020 framework. It is a flagship project that develops a Beyond 5G (B5G)/6G vision and an intelligent fabric of technology enablers connecting human, physical and digital worlds.

This document is the first deliverable of Work Package 5 (WP5). This first deliverable, D5.1, includes a gap analysis of existing architectures and proposes the direction the 6G architecture, i.e., the architecture transformation. We propose novel architectural concepts enabling intelligent distributed networks, new network topologies for addressing new use cases, and architectural enablers for cost-efficient deployment of 6G networks. In addition, D5.1 defines the scope of subsequent deliverables, defining the content and work in the so-called WP5 enabler tasks Intelligent networks, Flexible networks, and Efficient networks.

1.1 Objective

The objective of this document is twofold. The first objective is to perform a gap analysis of current mobile architecture and thereafter point at the architectural direction for a possible 6G architecture, i.e., the goal of the 6G architecture. The second objective of the document lays out our initial scope about architectural enablers for fulfilling 6G architectural direction.

1.2 5G network architecture background

There is a continuous increase in traffic in cellular networks and has been at least since 2G. Hence, mobile operators need to continuously update their existing networks to match the growing data demands. Since deploying a network takes a lot of effort, operators demand continuation and backwards compatibility from the new functionality added to the networks. The need for backwards compatibility also affects decisions regarding design of future network architectures.

A cellular network has two major parts: A Radio Access Network (RAN) and a Core Network (CN) [Sau21]. The RAN handles all radio related functionalities, and the CN is responsible for, e.g., switching and routing calls and data connections to external networks. From the beginning, traffic was Circuit Switched (CS), which means that a dedicated connection was established between the endpoints of phone calls using the Public Switched Telephone Network (PSTN). Important network elements in the CS CN were a Mobile service Switching Centre/Visitor Location Register (MSC/VLR), a Home Location Register (HLR), and a Media Gateway (MGW). The HLR is a database located in the subscriber's home system that stores the master copy of a user's service profile. This profile also includes information about the subscriber and all allowed services, forbidden roaming areas and information on supplementary service information. The MGW can be used for transmitting and converting the User Plane (UP) traffic in both a CS CN and a Packet Switched (PS) CN as a border element between different kinds of networks.

While the first 2G users were happy just to be able to make phone calls, the internet evolved. As more and more mobile users saw a need for accessing the internet, 2.5G was introduced with support for IP networks. This was referred to as General Packet Radio Service (GPRS). Thus, mobile phones now had support for both a CS and a PS network. The CN was updated with a Serving GPRS Support Node (SGSN) and a Gateway GRPS Support Node (GGSN), see Figure 1-1. The main task of the SGSN is to deliver data packets from and to the UEs within its geographical service area.

When 3G was deployed it had both CS and PS connectivity from the beginning. The PS traffic was handled via the GGSN and SSGN introduced for 2.5G (see Figure 1-1). Even though 3G did have support for PS traffic from the beginning, the capacity was limited (max rate was 384 kb/s). There was a clear trend in the industry to move towards PS, so with the sixth release of the 3rd

generation partnership project (3GPP) specification (Rel-6) high speed packet access (HSPA) and enhanced uplink (EUL) were introduced. With HSPA the theoretical downlink (DL) bitrate was 14 Mbps.

The changes to the CN that provided the big improvements to performance for 3G were direct tunnelling, allowing UP to bypass SGSN, and evolved HSPA, in which UP data further bypasses the RNC and offloads data directly to the Internet via the GGSN. When the direct tunnelling and evolved HSPA solutions are combined, only two network elements remain in the UP, which increases the flexibility in network topology and allows the SGSN node to be optimized for Control Plane (CP). Direct tunnelling was not required to meet the performance of evolved HSPA; however, this was the first step towards system architecture evolution (SAE), which is the base for 4G.

4G or UMTS Terrestrial Radio Access Network (UTRAN) Long Term Evolution (LTE) and SAE define an all-IP network. In LTE/SAE there are many new CN nodes, such as the Mobility Management Entity (MME) and the SAE gateway (S-GW and P-GW), see Figure 1-1. The only remaining RAN node is the eNB. The SAE includes a GW towards the actual network (S-GW), but it also includes a GW to external networks (P-GW). Both GWs process UP data and handle tasks related to mobility to networks with other 3GPP Radio Access Technologies (RATs). The main driver for having both an S-GW and a P-GW was to support roaming (with P-GW in the home network), and possible regional mobility (with a P-GW anchor. Note that co-deployment of S/P-GW is possible (and frequently used). The Mobility Management Entity (MME) handles CP signalling, especially for mobility and idle mode handling. Subscription-related information is handled by a node Home Subscriber Server (HSS), which is similar to HLR for 2G and 3G.

A major benefit of going for all PS is that the LTE RAN does not need to support the extra complexity of supporting two core-networks as required by 3G RAN. A drawback of providing a PS-only network is that interworking with previous 3GPP networks can be complex, e.g., in case an LTE UE receives a phone call (CS) from a 2G phone the LTE device has to switch to 3G in order to be able to answer the call.

Finally, in the 5G CN (5GC) the CP functionality has been slightly reorganized, see Figure 1-1. Instead of the MME and S-GW in EPC, the functionality has been divided between a Session Management Function (SMF), and an Access and Mobility management Function (AMF). There can be multiple SMFs associated with one UE [23.501]. The EPC functionality for access/mobility from the MME are reallocated to the AMF to contain all access and mobility functionality in one node. Functions related to UP processing, i.e., P-GW in 5G, are handled by a User Plane Function (UPF).

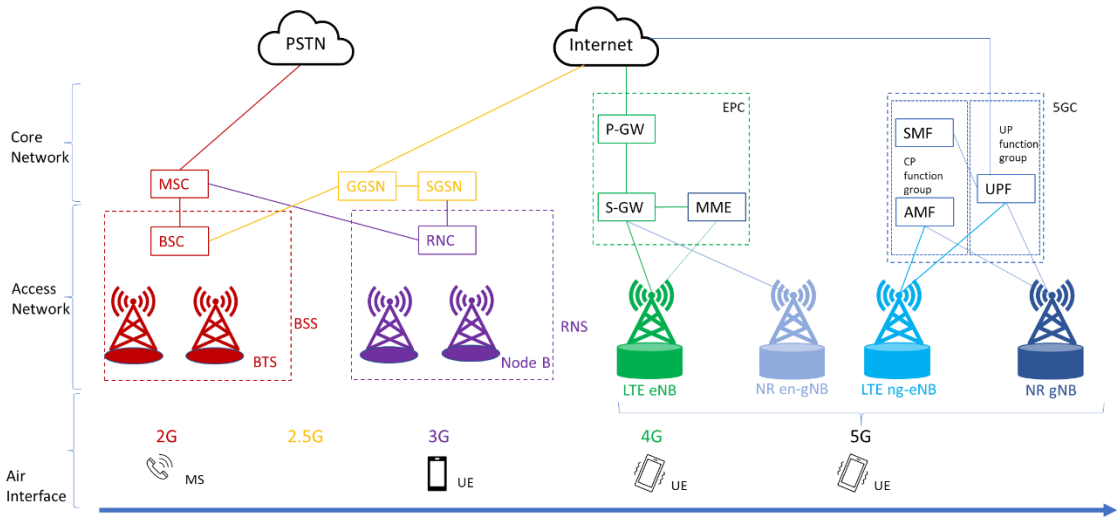

Figure 1-1 How architecture has changed with time. [adapted from 3g4ghist]

Already with 5G there is the possibility to place RAN functionality in different locations in the 5G RAN architecture, e.g., distributed RAN (DRAN) or centralized RAN (CRAN), see Figure 1-2. The option that is most appropriate for a particular deployment primarily depends on the type of deployment area (urban, suburban, or rural) and the availability of fibre. For both DRAN and CRAN, it is possible to add a Virtual RAN (VRAN) by implementing a higher layer split (HLS) where the gNB is partitioned into a central unit (CU) and distributed units (DUs). This is known as HLS-VRAN.

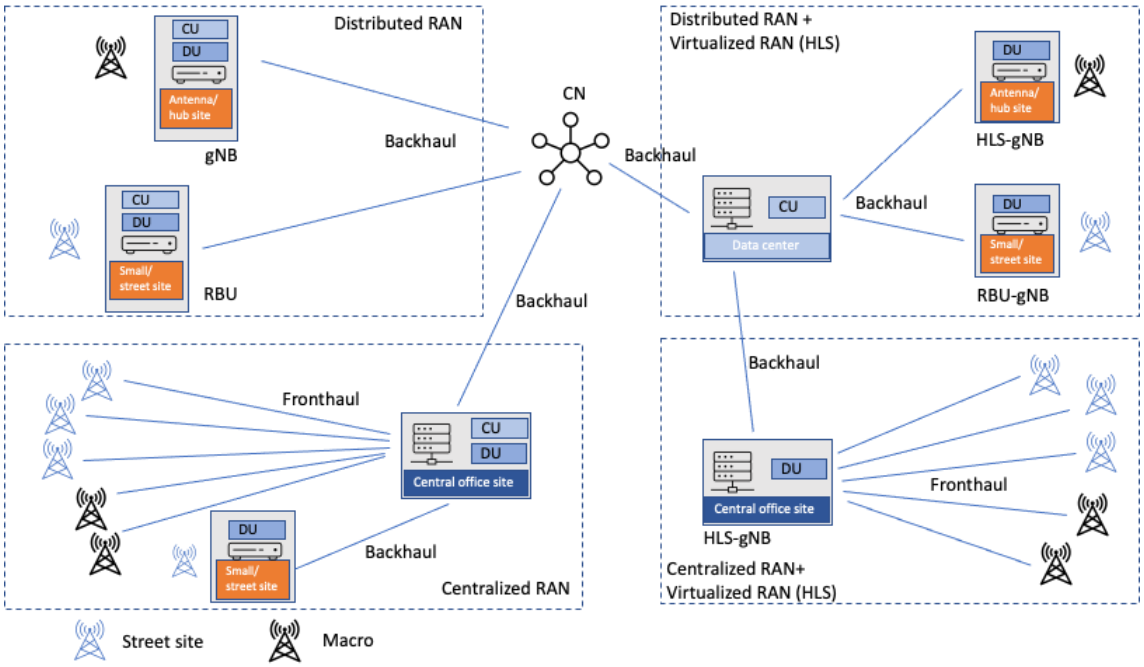

Figure 1-2 Different types of deployment (adapted from [EFA+19])

The commercial introduction of 5G was based on the so-called 3GPP Rel-15 E-UTRA NR Dual Connectivity (EN-DC). This network topology, also known as Non-Standalone (NSA) 5G, heavily leverages the network equipment already deployed for the 4G Evolved Packet System (EPS), i.e., the EPC and the LTE RAN (E-UTRAN), see Figure 1-3. With respect to the previous pre-Rel-15 LTE networks, EN-DC allows to increase the user data rate by using radio resources provided by the NR base station tightly inter-working with the LTE network. In this sense, EN-DC is basically a network topology suited for enhanced Mobile Broadband (eMBB) applications.

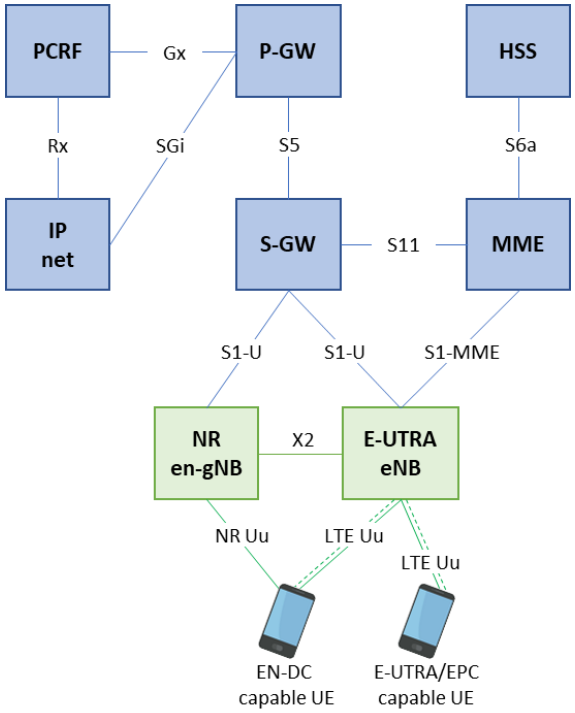

Figure 1-3 - E-UTRA NR Dual Connectivity (EN-DC) architecture

EN-DC allows operators to rapidly deploy 5G even if it cannot fully support all the 5G-specific services and features such as Ultra Reliable Low Latency Communications (URLLC), network slicing, etc. EN-DC also allows operators to use EPC instead of 5G Core network (5GC). However, 5G addresses not only eMBB but also URLLC and massive Machine Type Communications (mMTC) use cases, hence allowing operators to address, via standardized solutions, the new emerging market needs coming from the so-called Verticals, going beyond eMBB services. Especially for URLLC services, additional 5G network topologies have been defined making use of two main key enablers: the new NR RAT and the new 5GC. Some Asian and American operators are now starting to deploy Standalone (SA) 5G networks [5GSA] [5GSMA], see Figure 1-4, migrating their EN-DC based networks directly to 5G SA. Note that operators have different strategies concerning the migration path towards full 5G networks, some of which have extensively been analysed in [5GSMA] and in 3GPP [38.801].

Figure 1-4 also shows other options such as the NG-RAN E-UTRA-NR Dual Connectivity (NGEN-DC), where NG-RAN stands for Next Generation RAN, as well as NR-E-UTRA Dual Connectivity (NE-DC). NGEN-DC has the same configuration of EN-DC with the only difference that 5GC is used instead of EPC, hence there is an LTE anchor node (i.e., the Master Node) which manages both CP and UP connectivity to the UE and a NR node (the Secondary Node) which provides additional UP resources to the same UE. In NE-DC, instead, it is the NR node that plays the role of the Master Node while the LTE node acts as the Secondary Node, both RATs still connected to the 5GC. Note that Figure 1-4 does not include the NR-DC case (dual-connectivity between two NR nodes) and omits the case where non-3GPP accesses connect to 5GC. Moreover, Figure 1-4 shows the 5GC represented by means of the so-called Service Based Architecture (SBA) design approach, introduced by 3GPP on top of the traditional reference point and interface approach in which each pair of Network Functions (NFs) communicates with each other using a pre-established peer-to-peer signalling interface; in fact, from an architectural viewpoint, one of the most significant changes with respect to the EPS is in how core network functions communicate to each other. With SBA, 5GC NFs, e.g., AMF and SMF, Policy Control Function (PCF), etc., can be implemented as a set of software-defined services; each service is provided by a service producer and can be consumed by one or more service consumers, hence allowing 5GC-specific procedures to efficiently expose and consume services. Different messaging models can

be used: for simple services or information requests a request-response model can be used while, for more advanced and complex processes, the framework supports a publish-subscribe (and notify) model. Finally, another intrinsic SBA feature is that it copes with NFs' load distribution by design, in the sense that, during the process of NF selection, the dynamic load of the candidate NF instances is taken into consideration.

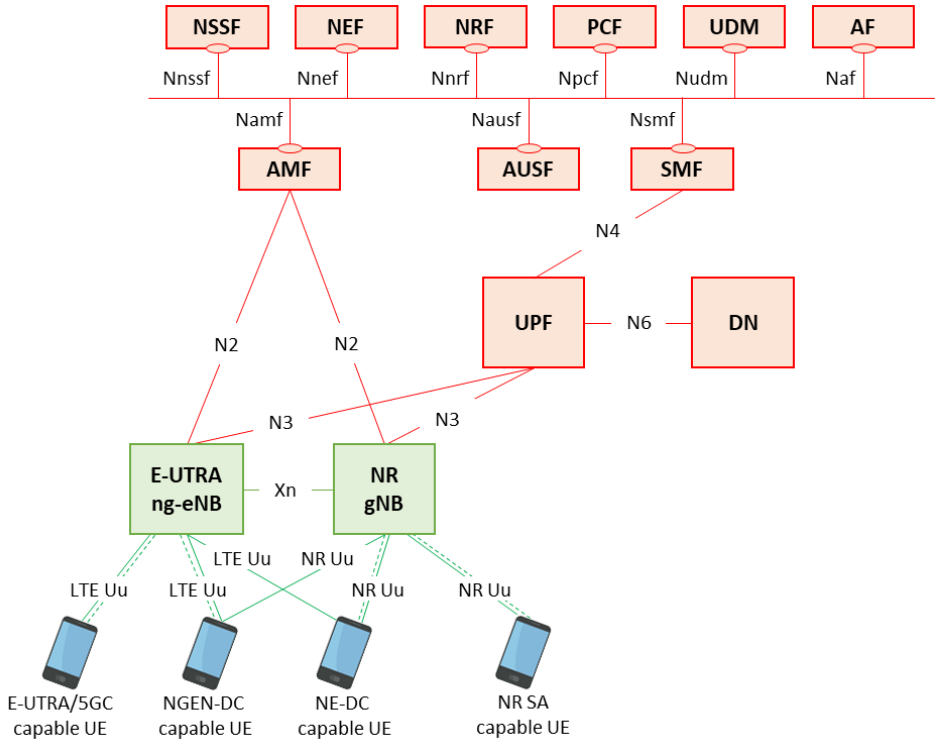

Figure 1-4 – 5GC-based network architectures leveraging both NR and E-UTRA

It is worth to note that architectures in Figure 1-3 and Figure 1-4 are able to interwork in a standardized manner by means of 3GPP-defined interfaces, e.g., the N26 interface connecting the MME of the EPC with the AMF of the 5GC [23.501][2]. However, at the time of submitting this deliverable, such interworking has only been specified by considering the traditional reference point and interface approach as well as by introducing the so-called "combo-nodes" implementing EPC- and 5GC-specific network functions simultaneously, e.g., the "UPF + PGW-U" logical network entity as in Figure 4.3.1-1 of [23.501] coping with the UP data of the UE during e.g., inter-system mobility. At least for Rel-18 there are no plans to standardize the interworking between the EPS and the 5G System (5GS) interworking by means of the SBA approach due to the significant effort that such standardization activity might require, since the EPC network entities were architected to be implemented on physical nodes that were virtualized afterwards, hence not designed to be virtualized from the outset as for the 5GC.

1.3 Structure of the document

As said in Section 1.1, there are two main objectives for this document, and this is also reflected in the structure of the document. The methodology of the document is described in Chapter 2. The first part of the document corresponds to the first objective. The first part includes Chapter 3, which describes the technology trends that may affect the 6G architecture, Chapter 4, which

[2] The 3GPP standard also supports inter-RAT handover without the N26 interface.

presents a gap analysis of current architecture, and Chapter 5 which presents a possible 6G architecture transformation, i.e., the direction and principles we think is needed for a successful 6G architecture. The second part of the document, (Chapter 6 , Chapter 7 and Chapter 8) describes the initial architectural enablers necessary to fulfil the 6G architectural principles. Finally, we summarize the documents with the new architecture KPIs based on the architectural enablers in Chapter 9. A brief description of the proof of concepts that belongs to WP5 are found in Chapter 10, and the conclusion is found in Chapter 11.

The main terminology used in this document is described in Annex A.1.

2 Methodology

The methodology for this document is to first investigate possible trends that may affect the architecture, see Figure 2-1. Thereafter we identify gaps of the current architecture or possible necessary improvements due to new use cases defined in [D1.2]. These gaps (or improvements) form the basis for the architectural direction, or as we call it, the architecture transformation. The second objective of this report is to initiate the scope of the so-called enablers of the 6G architecture. These enablers are the important components necessary to achieve the 6G architecture transformation. As shown in Figure 2-1, the architectural transformation will influence the enablers, but the enablers will also influence the overall architecture direction.

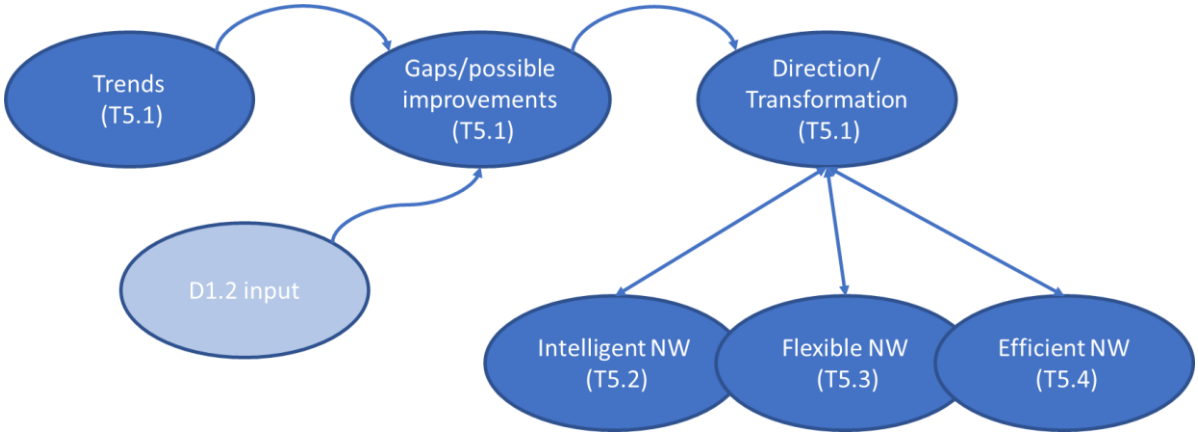

Figure 2-1 Methodology steps of this document

2.1 D1.2 Use cases

D1.2 identified 23 use cases and clustered them into 5 families, see Figure 2-2. The use case family *sustainable development* addressed the challenges associated with sustainable development. It includes use cases such as global service coverage, energy-optimized infrastructures, and e-health for all. *Massive twinning* use case family is the use of digital twins to represent, interact and control actions in the physical world. The *telepresence* use case family covers immersive telepresence for enhanced interactions, involving mixed reality or merged reality, providing extreme and fully immersive experience. The use case family *robots to cobots* includes how robots are interacting with robots. The use case family *local trust zones* include in-body networks to wide area deployment of sensors networks.

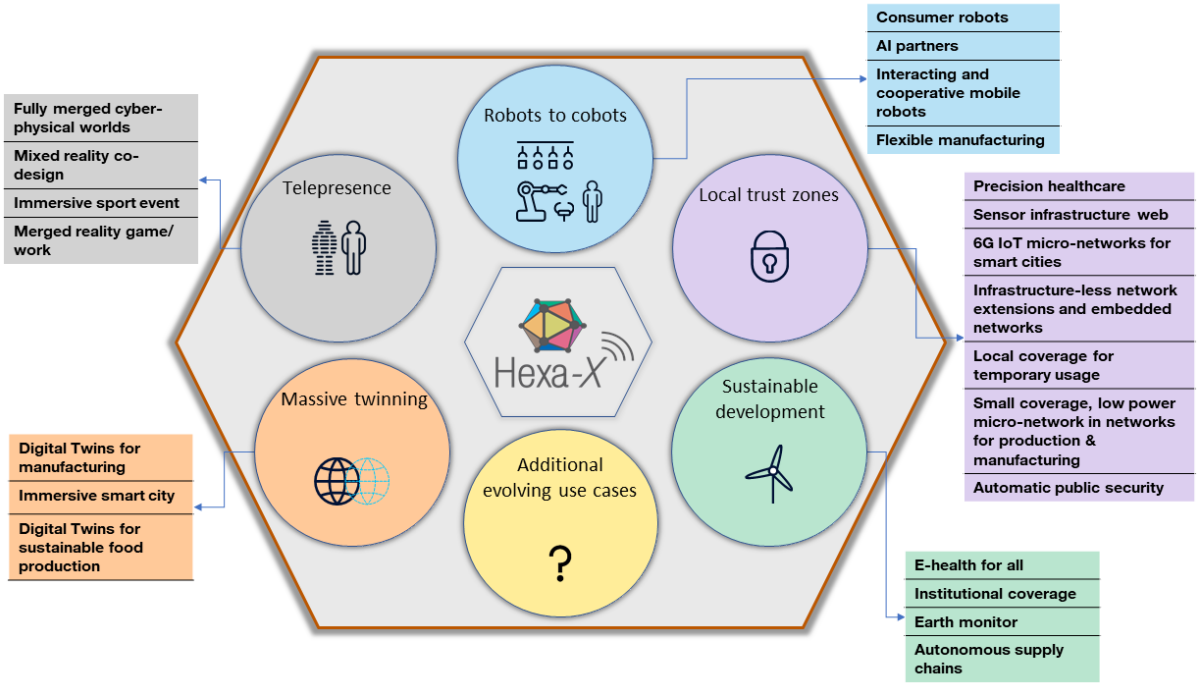

Figure 2-2 Use cases from [D1.2]. There are 5 families of use cases, and in total 23 use cases.

The five use case families and their impact to the architecture are summarized in Table 2-1. The middle column indicates the expected impact the use case families have on the 6G architecture. The right-most column shows the impacted WP5 Tasks for the use case family, and also specifies some of the specific use cases that impacts the WP5 task. Which specific use case that impacts the tasks are described in the enabler chapters, i.e., Chapter 6 , Chapter 7 and Chapter 8.

Table 2-1 Use case families from [D1.2]

Use case family	Expected impact on architecture	Impacted WP5 Tasks and an example of the use case in parenthesis.
Sustainable development	High	High: Task 5.3 (NTN global coverage). Medium: Task 5.4 (signalling with IoT, energy savings) and 5.2 (AI)
Telepresence	Medium	Medium: Tasks 5.2 (connected intelligence)
Robots to cobots	High	High: Task 5.3 (Network of networks). High: Task 5.2 (Robots will utilize the connected AI capabilities offered by 6G for situation-aware cooperation and collaboration and assistance)
Local trust zones	High	Task 5.4: sharing of sensor data, computing Task 5.3: ad-hoc networks
Massive twinning	High	High: Task 5.4 (massive sensors and signalling). High: Task 5.2 (Immersive smart city, AI, actuators, CaaS, AIaaS).

Moreover, 6G is going to host future Industry X.0 paradigms, with intelligent management of distributed industrial factories, with hundreds of collaborating robots. These use cases have already been popular during 5G design and they will become massively employed in future 6G networks.

2.2 D1.2 Performance indicators

In addition to this, [D1.2] has defined several Key value and performance indicators (KVIs/KPIs) that may be relevant to the architecture. A summary of the KPIs is listed in Table 2-2. Several of these KPIs will affect the architecture evaluation, such as the energy efficiency and flexibility.

Table 2-2 Key value and performance indicators from [D1.2]

KVIs and KPIs	Expected impact on architecture	Related WP5 task
Advancing on existing KPIs		
Data rate	Low	NA
Capacity	Medium	NA
Localization	Medium	All tasks
Connection density	High	All tasks
Redefinition of existing KPIs		
Service availability	Medium	Task 5.3
Deterministic services and dependability	Medium	NA
Coverage	High	Task 5.3, see sustainable coverage KPI in [D7.1] also
Network energy efficiency	Medium	Task 5.4
New KPIs		
Integrated sensing	Medium	Task 5.4
Local compute integration	High	Task 5.2
Integrated intelligence	High	Task 5.2
Embedded devices	Low	NA
Flexibility	High	Task 5.3

To summarize, the main input from the use cases in [D1.2] is the wide range of applications the use cases represent. To support the wide range of use cases, the 6G network architecture must be a combination of *intelligent* (see Chapter 6), *flexible* (see Chapter 7) and *efficient* (see Chapter 8). In addition to this, the KPIs also affect the architecture, or more precisely, how it shall be evaluated. This will be investigated further in Chapter 9.

3 Trends and State-of-the-art

This chapter identifies several trends that are important for a 6G architecture. These trends are later used for the gap analysis in Chapter 4. To this end, the new emerging applications and use cases are recalled from [D1.2] as well as new technical and architecture-specific trends (e.g. [SAP+21], [ZFW+19], [6GSam20]).

3.1 Standardisation trends

3.1.1 AI and management

The future of the network architecture might comprise fully end-to-end machine learning and model accessibility that would enable exploitation of highly autonomic networking. This means

taking advantage of AI/ML capabilities such as employing unsupervised machine learning algorithms and reinforcement learning, in addition to supervised learning. The architectural aspects of autonomic networks based on AI/ML are dealt with in standardisation efforts of ETSI Generic Autonomic Networking Architecture (GANA) [E16] and ETSI Industry Specification Group (ISG) Experiential Networked Intelligence (ENI) [E19]. In parallel, 3GPP SA5 has also started working on ML/AI for 5G [210520].

The architectural aspects of autonomic networks based on AI/ML are dealt with in standardisation efforts of ETSI Generic Autonomic Networking Architecture (GANA) [E16] and ETSI Industry Specification Group (ISG) Experiential Networked Intelligence (ENI) [E19].

Regarding design and standardisation of an intelligent architecture, the work of ETSI ISG ENI started around 2017 with the publication of a white paper [EE17]. The main scope of ETSI ISG ENI is to design an intelligent architecture for network management, considering context-aware policies to adjust network service provisioning according to users' needs, environmental conditions, and business goals. By considering the general architectural characteristics, the ETSI ENI system considers the heterogeneity of the existing and future network infrastructure hardware together with the full virtualisation capabilities obtained via the ETSI NFV MANO architecture. On top of that, there is the AI of ENI, which interacts with and manages the specific assisted systems distributed across the network.

The white paper [EE17] defines a list of the requirements for the ETSI ISG ENI architecture. These requirements are important to evaluate and control how AI works within the network and the applications to improve network operations, management, and service provisioning. Finally, the use cases that have been identified for the ETSI ENI system can be categorised into five main groups: infrastructure management, network operations, service orchestration and management, assurance, and network security.

At the same time, ETSI has also been investigating a reference model for autonomic networking, cognitive networking and self-management of networks and services. This specific architecture is called ETSI GANA, developed in 2016. This architecture is directly inspired by the idea of self-organizing networks (SON). It provides a more general reference able to interoperate with complementary technologies such as SDN, NFV, and big data analytics for autonomic management and control (AMC). The idea of AMC relies on the definition of the decision-making-element (DE), which is an autonomic function comprising a cognitive control-loop in centralised/distributed management and control planes. The DE owns self-* features such as self-configuration, self-optimisation, self-healing, etc. Each DE is an adaptive entity, which dynamically monitors and manages its respective managed entities. Practically, a DE is placed within a network node at a specific layer of the protocol stack. Additionally, each DE can be either a hardware-based or a softwarized entity.

This brief excursus on ETSI ENI and GANA has been important to underline the concept that the complete integration of AI into 6G cannot neglect the experience provided by ETSI in the recent years. Significant parts of research on intelligence in 6G have not even been mentioning these activities. However, ETSI ENI and GANA architectures will be pivotal for 6G standardisation as ETSI MANO SDN-NFV architecture has been fundamental for 5G.

3.1.2 RAN and CN

The trend of using SBA, which gives flexibility with respect to the traditional reference point and interface model, is expected to be further enhanced for the representation and interaction between network domains, e.g., between RAN and CN. In fact, without SBA, whenever a new function is introduced in the system, the existing NFs need to be enhanced to support the new functionality and also a new peer-to-peer interface needs to be defined between the new function and the existing NFs that communicate with it. Conversely, with SBA, the services provided by a service producer, although initially defined for a specific service consumer (or a set of service consumers), can later be also made available to additional consumers, if needed. The adoption of

SBA hence allows for better scalability with 'plug-and-play' approaches, reducing cycles for service innovation by following the Continuous Integration/Continuous Delivery/Continuous Testing (CI/CD/CT) pipeline for a quickly evolving system architecture.

As seen in Figure 1-3 in Section 1.2, the 5G system relies on the coexistence of two RATs, i.e., the evolution of LTE from 3GPP Rel-15 and NR. Therefore, as also stated in [5G6GNTT], an open question is whether the future 6G mobile system will follow this approach in the sense that the 5G RATs evolution may coexist with a brand-new RAT, enabling a stepwise 6G deployment heavily leveraging the continuously evolving 5G networks.

3.1.3 Cloud and SBA

In 3GPP, some effort was made to better support the cloud native implementation with the SBA in the 5G CN. This cloudification trend is likely to continue for 6G with inclusion of service-based interfaces also for UE and RAN. The key challenge is to design a future 6G architecture, which is able to fully utilize the cloud platform with regards to speed of development, reuse of common cloud components, and balancing the need to standardize critical business interfaces with the fast evolution of IT tools.

The current trend in the standardisation efforts is the harmonization of several Multi-access Edge Computing (MEC) architectures into a common one. An example is synergized harmonization of ETSI MEC and 3GPP SA6 Edge Application architectures [ETSI20]. However, the evolution of the edge is going much further: the whole architecture and its functions are designed based on reusable and composable microservices that will enable dynamic workload scheduling to optimal execution points in a hierarchy of data centres across the network, matching the latency and scalability needs.

3.2 Cloud trends

Cloud transformation of businesses, services, and networks is a strong trend that is likely to continue in 6G. There are two trends in play when applying cloud technology into networking. Diversification of deployment by specialization is going to be pivotal. Additionally, the competition, that is mostly driven by over-the-top cloud providers, and the harmonization of cloud computing standards, driven by ETSI and 3GPP, are also fundamental trends.

Another trend is network control using cloud solutions. The Open Network Foundation (ONF) is developing an enhanced open-source stack to provide true network control, zero touch configuration, and verifiable/secure networking. Micro Open Network Operating System (µONOS) [O21], one of ONF's announced projects, is the next generation SDN platform. µONOS is a cloud-native platform using micro services in containers based on Kubernetes and edge-cloud. Its goal is to provide a comprehensive platform for operations, covering aspects such as configuration, control, monitoring, verification, live update, and diagnostics for 5G RAN edge.

Furthermore, micro service-based architecture is also going to evolve towards more general multi-agent systems [ABG+21], where intelligent softwarized agents (in-network functions and operations, or services) will collaboratively work in logical chains. These agents will still work in software environments like containers, running in servers.

MEC federation mechanisms are expected to provide geographically distributed services (e.g., MEC-enabled V2X, AR/VR gaming), by leveraging cooperation among several Mobile Network Operators (MNOs) [ETSI21]. By allowing cooperation among multiple MNOs' MEC systems through well-defined interfaces, a MEC system participating to a MEC federation can access services and resources provided by other MNOs' MEC systems (i.e., through east/westbound interfaces). This is accomplished through the federation manager running in the MEC system, which may communicate with other MNOs' federation managers either directly (i.e., peer-to-peer topology) or via a federation broker acting as intermediary (i.e., star topology) [GSMA20]. This

raises opportunities, as well as challenges. For instance, an AI-based V2X service running on MNO A's MEC system can be trained using real-time information coming from vehicles operating in its own domain, as well as from vehicles served by MNO B's MEC system, hence improving its local AI model to perform better predictions. However, interactions among different MEC domains must preserve data integrity and privacy, as well as keep communication overhead low.

Serverless computing has emerged as a compelling paradigm for the deployment of cloud applications, largely due to the shift of enterprise application architectures to containers and microservices. It is an architecture in which a service provider accepts code from a customer, dynamically allocates resources to the job and executes it. The main benefit of a serverless approach compared to the classical cloud infrastructure is the operational cost. In cloud computing, a customer rents a certain number of resources from a provider. The amount of rented infrastructure resources could be overestimated because peak traffic can vary over time, thus preventing multiplexing gains and implementation of energy efficiency-oriented resource allocation strategies. This means that the customer can potentially have paid for resources that in the long run are not fully used, resulting in poor investment, see Figure 3-1.

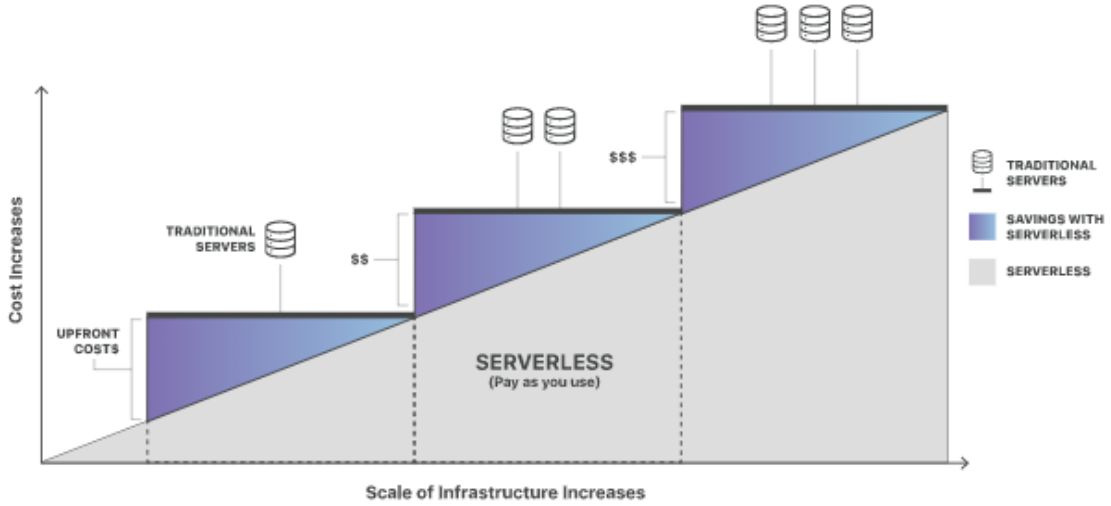

Figure 3-1 Cost Benefits of serverless [SB].

With a serverless approach, the customer can pay based on the actual usage of infrastructure resources. In average, unless very convenient cost models are implemented for the traditional resource rental approach, this results in a more efficient cost infrastructure / resource usage ratio, as shown in Figure 3-1.

3.3 AI trends

Artificial intelligence (AI) and specifically machine learning (ML) have the potential to impact the network architecture in almost every aspect, thanks to the enormous development of computing power, the power of cloud-based computing, and the availability of data to be mined and processed. For the purpose of providing precise analytics, data collection processes (e.g., how data is collected) is not the unique relevant point, but the quality of the data should also be considered This means that consistency, integrity, accuracy, and completeness of the collected data should be ensured to have efficient or minimal pre-processing to reach most precise and timely analytic outcome.

Federated learning (FL) is an ML technique that leverages input from multiple sources, which collaborate to train their local machine-learning models without exposing their own raw data. With respect to FL, MEC offers a low-latency alternative to cloud computing by hosting a central server and allowing that server to leverage local context information (e.g., UEs' positions, channel

quality information, etc.) that would otherwise be unavailable to them. This – on one hand – raises different privacy issues, and – on the other hand – promises increased effectiveness through context-awareness. [LLH+20] presents an overview of opportunities and challenges for MEC-based FL, with the dual focus of FL *at* MEC (i.e., using MEC to run FL) and FL *for* MEC (i.e., using FL to optimize MEC functions).

Parallel to FL, the concept of explainability in AI has been receiving more and more attention, towards achieving trustworthy AI [HLE19], which is important in several fields (institutional and policy-making, multi-party services, critical services): **eXplainable artificial intelligence (XAI)**, i.e., the set of techniques that ultimately allow an audience to receive "*the reasons of a model to make its functioning clear or easy to understand*" [ADS+20] has thus been investigated, with the purpose of either devising ML models, which are explainable *a priori* (the so-called "transparent-box design" strategy), or adding post-hoc explainability ("explaining black-box" strategy).

6G applications will drive the need for an intelligent SON aiming at implementing **autonomous networks** being able to manage network operations and resources in such a way that the network's KPIs are ensured under highly dynamic and complex environments. The "intelligence" is ensured by AI so that the autonomous (sub-)networks will be able not only to adapt their functions but also to sustain their resource usage and management.

3.4 Device trends

Some devices will have limited processing capacity, no input device nor a display. Some devices may consist of disaggregated functionality distributed across different processing points with multiple interfaces, making the logical device merely as a network rather than a single monolithic physical device. End devices may join another device in a peer-to-peer manner as well as hierarchical manner forming a completely independent "subnetwork". Also, there is a trend in 5G that functionality of wearable devices gradually replaces those of smartphones and this trend is likely to continue also in 6G, where devices such as smart wearables, integrated headsets and smart body implants might bring the end of the smartphone as known today [SBC20].

5G addresses communication between machines, whereas 6G will look more closely at cooperation between humans and machines in both natural and virtual environments. Examples of the types of communication interfaces are, e.g., **human machine interface** and **brain computer interface**. For example, while today's internet democratizes information sharing, the **Tactile Internet** has set out to democratize access to skills and expertise to promote equity for people of different ages, cultural backgrounds, or physical limitations [FLS+21].

3.5 Network flexibility

Network flexibility is a very wide area but typically means that a network can be deployed for many use cases, and for different architectures. One network flexibility aspect is the **modularization** trend, that is, functional decomposition of the RAN – it is worth to note that network flexibility deals also with hardware/software decoupling in every network element, i.e., the so-called vertical disaggregation. As part of the modularization, 3GPP has standardized a split of the RAN into a Central Unit (CU) and a Distributed Unit (DU), with the CU being further decomposed into its CP and UP counterparts, that is, CU-CP and CU-UP respectively [38.401]. Another example is the O-RAN Alliance, which attempts to further modularize the RAN, as operators are pushing towards 5G deployments based on standardized open interfaces and programmable, software-driven virtual RAN (vRAN) architectures [ORAN21]. Following this trend of modularization, it could be expected that 6G will see an even finer decomposition of functionalities in the RAN, possibly with the need to define new open, standardized interfaces to ensure multi-vendor interoperability. As the 6G should be based on a lean architecture, the number of new interfaces should be kept to a minimum, ensuring deployment flexibility and

lowering the Total Cost of Ownership (TCO). However, a high-level of decomposition may bring design complexity, may limit the degrees of freedom for implementation, may require a higher number of message exchanges and, potentially, may result in higher latency. It is also critical that modularization is done in the right way, ensuring a good separation of concerns while avoiding unclear division of responsibilities between different entities. Such unclear division could have a negative impact on performance, lead to duplicated functionality and increased implementation complexity. It is therefore not clear in what direction the modularization trend may lead. Another trend in this area is **3D networking**, which goes more and more towards the seamless inclusion of satellites and aerial platforms in 2D terrestrial communications. Modularization and network softwarization may be useful for enabling 3D networking. As an example, the functional split of RAN will enable placement of virtual network functionalities on aerial platforms such as High-Altitude Platforms Stations (HAPS) or in Low Earth Orbit (LEO) nanosatellites (e.g., cubesats) [BGS+20, BBG+20].

Nowadays, a steadily increasing number of services can be observed that require mobile network communication. These services require subnetworks with ad-hoc available communication or highly reliable communication y. Further on, there is a trend for services requiring multicast/broadcast/geocast transmission modes, services requiring direct D2D communications, and/or services requiring D2D plus relaying. These networks may be **autonomous subnetworks** where different user devices will connect to each other to form subnetworks that provide some specific service or function. A common factor to these subnetworks is that the service that they provide critically depends on the subnetwork connectivity between the subnetwork parties. Such subnetworks are managed by the infrastructure network even though they are autonomous and can work without continuous connectivity to the infrastructure network.

Network programmability and software defined networking SDN arose in the response to an "ossified" internet that made innovation a daunting task when it comes to how the network is evolved, controlled, and managed. SDN technology enabled faster paced innovation and management of the networks and enhanced network programmability with open interfaces, making it possible to customize packet forwarding based on specific header fields. towards a model-driven management of transport network connectivity [FGS20]. In SDN, apart from control and data plane, there is a management plane programmability obtained using standard open interfaces that permits to increase level of network automation and provides openness to the network. Network flexibility was further increased with the introduction of data plane programmability, enabling the customization of packet processing pipelines in forwarding devices [FGS20]. This trend of programmability is likely to affect mobile systems in the 6G timeframe, both for the network as well as the end user devices. In emerging highly heterogeneous radio and network technologies, the coordination across different protocol stacks is getting harder and harder and less flexible. The complete softwarization and programmability of future 6G networks may lead to support of flexible and adaptive protocols and network layers [FGS20].

Another trend is the concept of **cell-less** architecture, also known as cell-free, where a UE communicates with **cooperative base stations/access points** via CO**ordinated MultiPoint** (COMP) transmission and reception techniques instead of connecting to a single base station, enhancing connectivity performance, and lowering the latency of a traditional handover process. One example of this is a distributed MIMO network, where the user is connected to multiple access points belonging to the same logical entity. To some extent, such approach has already been used in 3GPP Rel-16 with the introduction of the Dual Active Protocol Stack (DAPS) handover aiming at reducing the handover interruption time [38.300]. The cell-less architecture (or similar concepts) is regarded as very important in 6G due to the highly heterogeneous deployment of different communication systems, i.e., the network of networks concept, where UEs supporting a number of heterogeneous communication technologies move seamlessly from one network to another by exploiting multi-connectivity techniques, hence avoiding the handover issues in terms of failures, delays, data losses and "ping-pong" effect. By continuously choosing the best link from the available heterogeneous links (e.g., mmWave, THz and OWC) in an automated manner, the user will be connected to the network as a whole (i.e., via multiple

complementary technologies) and not to a single cell, hence going beyond the concept of cells in wireless communications. However, due to e.g., transport network or processing limitations, some higher layer clustering of access points might be needed which, in turn, might require a CP protocol similar to the Radio Resource Control (RRC) as in today's networks to properly manage UE's mobility across clusters of access points. Finally, according to the specific use case, the UE may also concurrently use different network interfaces to exploit their complementary characteristics, e.g., the sub-6 GHz layer for the CP and a THz link for the UP [GPM+20].

Private networks or Non-Public Networks (NPNs) in 5G are designed to serve only one factory or village without necessarily guaranteeing mobility to other networks [FGS20]. For example, 5G Campus solutions can be deployed large operators to procure their own 5G NPN network in addition to public 5G, which is then under the full control of the operator. This allows the operator to customize the network with technologies from specific country/regions. Small operators can also benefit from 5G Campus solutions. For example, a 5G Campus can cover an entire region, where companies train robots. The further development of campus network architectures could be an important component of 6G research. the ability of interacting with public land mobile networks, thereby resulting in a combined infrastructure. This can enhance the capability to allow end-to-end service management and mobility for devices without dual SIM connectivity. NPN can combine the use of multiple wireless access technologies, including radio but also WiFi-x and LiFi [802.11bb]. An example in this context can be the work from 5G-CLARITY project [CGG+20].

3.6 Performance trends

It is expected that new applications such as Augmented Reality (AR) and Virtual Reality (VR) will keep the need for higher bitrates growing in an exponential rate [D1.2], motivating a need for more spectrum resources and subsequent further exploration of frequencies beyond 71 GHz as currently considered in 3GPP [38.807], but also higher reliability, which will be more challenging to meet at high frequencies. For both wireline and wireless communication there is a trend towards using more than one path to increase both throughput and reliability. The 5G standard includes both dual connectivity between LTE and NR (EN-DC, NGEN-DC, NE-DC), between two NR nodes (NR-NR dual connectivity), NR carrier aggregation and supplementary uplink.

One performance trend is the recent development of new transport protocols and congestion control algorithms. In 5G, the available transmission rate will vary significantly in short timescales due to the new millimetre-wave spectrum. This means that the transmitting endpoint must adapt its transmission rate to the available capacity. The transmission rate of the endpoint is determined by the transport protocol (i.e., TCP), and, in particular, by the congestion control algorithm. Current congestion control algorithms find the adaptation to large changes in capacity in short timescales particularly challenging, since they are designed to converge to the available capacity without excessive overshooting to avoid creating excessive congestion while doing so. Some papers [MFW19], [ZMF+16], [KTY20] have analysed the performance of different congestion control algorithms on these highly variable conditions and they empirically observe that existing transport protocols consistently fail to seize available capacity when large fluctuations occur. They also observe that the performance of different congestion control algorithms varies greatly, and that more modern congestion control algorithm such as Bottleneck Bandwidth and Round-trip propagation time (BBR) outperforms older ones such as Cubic, in some cases in several orders of magnitude [KTY20]. This in turn creates severe unfairness conditions between the different protocols being implemented in different endpoints operating in these conditions, which penalize the users of the underperforming congestion control algorithm.

Another strong trend is the use of aerial base stations or HAPS and **Non-Terrestrial Networks** (NTNs) for mobile communication. Both HAPS and NTN have been discussed for a long time. Already in 3G there were discussions for a satellite component for UMTS (see e.g. [ESES01],

[NKE+04]). However, aerial base stations have never been fully integrated in the terrestrial UMTS and LTE networks. With the development of drones and cheaper satellites, the NTN is a trend 6G must support from the beginning [D1.2]. A consequence of the new NTN and HAPS trend is probably that there will be a new metric for the Spectral and Energy Efficiency (SEE), which takes the "3D nature" of 6G into account, along with new procedures for the network-UE connection and mobility management. Moreover, considering that on different altitudes the number of neighbouring aerial base stations will vary, there is the need to define new measurements methodologies for e.g., mobility and Radio Resource Management (RRM), taking such aspect into consideration. The integration of NTN components allows 3D coverage, enabling high-altitude communication scenarios. Further on, 3D networks will open use cases in which satellites, UAV, rovers, landers, orbiters, etc. will be deployed for future application such as aerospace communications, interplanetary communications, planet exploration, Internet of Space Things (IoST), etc. Additionally, space intelligence, meaning the capability of building a networked cognitive space environment, is able to support the endeavours of space technology researchers and developers. This will imply the Internet of Remote Things (IoRT) and IoST [AK19].

Electromagnetically active surfaces (e.g., metamaterials) **and environments** that include man-made structures such as walls, roads, buildings, etc. is also gaining momentum. According to [HHA+20], smart surfaces can serve either as a transceiver or a reflector: the former case refers to an active system, where energy-intensive RF circuits and signal processing units are embedded in the surface; such system is also termed as **Large Intelligent Surface** (LIS) and represents a natural evolution of conventional massive MIMO systems, obtained by accommodating more and more software-controlled antenna elements – each element could be sub-wavelength (i.e. "small") – onto a two-dimensional surface of finite size. On the other hand, a reflector made with a smart surface acts as a passive metal mirror or "wave collector" which can be programmed to change an impinging electro-magnetic field in a customizable way; compared with its active counterpart, such reflector – also termed as **Reconfigurable Intelligent Surface** (RIS) or **Intelligent Reflecting Surface** (IRS) – is composed of low-cost passive elements that do not require dedicated power sources: their circuitry and embedded sensors can be powered with energy harvesting modules, with the potential of making them truly energy-neutral.

3.7 Traffic and services

In today's 5G networks, operators are experiencing the massive, distributed small data, e.g., in the context of IoT applications, and this is likely to continue also in 6G. Therefore, there will be the need to identify new techniques that go beyond classical big data analytics to enhance network functions' performance and provide new services, e.g., ML-based techniques. Further on, 6G is expected to support multiple functions including **Communications, Computing, Control, Localization, and Sensing (3CLS)** in a convergent way (i.e., jointly, and simultaneously). The 3CLS may improve the experience of services such as eXtended Reality (XR) and Connected Robotics and Autonomous Systems (CRAS) compared to what can be achieved with current 5G networks. As indicated in [SAP+21], for both XR and CRAS the service KPIs are a blend of traditional URLLC and eMBB as conceived for 5G, in order to simultaneously fulfil stringent and challenging requirements in terms of data rate, reliability and latency, hence reaching a balance among KPIs that might not be achievable with current 5G networks. Since the distinction between eMBB and URLLC will no longer be sustainable for applications such as XR and CRAS, still in [SAP+21] it is proposed to define a new service class called Mobile Broadband Reliable Low Latency Communication (MBRLLC) that allows 6G systems to deliver any required performance within the rate-reliability-latency space.

3.8 New spectrum

A consequence of the exponential growth in wireless data traffic is that spectrum in the THz band needs to be considered for 6G [D1.2]. Alongside communications, THz technology can bring significant advances to the areas of imaging, sensing, and localization [SSA+20]. Finally, as the antenna size shrinks quadratically as the radiation frequency increases for a given gain, there is the possibility to significantly mitigate packaging issues associated to the on-chip antennas placement, hence realizing cost-effective and compact THz transceivers [MS21]. Despite these advantages, multiple issues arise from the THz communications such as excessive signal attenuation, rapid channel fluctuation and severe propagation loss which reduce the communications' range [SAP+21]. The available transmission rate will vary significantly in short timescales and the transmitting endpoint must adapt its transmission rate to the available capacity; band splitting and bandwidth reduction are also observed due to the frequency-dependent molecular absorptions. Finally, engineering-related challenges should also be considered, e.g., high computational power for supporting the extensive bandwidth.

Another spectrum trend is the Optical Wireless Communication (OWC) technology, which exploits infrared (IR) and visible light spectrum [CHI+18]. Of these, the Visible Light Communication (VLC) may be the most promising due to the technology advancements of Light-Emitting Diodes (LEDs), the use of free and unlicensed spectrum (with extensive bandwidth, at THz-level), no emission of electromagnetic radiation and high immunity to other potential electromagnetic interference sources [SAP+21]. Moreover, VLC also ensures communication security and privacy as it outperforms in indoor due to the fact that the VLC signal cannot penetrate walls and other opaque obstructions, even though VLC has been demonstrated to be effective also in outdoor scenarios such as V2X applications and underwater communications [SAP+21] [CHI+18]. Finally, VLC can rapidly establish wireless networks and does not need expensive base stations since it uses illumination light sources which, however, might require significant backhauling capacity. VLC has also some limitations: low data rates in Non-Line of Sight (NLOS) conditions, visibility of light when illumination is not required, and it cannot perform long-distance communications. There are also optical-based network architectures such as the all-photonic RAN [ZFW+19] which is gaining attention as it brings agility and flexibility to the air interface with improved system efficiency.

4 Why do we need a new architecture?

Based on the input from [D1.2] (see Section 2.1) and the trends affecting the architecture (see Chapter 3), this Chapter identifies gaps of current architecture and lists several key components for the 6G architecture:

- Enabling AI,
- Programmability,
- Architecture for network of networks,
- New protocols for new 6G spectrum,
- Cloud softwarization and service-based architecture,
- Continuum Orchestration.

In addition to the list above, we have the sustainability and regulations as important aspects to be considered too.

4.1 Enabling AI in 6G

Thanks to the enormous development in computational resources and cloud computing, as well as due to the ever-increasing amount of available network and application data, AI can now be

applied to almost every aspect of mobile networks, enabling automated network operation and user application/service support. However, to be able to harvest from the benefits of AI, 6G systems need to be AI and computation pervasive, which calls for the 6G architecture to be data-driven. It is our vision that 6G network will leverage AI for optimising the 6G air interface (e.g., physical layer configuration; mobility and resource management; QoS assurance) but also to transform a 6G network to a powerful distributed AI platform. Hence, the AI as a Service (AIaaS) concept [D1.2] will be a key 6G enabler.

To enable AI pervasiveness to 6G networks, several aspects need to be considered during system design, taking into account a number of end users, network operator and service provider needs. Firstly, AI functionality needs to be implemented beyond the scope of today's Network Data Analytics Function (NWDAF) for the purpose of both network automation and at user application level (see section 4.1.1); such functionality needs to be utilised in a communication-efficient way (e.g., performing high-quality model training with minimal -radio and wireline- signalling focusing on high-quality data), also ensuring that both network and user data security and privacy is preserved during model training and inference. Learning architectures, which are both communication-efficient and producing comprehensible AI-based decisions, need to be flexibly instantiated across the network, accommodating heterogenous devices and data structures (see Section 4.1.2). Furthermore, distributed AI service approaches need to be developed, replacing "monolithic" centralised approaches, which are resource inefficient and thus costly (see section 4.1.3). Also, the whole cloud-edge-device continuum needs to be exploited efficiently for AI-based network orchestration and service management, calling for specific architectural enablers, i.e., interfaces and proper protocols (see section 4.1.4).

Additionally, the integration of AI in network management and operations will open the possibility to ‘anticipate’ the future network behaviour allowing to potentially decrease latency to ‘negative’ values. This paradigm is the so-called *anticipatory networking* [SF21, BCH+17]. However, the price to pay will be that these predictions will not be deterministically accurate. Thus, the architecture will have to deal with a critical trade-off between latency and reliability or resilience.

4.1.1 Network Data Analytics Function (NWDAF) and proposed enhancements towards AIaaS

The NWDAF was first introduced in 3GPP Rel-15 and further extended in Rel-16 and Rel-17 as part of Enablers for Network Automation for 5G (eNA) work items (see [23.288]). NWDAF addresses operator needs for supporting analytics based on data collected from Network Functions (NFs), Application Functions (AFs), and Operations, Administration, and Maintenance (OAM) functions.

With NWDAF, a data collection architecture is introduced in the 5GS supporting subscription to data delivery and to request a specific report of data for a particular context from any NF such as: AMF, SMF, PCF, Unified Data Management (UDM), AF (directly or via Network Exposure Function (NEF)) and OAM. Specific to the data collection from OAM, the NWDAF may collect relevant management data from the services in the OAM as configured by the public land mobile network (PLMN) operator. This includes, e.g., performance measurements from NG-RAN and 5GC (see [28.552]), as well as 5G end-to-end KPIs (i.e., KPIs with impact on the end-user (see [28.554])).

The analytics information supported by NWDAF are either statistical information of the past events, or predictive information. The consumers of the analytics information, namely 5GC NFs and OAM functions, decide how to use the analytical data provided by NWDAFs.

Rel-16 analytics include the following information about (see [23.288], clause 6.3 to 6.9):

- slice load level;

- observed service experience (derived from an individual UE, a group of UEs or any UE in an application or a set of applications);
- NF load (resource usage, NF instance load);
- network performance;
- UE mobility;
- UE communication (traffic characteristics for a given UE or a group of UEs);
- abnormal UE behaviour;
- user data congestion (data throughput per IP packet filter or application identifier); and
- QoS sustainability (QoS reporting based on thresholds for a given location area).

NWDAF architecture and functionality was further enhanced in 3GPP Rel-17 as part of eNA Phase 2 study and work items (see [23.700-91] and [23.288]). Rel-17 analytics includes the following (see [23.288], clause 6.10 to 6.14):

- dispersion (location or network slice where UEs disperse most of their data volume and sessions transactions (i.e., mobility management and session management messages));
- WLAN performance (quality and performance of WLAN connection of UE according to UE location and SSID);
- session management congestion control experience (list of UEs experiencing high, medium, or low session management congestion control);
- redundant transmission experience (packet transmission performance experienced on N3 interface with or without enabling redundant transmission);
- DN performance (UP performance for a specific edge computing application).

The Rel-17 NWDAF enhancements include: (i) support for multiple NWDAF instances; (ii) NWDAF logical decomposition into two logical functions such as Model Training logical function (MTLF) and Analytics logical function (AnLF); (iii) trained ML model sharing between multiple NWDAF instances (currently, 3GPP does not intend to standardise the sharing of models across different vendor environments as specified in [23.288] clause 6.2A.0); (iv) support for UE data as an input for analytics generation and (v) increased efficiency of data collection and analytics exposure by introducing a Data Collection Coordination Function (DCCF), Analytics Data Repository Function (ADRF), Messaging Framework Adaptor Function (MFAF), and Messaging framework (Messaging framework is outside the scope of 3GPP). These possibilities of instantiating a NWDAF in the network are illustrated in Figure 4-1.

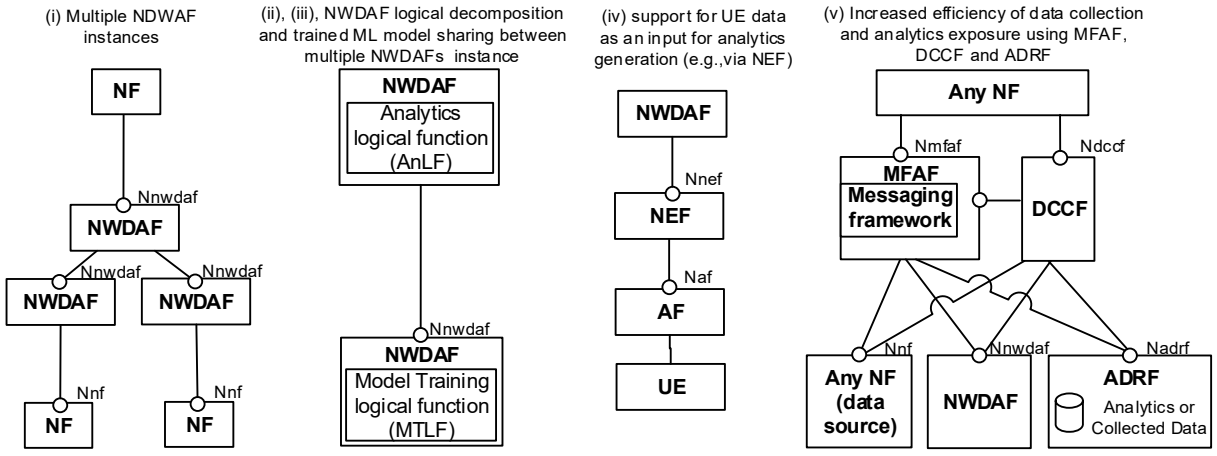

Figure 4-1 Possibilities of instantiating a NWDAF in the network.

Further information can be found in references [23.501], [23.502], [23.503], and [23.288].

As described above, the 3GPP NWDAF provides enablers for network automation. 6G will support new use cases beyond network automation requiring e.g., AI agent instantiation at the UE side. This also means that new application-specific data sources and analytics need to be supported (e.g., sensors, cameras, etc.). Furthermore, for an operator to support AIaaS as described in [D1.2], the 6G data collection and analytics exposure architecture needs to be enhanced compared to 5G in supporting any application including 3rd party applications. Indicative scenarios to be addressed by 6G networks are the following:

i) A UE (which processes generated data locally) wants to share a trained ML model with others.

ii) Same as (i) but a UE with limited computing capabilities (e.g., IoT devices, AR headsets) wants to offload some AI computation tasks to the network edge.

iii) A UE with limited computing capabilities (e.g., IoT devices) sends its raw learning data to the network edge for processing upon request issued by an AI agent instantiated at the network.

4.1.2 Motivation and gaps of federated and explainable AI

4.1.2.1 Background on collaborative and explainable AI

Collaborative AI, see Figure 4-2, may be implemented by means of the following "flavours" distinguished by the level of information shared by contributing data sources (i.e., applications running on connected devices) and the existence or absence of a global model that is updated for efficient decision making:

a) centralised learning,
b) decentralised incremental learning,
c) federated learning (FL).

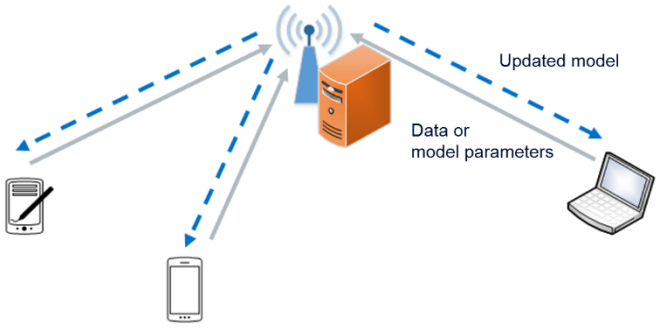

Figure 4-2 Concept of collaborative AI, where either "raw" data or ML model parameters are communicated to a central entity (e.g., edge cloud server).

Centralised learning is implemented by uploading all data from each connected device to an application server (possibly located in the cloud) to train a model that is distributable and applicable to all devices. Decentralised incremental learning consists of learning locally at the device side and sharing only the model parameters with, for instance, an application server running at the (edge) cloud which stores them locally. In this case, a library of local device models is centrally available.

The concept of FL aims to incorporate the advantages of both centralised ML (i.e., generalisation capability) and decentralised incremental ML (i.e., communication efficiency and data privacy preservation). FL refers to the case when multiple network nodes collaboratively train a global/ aggregated ML model in a distributed and iterative fashion (until convergence of the global model) by only exchanging local model parameters instead of "raw" data [Ekk20]; the global

model (e.g., built according to the FedAvg algorithm [MMR+17]) can then be shared to individual network nodes, i.e., the local model contributors or other network entities. The concept of FL was first conceived by the seminal paper by McMahan et al [MMR+17].

A key impediment to the use of AI-based systems is that they often lack transparency – AI/ML systems appear as "opaque boxes". For instance, AI/ML-based inference may be of high quality, but it lacks justification (i.e., questions are raised such as "why has this resource allocation policy been recommended by the AI agent?"). This issue is addressed by the research field eXplainable AI (XAI).

According to [AB18], there are basically four reasons for the need to explain AI systems, i.e., (i) explain to justify; (ii) explain to control; (iii) explain to improve and (iv) explain to discover. In terms of XAI methods taxonomy, explainability methods can be categorised as either: (i) complexity related, i.e., design an algorithm that is inherently and intrinsically interpretable, or use a reverse engineering-based post-hoc explanation; (ii) scope-related, i.e., understanding the entire model behaviour or understanding a single prediction or (iii) model-related methods. i.e., model-specific or model-agnostic interpretability.

4.1.2.2 Challenges to be addressed by 6G networks

Based on a first discussion in [Ekk20], in what follows, a number of identified gaps is provided in this section, which future networks need to address to support FL setups. Some of the identified gaps are the following:

1. *Communication efficiency:* AI models are typically trained iteratively, hence a large amount of data needs to be transferred on a continuous basis.
2. *System heterogeneity:* contributing devices will be quite heterogeneous regarding their communication/storage/processing capabilities, but also regarding the type and amount of data traffic they send or receive, as well as environmental/network conditions experienced by the devices themselves. For example, format of training data may vary, calling for adaptation or conversion of data transferred among the involved entities.
3. Data contributor availability and synchronisation: AI mechanisms rely on data coming from connected devices. Such mechanisms need to be able to work properly even when devices are experiencing connectivity issues (e.g., sudden drops) or temporary lack of device synchronisation.
4. Statistical heterogeneity: devices, in general, generate and collect data which fail to satisfy the statistical conditions of being independent and identically distributed (i.i.d.); of course, this may even be an asset, e.g., when it comes to designing air interface functionalities with AI/ML tools, but it may not be equally welcome across all inferencing tasks - such implementations may be further susceptible to e.g., data poisoning attacks. The distribution of data contributed by different network nodes is transparent in today's networking.
5. *Privacy preservation:* malicious applications can obtain sensitive information from an anonymised dataset (e.g., via reverse engineering), especially for devices producing "outlier" data (i.e., data that can be easily distinguished among the others). The 6G network will need to provide mechanisms to prevent the above "outlier attacks".

In addition to the above, the synergy of MEC with 6G will radically alter trade-off points, possibly making existing protocols inefficient: for instance, if FL is performed by MEC applications, executing on a MEC host, then the communication among FL-agents and a central federator becomes intra-MEC, and does not consume resources from the RAN. Note that this also allows lightweight devices to participate in FL, which extends the FL paradigm towards *objects*. However, data communication between the user equipment and the FL-agent MEC application consumes RAN resources and may become a problem. The work towards a new architecture shall investigate which functions of an FL application can be conveniently offloaded to the MEC, and which should stay on the user equipment.

With regards to the learning synchronisation issue mentioned above, problems such as asynchronous versus synchronous learning may become less of an issue in a MEC-enabled environment, where FL is done by MEC applications. In fact, asynchronous learning is preferable with the expectable high heterogeneity of devices connected to the 6G network, some of which may slow down a synchronous computation. However, when FL functions are delegated to the MEC system, computational capabilities of participants can be balanced, hence making synchronous learning effective. The same shift could take place with radio resource scheduling: depending on where functions are located, different amounts of data might need to be sent over the air and the radio resource scheduler must act accordingly.

Privacy issues may arise if MEC applications handle user data, while residing on the same MEC host, since data belonging to different users are normally physically separate, which fosters the confidentiality of this data.

Moreover, if the FL application is meant to work in a federated MEC environment in which multiple MNOs will collaborate to provide ubiquitous services (see section 3.2), network protocols designed for FL should comply with inter-MEC (east/westbound) interfaces and keep into account the federated-MEC architecture. Challenges are related to, e.g., privacy and confidentiality issues, and service migration and discovery under limited information sharing among operators.

Finally, it is foreseeable that FL will enjoy a massive boost as the number of connected objects is expected to increase in 6G. This boost will trigger a question that so far has not been thoroughly investigated, namely, how to decide who federates with whom, how, and when, i.e., *orchestration* of FL federations. For example, given N users that may be able to federate in the same FL-based application, it may not be optimal that they all do, not only for reasons of scale, but also to make the learning process more effective.

Regarding the explainability feature, there exists a fundamental trade-off between maximising inferencing accuracy and the level of explainability for a given ML model, as discussed in [D4.1].

As explained in [AB18], not every output/ decision of an AI system needs to be explained, as this highly depends on the involved application and its purposes. In any case, today's networks do not support any mechanism for requesting/issuing justifications of ML-based output. As explained in [AB18], not every output/ decision of an AI system needs to be explained, as this highly depends on the involved application and its purposes. In any case, today's networks do not support any mechanism of for requesting/issuing justifications of ML-based output.

In federated and XAI systems, challenges specific to the explainability feature need to be addressed (e.g., rule transfer, if rule based XAI is chosen, or protocol features related to request/provision of explanations). More to the point, the specific techniques used for XAI may themselves warrant a different approach with respect to the above-mentioned FL protocols: for instance, the triggers for model updates may be different, as either time-triggered or event-triggered model learning may become preferable, hence different interaction paradigms among the entities involved in the FL and XAI process (e.g. who initiates the communication with whom, and when) would need to be embedded into fundamentally new protocols.

4.1.3 Serverless AI

The combination of the serverless paradigm with AI techniques and algorithms can become a key enabler for 6G network automation. Indeed, already with 5G, the distribution of network functions and applications at the edge supports the requirements imposed by latency-sensitive services and use cases. This migration to the edge is expected to be more and more pursued with 6G, with the aim to integrate and use AI to address the challenges for extreme network capabilities and improve efficiency, service experience and network automation. Indeed, as 6G networks are expected to be increasingly complex, heterogenous and dynamic, AI can leverage on advancements of data processing technology and the availability of data. This significantly contributes to address the 6G requirements in terms of advanced radio management, intelligent traffic control, network

security and automated management and orchestration with predictive optimisation and continuous service delivery. Therefore, AI frameworks need to evolve from current highly structured, controlled, and centralised solutions towards more flexible, adaptive, and distributed architectures built of highly interconnected and cooperative AI agents and functions. AI/ML agility is required for bringing algorithms to data, instead of vice versa, and process data collected at the edge to reduce the need of high bandwidth data transfers. All this must be combined with improved data security and privacy.

Therefore, effective, and efficient solutions for distributed AI that leverage edge computing will be required for 6G. In this context, the adoption of a serverless computing paradigm can further improve AI agility and flexibility in cloud-native environments. This way AI agent and function spawning can become event-based and their execution limited (in time) to the scope of the function itself. As a consequence, this allows to use and consume computing resources at the edge in a much more efficient way. In addition, a serverless AI approach abstracts away the AI functions execution environment from the function code itself, drastically simplifying the development and deployment process. The execution of serverless AI functions can be invoked through dedicated per-function API (e.g., periodically or triggered by external events), or it can be triggered by an event source that is properly configured to activate the function in response to events such as data availability, alarms, etc. This allows to go far beyond pre-scheduled or continuous execution of AI functions at either edge or cloud locations, enabling to flexibly compose AI/ML pipelines as concatenation of event-based AI functions, thus making their management and orchestration more agile.

Therefore, the serverless AI approach can be considered a key enabler for the AIaaS paradigm, as AI functions and agents can be truly delivered and exposed as on-demand services, possibly spawned and executed as event- or API-triggered chained pipelines of serverless functions. Moreover, from an operator perspective, serverless AI would also ease onboarding and management (including verification) of algorithms and functions from third-party providers, as developers could just provide their function code independently of the execution environment. On the other hand, the serverless AI approach would help operators in implementing and deploying AI-based network management solutions and fulfil their 5G and (future) 6G network automation requirements. Indeed, it would enable the transition from a static NWDAF deployment (e.g., linked to the lifecycle of other CN CP network functions) to more agile and flexible event-based serverless execution of distributed analytics logical functions at the edge following the NWDAF functional split envisaged with 3GPP Rel 17 [23.288].

While cloud-based commercial solutions are available in the market to implement the serverless paradigm (e.g., AWS Lambda [AWS] and Google Cloud Functions [GoClo], among others), the integration of serverless technologies within the operators 5G infrastructures for in-network AI is still at its very early stages. Moreover, a wide set of open-source serverless platforms are already available and well suited for their use at the edge, like Apache OpenWhisk [ApOp] and Knative [Knat], which are relying on Kubernetes to automate deployment and execution. Such platforms can be leveraged to realize the serverless AI approach in the orchestration and operation of 5G and future 6G systems. Some research projects are also investigating this particular area, like the H2020-ICT-52 AI@Edge project [H2020], which aims at delivering a secure and reusable serverless AI platform for edge computing in beyond-5G networks. Therefore, the serverless AI approach will require a coordinated control and management of heterogeneous computing resources across the different 6G edge and core domains. In turn, this will call for an evolution in the concept of virtual infrastructure management, making it able to seamlessly manage serverless and other computing resources end-to-end. In addition, the operators' 6G network orchestration platforms will need to evolve to transparently manage the deployment, coordination, and execution of in-network AI in support of both AIaaS and network automation.

4.1.4 AI/ML applied to the network orchestration

In previous generations, regular non-AI/ML algorithmic approaches have been sufficient to address the conventional management and orchestration requirements of mobile networks. This is because these regular algorithmic approaches are typically based on a well-defined network infrastructure data (e.g., network traffic metrics or configuration parameters), which is more limited in scope and according to specific formatting rules.

However, the introduction of the extreme-edge[3] domain into the orchestration scope brings new challenges. One of them is the high heterogeneity of devices and technologies in that domain, together with the potentially huge number of devices in the extreme-edge scope. Also, the integration of those extreme-edge devices raises the option of gathering data from a new and diverse dataset very close to the end-users.

These new challenges are difficult to address using regular algorithmic approaches due the heterogeneity and the big amount of data in this new scenario. However, AI/ML techniques have demonstrated to be a valuable resource for this kind of situations [AMA19] [DKV18] [KKI+21] [RRV+19]. Also, AI/ML techniques can be used for implementing new management and orchestration strategies, e.g., implementing proactive orchestration approaches based on forecasting end-users' behaviours. In general terms, AI/ML techniques would help here as an operation automation enabler, correlating heterogenous data from different orchestration domain, e.g., core, edge (including RAN), transport network and extreme-edge domain. Also, AI/ML techniques can be useful to support the management and orchestration processes in an automated manner, thus ensuring a proper use of resources.

These AI/ML techniques [EB5G] could be applied in many ways in practice, e.g.:

i. Supervised learning algorithms could be used to trigger management and orchestration actions based on complex pattern recognition, classification tasks, time series forecasting or image processing [CCD08].
ii. Unsupervised learning algorithms could be used for data clustering, information extraction or anomaly detection [JCD+21].
iii. Reinforcement Learning (RL) algorithms could be used for implementing automated control loops based on complex data processing [JCD+21].
iv. Federated Learning (FL) techniques could be used to implement collaborative machine learning using highly distributed training data sets [YLC+19].

Based on that, possible management and orchestration actions include:

a. Proactive scaling actions based on traffic forecasts.
b. Automated NF placement actions (in the core, edge, transport, or extreme-edge domains).
c. Closed-Loop Automation.
d. Automated healing actions (in specific NFs).
e. Proactive Alerting towards the OSS/BSS teams.
f. Hidden patterns discovery, to proactively assist network planning and sizing.
g. Proactive incident analysis.
h. Support to network slices provisioning (instantiation and configuration).

To apply these AI/ML techniques in B5G/6G networks different approaches can be evaluated. On the one hand, they could be applied locally within each network domain (core, edge, transport, and extreme edge), bringing the possibility to perform management functions event at a very granular level (e.g., for single NFs or devices); on the other hand, they could be also applied in a more holistic way integrating data from the three different domains to accomplish self-network decisions and by mixing applications data and infrastructure metrics. This would be in fact the

[3] This is also known as "far edge" or "deep edge" [Lem20] [WG20].

basis for implementing the "continuum orchestration" concept, i.e., integrating data from the different orchestration domains and implementing orchestration strategies based on that.

The possible ways of implementing the continuum orchestration concept (Section 4.6) are one of the main areas of work in Hexa-X project, specifically in the context of WP6 (Intelligent orchestration and service management for future B5G/6G networks). The first deliverable of WP6, namely [D6.1], introduces and identifies the gaps, features, and enablers for B5G/6G service management and orchestration.

4.2 Programmability

While programmability has been a feature of network devices for a long time, the past decade has seen significant enhancement of programming capability for NFs spearheaded by the SDN paradigm as well as the ongoing trend towards softwarization and cloudification. On the one hand, there are now many more APIs and standardized programming interfaces towards NFs than ever before. This allows 3rd party developers to interact with the network in new ways. On the other hand, the capability to program is no longer confined to the CP software but has been introduced into (hardware) data planes as well using Smart Network Interface Cards (SmartNICs) and switches. A key candidate technology for this is P4 domain-specific language and the functional abstractions [BDG+14]. The reusability and flexibility through programmability is of particular importance at edge and extreme-edge locations where deployments have a limited footprint (i.e., subject to limited hardware types and models) and therefore need to be flexible to support a wide range of functions and use cases with diverse performance requirements. For 6G, this trend is expected to continue and even accelerate. However, many open questions remain as competing concepts exist, and actual deployments are mostly limited to trials. Therefore, key research areas in this space include:

- The right level of abstraction of infrastructure for application developers, especially when direct hardware access is currently the norm.
- Operational practicalities of rolling out functional changes of networking devices (not just configuration) automatically in the field in alignment with CI/CD pipeline methodology.
- Performance and security implications of non-integrated programmable NFs with a larger attack surface due to the exposure of more functionalities via APIs.

SDN provides programmable control framework that has enabled faster pace of transport network innovation. One limitation in initial embodiments of the SDN principle was programmable control of fixed protocols. Recent advancements in data plane programmability such as the invention of P4 programming language has enabled the possibility to introduce new data plane protocols by just compiling and running a program. This enables to control unlimited number of data plane protocols rather than being limited to only standardized ones. SDN principles have recently been considered in the scope of mobile networks as well. A limiting factor in achieving the full potential of network adaptability in mobile networks is that introducing features having impact on air interface protocols (and, in general on UE behaviour) is not possible without going through standardization procedures since changes should be applied to both UE and gNB. To unlock this potential, 6G calls for a programmability framework that gives the possibility not only to program specific features in a single network entity, but also to program more innovative features having impact on air interface protocols. An example could be having a tailored control-plane procedure for industry sensors/IoT devices connected to a dedicated network to optimize the performance in a factory automation scenario. This motivates for an abstraction level that enables UE programmability with a right balance between the scope and pragmatism to enable flexibility and adaptability at the edge.

4.3 Network of networks

One of the goals with 5G was to include new use cases such as massive MTC and critical services with URLLC [MII17-D24]. As mentioned in Section 3.5, there are many trends (e.g., sub-THz, private or non-public networks (NPN), autonomous networks) indicating that this will be even more important to 6G. The exponential advancement of technology with the various infrastructures and the diversified applications in different sectors (AR, VR, digital classes, remote health services, digital twins, holographic projection etc.), will stress the networks and push for a new 6G architecture, namely the envisaged "network of networks" [D1.1] as described in [D1.1].

"Network of networks" is a new architecture to support limitless connectivity and global coverage, to enable diverse services by means of big data and AI-powered backends, as well as embedded intelligence. This approach will enable the integration of a) Non Terrestrial Networks (NTN), b) Flexible multi-connectivity (Flexible MC) and/or combined cell/multi-point transmission, c) Sub Terahertz nodes (Sub-THz) and Visible Light Communication (VLC), d) L1/2-mobility, Distributed Multiple-Input and Multiple-Output (D-MIMO) and Multi-Transmit Points (Multi-TRP), e) Device to Device communications (D2D) and mesh topologies, and f) Campus/private networks, together with architectural/technology enablers to manage the locally created ad hoc networks, in coordination with the infrastructure, as well as to distribute their functionalities between them and at the edge.

Flexible topologies will be the key for addressing challenges like global service coverage, lower latencies, higher reliability for certain relevant cases, security, and decentralization, and handling extreme experiences. Flexible topologies will be realized by studying and exploiting the dynamic integration of the technologies aforementioned above, e.g., the (a) advanced networking technologies (mesh, nano technology networks non-terrestrial networks, D2D, etc.), (b) (sub) Terahertz by using ultra-high frequencies in base stations, (c) new computing paradigms and more. Trustworthiness of the system, dependability, sustainability, limitless connectivity, and inclusiveness are the main 6G values, having the target achievements of higher data rates, lower EMF emissions, lower latency, and higher capacity. Furthermore, there is probably a need for a functionality allowing for automated addition, configuration and optimization of new network entities when introduced within the network, significantly reducing the operator's effort for network planning [6GSamS]. Other aspects to consider are related to enhancing mobility support for mobile entities belonging to transportation systems but that can be part of the cellular network, considering the speed of these systems and to improve the service continuity for user devices served by the network entities, which in turn could be moving and are connected to the cellular network through wireless connections. Such aspects have already been addressed in 5G, but it is foreseen that they will be of higher importance in future 6G networks, hence requiring the above-mentioned enhancements/improvements.

4.4 Cloud and Service-Based Architecture

As shown in Sections 3.1 and 3.3 with numerous examples, cloud-based solutions are a strong trend as of today. 5G CN now supports cloud-native implementation of the service-based architecture (SBA). Further on, 5G employs concepts such as separation of UP and CP functions, network slicing, convergence of fixed and mobile communication (and non-3GPP access), local breakout mechanisms, support for a wide range of frequencies, etc. However, some areas of improvements identified here with respect to the current 5G networks, for example the functional allocation and procedures may prevent full integration of cloud-native network functions across all domains and layers. The current 5G architecture applies service-based approach in core network [23.501], [23.502] and defines network functions applying service-based principles, but here the scope is only for the CN omitting RAN and the management system. Service based approach has also been adopted in the management system (SBMA, Service Based Management

Architecture), with different management services federated together following service-consumer producer patterns. However, SBMA and the SBA as applied in the CN differ in the way how SBA is applied: CN builds service discovery around Network Repository Function (NRF) where as SBMA doesn't have such a service explicitly but multiple options instead. Further, SBMA [28.533] doesn't define NFs but APIs only counter to CN CP. ETSI NFV is also moving from an interface-centric solution design towards a service-based approach from release 4 onwards [ENFV04]. RAN functionality is still defined by using classical means based on the concept of network entity and peer-to-peer interfaces rather than network functions and service interfaces as in CN. As the cloudification continues within all subsystems (i.e., RAN, CN, and management) the overall architecture should be revisited to ensure architectural consistency, streamlined introduction of new features and simplicity of customization. Planned improvements include better cross-plane and cross-domain interactions, particularly for data collection for analytics and AI/ML needs. The 6G architecture must be more flexible to accommodate new types of end user devices and access network topologies which calls for dynamic functionality upgrades and function distribution to match changing deployment needs.

4.5 Softwarization

The concept for network of networks as foreseen for 6G implies rethinking convergence of architectures to be more cooperation between network segments and technologies. A cross-functional obvious convergence that is already engaged is the Network-Cloud convergence. This convergence aims at softwarizing E2E functionalities according to IT and service-based principles for greater flexibility and dynamicity in terms of design, deployment, configuration, and reconfiguration, etc. These principles bring software-based approaches to the centre of the network function design but limited to the CP only. However, applying these principles to management and user planes requires consistency of the design across the whole architecture, e.g., how APIs are designed, granularity of functionality, how functions are managed, etc... The current trend is to implement larger 5G functions and services as micro-services and multi-agent systems. Need for new APIs is an obvious consequence. A heterogeneous, fully-softwarized architecture, that is based on a multi-agent approach, will need to be designed to run on physical architecture seamlessly, efficiently, and effectively. The architecture must be able to support multiple technologies and network topologies. Effectiveness means to instantiate cloud native functionality across different domains even though the software components would be otherwise self-managed, and self-optimized.

4.6 Continuum Orchestration

Another reason why a new architecture is necessary is the realisation of the *Continuum Orchestration* concept, which implies the evolution of regular management and orchestration techniques towards the continuum consisting of the joint combination of different orchestration domains: core network, transport network, edge, extreme-edge, and other networks that can be external to the MNO (e.g., fixed access networks, private networks or hyperscaler networks). In pre-5G generations, management and orchestration resources were primarily focused on the core network. The 5th network mobile generation already makes possible the emergence of new architectures enabling the joint management and orchestration services and resources deployed on both: core and edge. However, the concept of "continuum orchestration" for B5G/6G networks takes this a step further by also including other resources as mentioned before (see Figure 4-4).

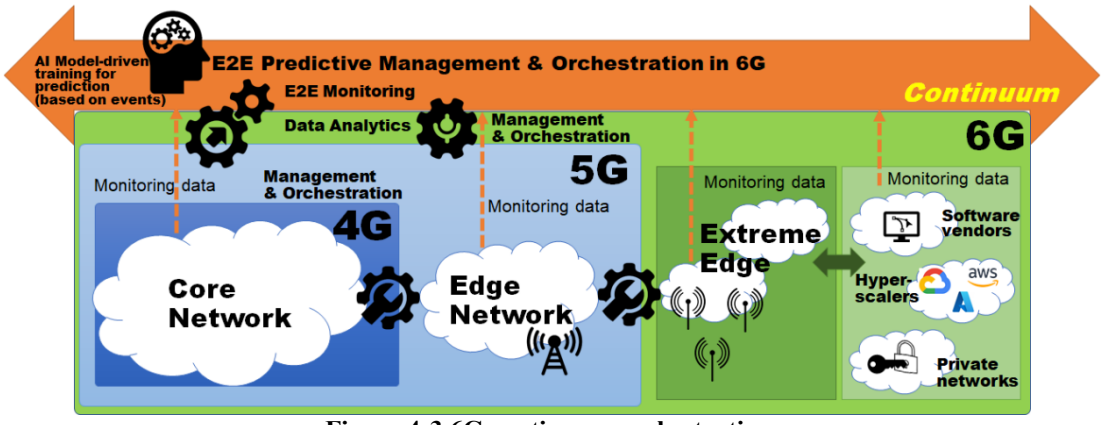

Figure 4-3 6G continuum orchestration.

This is aligned with one of the main aspects considered in the scope of the H2020 ICT-52 call, which is the long-term transformation of networks into a distributed smart connectivity platform with high integration with (edge) computing and storage resources. The Hexa-X approach to this regarding network management and orchestration is to provide a framework for dynamically supporting reliability and resilience in changing requirements providing the so-called "continuum" device-edge-cloud management to address mobility and resource utilisation. This work is mainly addressed in WP6, where one of the main objectives is to demonstrate algorithms for data-driven device-edge-cloud continuum management, and provide implementations of continuum management of device, edge, RAN, and cloud as one of the main measurable results.

It should be noted that the extreme edge domain is quite different from the core and edge domains that are within the management scope in 5G networks. These edge and core resources are typically deployed in strictly controlled data centres under a close supervision of the network operator, i.e., the network operator chooses the hardware and software resources to be used to host and run the network services and deploys them in controlled environments with strict access and monitoring rules.

However, the nature of "the extreme edge" is totally different: there can be a plethora of devices based on a wide variety of technologies, both hardware and software. They could be personal devices (smartphones, laptops…), and a huge variety of IoT devices (wearables, sensor networks, connected cars, industrial devices, connected home appliances, etc.). Although the diversity of the extreme-edge devices is already high nowadays, the forecast is that the number of devices continue increasing in the coming years, becoming very large when B5G/6G technologies are available [RAD21].

But it is not just a matter of diversity. The environment is quite different also. It cannot be assumed those end-devices will be in a controlled environment such as the current 5G core and edge resources. Also, it can be assumed they could be in a sort of random and error-prone environment where they could be unexpectedly switched on/off, reconfigured, damaged, or whatever (they are end user-owned devices)[4]. It can be stated that, these devices would behave in a highly asynchronous way, which would have to be properly handled by an evolved infrastructure management function.

So, this extreme-edge domain has certain features that make it quite challenging to include it in the typical network management and orchestration processes, namely: (i) high heterogeneity of devices and supporting technologies, (ii) asynchronous, random, and error-prone environment,

[4] Perhaps in certain cases a more controlled environment might be available (e.g., in factories, airports, vehicle fleets...) but in any case, it should be assumed that the control level on the extreme edge devices will typically be outside the network operator scope, contrary to what is currently the case.

(iii) massive in scale. So, certainly, it is expected this new “continuum” management and orchestration concept including the extreme-edge devices will impact the upcoming 6G network architectures.

One of the impacts would be related to the infrastructure management function, as it was already mentioned. Another one would be on the usage of AI/ML techniques, that can be considered as a key enabler to deal with the heterogeneity of technologies and devices, as well as with the high number of connected devices. Indeed, AI/ML algorithms can be of valuable help in dealing with the great heterogeneity of data and the formats in which they may be presented, helping to deal with the envisioned complexity and serving as computational support for a data-driven architecture.

An additional complexity level would be the evolution of paradigms that are in use in 5G networks. One of them is the network slicing paradigm [GSMA18], but with the possibility of extending the network slices to also cover the extreme-edge resources. Also, using AI/ML approaches to correlate data plane and infrastructure metrics would be used to improve slicing elasticity algorithms based on zero touch automation techniques, e.g., applying predictive approaches to orchestrate the network resources based on end-users’ regular behaviours or preferences.

Another one is to continue providing support to other actors, and not only the MNO. As 5G networks are starting to be, B5G/6G networks would be enabled to inter-operate with different stakeholders (e.g., vertical operators, hyperscalers, neutral host providers or managed service providers among others). This of course adds a new level of complexity that affects also the future 6G network architectures.

Considering all above, there is a need for a new architecture to support all connected resources across different management domains, including technology, network, and administrative domains.

4.7 Sustainability and regulations

Sustainability has become a major requirement to cope with environmental, economic, and social challenges in the 21st century. Achieving sustainability is described being essential to continue making human life on Earth possible for future generations. The development of 6G architecture mainly needs to consider the environmental dimension (reduced usage of natural resources, e.g., by energy efficiency), while new service enabled by a 6G architecture can help with addressing economic and social challenges. In 2015, the United Nations General Assembly set up a collection of 17 interlinked global goals that are referred to as Sustainable Development Goals (SDGs). Most relevant to 6G architecture development is the goal 12 on "responsible consumption and production", addressing the following aspects: good use of resources, improving energy efficiency, sustainable infrastructure, and providing access to basic services, green and decent jobs and ensuring a better quality of life for all [SUS].

On the regulations side, the Radio Equipment Directive (RED) [EU14] is a key regulation tool by the European Commission related to access of any type of radio equipment to the European Market. As stated by the European Commission, the Radio Equipment Directive: "*ensures a single market for radio equipment by setting essential requirements for safety and health, electromagnetic compatibility, and the efficient use of the radio spectrum. It also provides the basis for further regulation governing some additional aspects"* [RED14].

The wireless industry is used to working with the regulation framework under RED. For the certification of radio equipment, however, the industry is supported by the officially recognized European Standards Organisations (ESOs), i.e., ETSI, CEN and CENELEC. The ESOs are developing so-called “Harmonised Standards (HS)” (or Harmonised European Norms (HEN)), which translate the rather generic text in RED into very specific technical requirements and related conformance tests.

As of today, the key articles of RED containing essential requirements to be met by industry are Article 3.1 on safety and health and electromagnetic compatibility as well as Article 3.2 on the efficient use of radio spectrum. These articles are currently in force and must be met by radio equipment in Europe to achieve market access.

Still, at the time when RED was introduced, there was an additional article 3.3 introduced, including a number sub-articles 3.3(a), …, 3.3(i), which have not yet been invoked. Currently, the European Commission is working towards an activation of the following sub-articles of Article 3.3, related to the following new requirements:

- Protection of the network and its resources (Article 3.3(d)),
- Protection of personal data and privacy of the user and of the subscriber (Article 3.3(e)),
- Protection from fraud (Article 3.3(f)),
- Ensure that combination of HW & SW maintains compliance to RED (Article 3.3(i)).

The new requirements of RED are expected mandatory to be followed by the time of 6G developments will be worked on in key Standards Developing Organisations (SDOs) such as 3GPP.

Current network designs and architecture approaches do likely not meet the new requirements of RED because the RED requirements had not been defined when the systems were designed. It is thus important to consider architectural changes across the whole chain (network side, User Equipment, etc.) for future equipment, in particular 6G equipment, to meet the new upcoming RED requirements.

Related to above requirements, the protection of personal data (in the sense of information which are related to an identified or identifiable natural person, such as credit card number, number plate, appearance, etc.) and privacy the EU General Data Protection Regulation (GDPR) was put into effect in 2018. Personal data is not only data that directly can be used to identify an individual, e.g., name and email address, but also covers data than can indirectly, e.g., by combining different data, used to identify a person. Even pseudonymised data can fall under this definition, if someone can still be identified when analysing the data, therefore imposing stringent requirements on pseudonymisation methods and use of pseudonymised data [GDPR-PD].

According to the GDPR, the protection of personal data covers all stages from data collection, recording, organizing, structuring, storing, using, to finally erasing the data.

The following principles need to be fulfilled by all entities processing data [GDPR]:

- Lawfulness, fairness, and transparency,
- Purpose limitation,
- Data minimization
- Accuracy,
- Storage limitation,
- Integrity and confidentiality,
- Accountability.

Taking the above into account, the 6G architecture needs to be designed to handle the data protection security principles, even in deployments where confidentiality is not imperative, as well as the sustainability goals.

5 Architectural Transformation

In previous Chapter 3 we described the trends that may affect the architecture in 6G and in Chapter 4 we identified possible gaps of current architecture and key issues to consider for the 6G architecture design. In this Chapter we try to point at the direction the 6G architecture should move towards, what we call the architecture transformation. Since we are still in the early phase

of the Hexa-X project (and 6G research in general), we focus on the components and enablers (i.e., solutions) necessary for a 6G architecture. First, we give an overview of the proposed 6G architecture in Section 5.1. We define a so called 6G functional layered architecture and the 6G architectural principles. Thereafter we define the necessary 6G enablers for Intelligent, Flexible and Efficient networks in Sections 5.2, 5.3 and 5.4, respectively. Finally, we discuss the trustworthiness and sustainability in Section 5.5.

5.1 Overview of the Hexa-x 6G architecture domains and principles

In WP5 we have defined different domains of the architectures. The different domains of the architecture are used to group where different architecture functionalities will be placed and understand how they are related to each other. In addition to this, the domains also reflect the work in the different tasks. The domains are as follows:

- Network of networks deployment: represents the actual deployment used to serve the users in a specific network deployment.
- Network Services: common cloud platform than runs the network functions.
- Network functions: RAN and CN expose. These services are for example RAN scheduling mobility, session handling etc.
- Interfaces and transport: signalling between the different network nodes and NF, including UE

In addition to this, we have the management interacting with all the different domains.

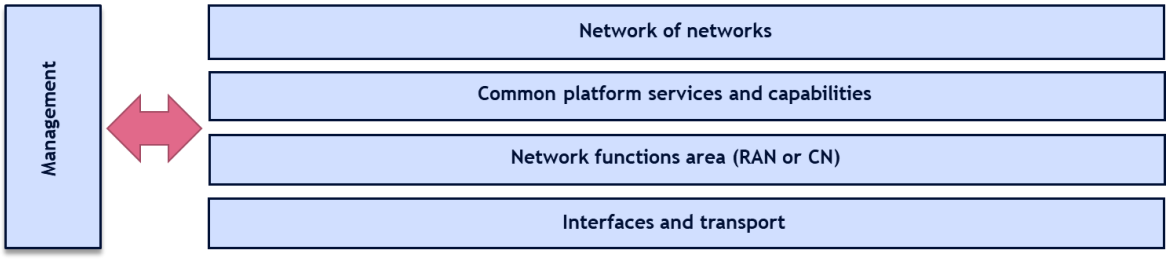

Figure 5-1 Architecture domains in WP5, used to place different enablers and understand how they are related to each other.

In addition to the layered architecture, we have defined eight different Architectural principles we believe the 6G architecture should fulfil, see Figure 5-2. They are sorted in colour for the different tasks in WP5: blue principles belong to Tasks 5.2, green to Tasks 5.3 and finally, orange to Tasks 5.4. The principles are explained below:

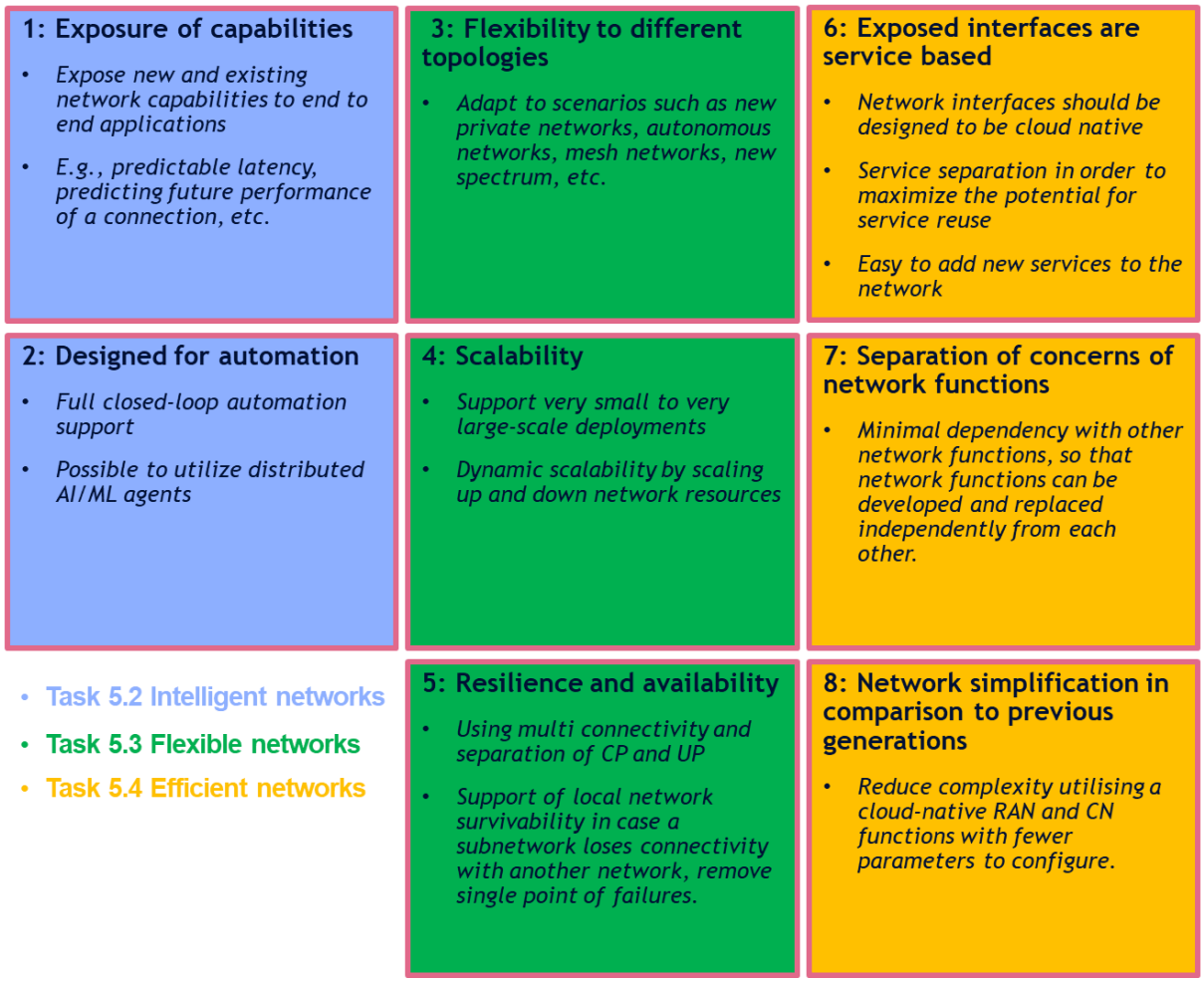

Figure 5-2 6G architecture principles, sorted in colour for the different tasks in WP5.

Principle 01: Exposure of capabilities

The architecture solution shall expose new and existing network capabilities to end-to-end applications. This means that the applications can be provided enhanced capabilities for various network features. This can for example be used for achieving low and predictable latency, predicting the future performance of a connection/application, predictive orchestration, etc.

Principle 02: Designed for (closed-loop) automation

The architecture should support full automation of network and service management operations, utilizing distributed AI/ML agents to manage and optimize the system without human interaction. Key features include observability and analytics as well as intent-based management.

Principle 03: Flexibility to different topologies

This principle is the ability of the network to adapt to various scenarios such as novel subnetworks such as non-public networks, autonomous networks, mesh networks, new spectrum, etc., without loss of performance and easy deployment. Addition of service capabilities and new service endpoints requires can be done in run-time without changes to existing end-to-end services.

Principle 04: Scalability

The system architecture needs to be scalable. The architecture shall support very small to very large-scale deployments, by scaling up and down network resources based on needs, e.g., varying traffic, utilizing underlying shared cloud platform.

Principle 05: Resilience and availability

The architecture shall be resilient in terms of service and infrastructure provisioning using multi-connectivity. This means that the architecture shall support separation of CP and UP and multi-connectivity as a method to provide service availability[5]. Further on, the architecture shall support of subnetwork resilience e.g., if a subnetwork loses connectivity it should connect with another subnetwork to remove single point of failures.

Principle 06: Exposed interfaces are service based

Network interfaces should be designed to be used in a cloud environment (i.e., cloud native), utilizing state-of-the-art cloud platforms and IT tools in a coherent and consistent manner (see [28.533] and [EGZ19]). Care should be taken to design proper service separation to maximize the potential for service reuse, and ease of adding new services to the network (plug-and-play).

Principle 07: Separation of concerns of network functions

The network functions have bounded context (separation of concerns) and all dependencies among services are through their APIs with minimal dependency with other network functions, so that network functions can be developed and replaced independently from each other.

Principle 08: Network simplification in comparison to previous generations

Streamline the network architecture to reduce complexity utilising cloud-native RAN and CN functions with fewer (well-motivated) parameters to configure and fewer external interfaces, to maximize innovation and reduce time to market.

The principles can also be placed in the different domains, see Figure 5-3 to show in which WP5 Tasks each principle belongs.

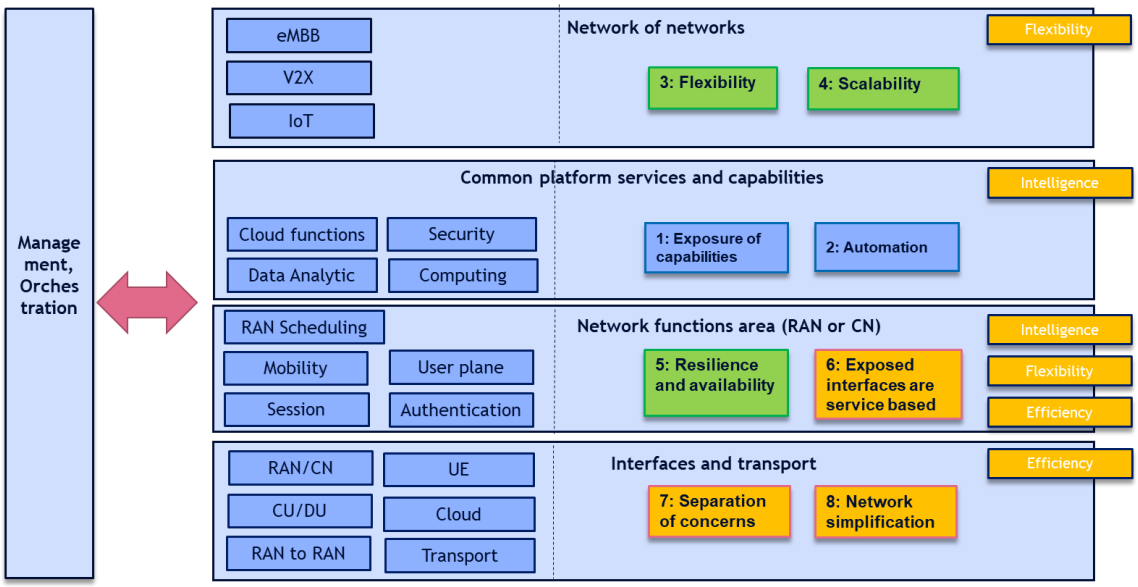

Figure 5-3 The architectural principles placed in the different architecture domains.

An alternative view to a potential 6G system architecture has been proposed by the Oulu 6G Flagship [O6GF], one of the early actors in 6G research. Although that work started somewhat prior to Hexa-X, many of the architectural principles listed in Section 5.1 are common in both. However, the starting point in [O6GF] work has been to identify the key set of high-level services and build upon them a service-centric functional 6G system model that can realize the envisioned new use-cases. It provides only the functional description of the services and their components,

[5] Availability is indicated by the percentage of time during which all required QoS parameters are satisfied.

leaving more freedom for the design of a logical network architecture at a later phase. See Annex A.3 for more information.

The following sections will briefly describe the architectural enablers for the different tasks. A more detailed description of the architectural enablers and how they intend to fulfil the 6G architectural principles is found in Chapter 6 (Intelligent networks), Chapter 7 (Flexible networks) and Chapter 8 (Efficient networks).

5.2 Research directions for Intelligent networks

5.2.1 Better support of AI

The scope of this subsection is to provide architectural recommendations on how a 6G system should efficiently support AI functionality needed for automated network operation. In other words, considering the AIaaS concept elaborated in [D1.2], the network architecture should be redesigned in order to support different software implementations of AI functionality, multiple AI agent setups and different learning architectures, including Federated Learning (FL). Beyond adding and embedding AI functionality into the architecture, the underlying NFs need to be adaptable to new conditions and environments that were not originally foreseen. This calls for network programmability, dynamic function placement and accurate means for data analytics for the basis of AI-based decision making and subsequent actions. These enablers are further elaborated in Chapter 6.

The architectural discussion about better supporting AI should also deal with the mutual dependency between intelligence and network operations, considering the satisfaction of KPIs. Since AI will be implemented in software environments, it will be more prone to failures, possibly contributing to reduce network reliability and availability by replication for example. Replication unfortunately leads to a synchronization problem. However, this drawback could be compensated by probabilistically predicting failure events which might result to the benefit of the so called ‘negative latency’ recovery, towards sub-millisecond delays.

Next, AI will have to also guarantee efficient, optimal, and continuous end-to-end orchestration. This might require a hierarchical approach to the design of the communication network architecture (in this sense, the nature of agents guarantees higher flexibility) [ABG+21]. The orchestration needs span over different domains (see more in Subsections 4.6 and 6.2) but still have a consistent view of their own scope. In such context, the coordination of a complex and distributed system will need the design of effective logic architectural interfaces to manage and synchronize operations and communications. To find more about this see Sections 6.2 (Network Automation) and 6.4 (AI-driven Orchestration). The communication aspects are divided into AI for communication (Sections from 6.3.1 to 6.3.3) and communication for AI and NF across domains (Section 6.6, Network Service Meshes).

Future 6G networks will make extensive use of AI algorithms, for various purposes and applications, some of which may not necessarily be known or foreseeable at the time of writing. We therefore should endow the network with flexible mechanisms to support federated/distributed learning of *collaborative* AI models (see Section 6.3). Federated/distributed learning is required to preserve data privacy and increase scalability. For example, in the case of FL, standardised management functions are required to discover the FL servers, the set of available FL applications, join/leave a FL process, exchange models (or parts thereof, especially in the case of rule-based AI models), request updates, etc. These management functions should be flexible enough to allow very different learning paradigms to be constructed (e.g., synchronous vs. asynchronous model updates, “black-box” as well as explainable models, etc.).

It is foreseeable that the devices that will need to participate in FL will be heterogeneous: from standard, high-end (processing-wise) handheld equipment to *things*, with very limited computation capabilities. This calls for to *proxy* constrained devices in their FL process,

relocating computation *inside* the network. This is enabled by the ever-increasing coupling of computation and communication in networks (e.g., MEC-enabled networks). FL mechanisms should thus be able to support both end-user-located FL agents and network-located FL agents seamlessly and simultaneously, possibly moving computing functions between network and user device dynamically. Refer to Subsection 6.3.3 to find more about the protocol implications of FL.

In a FL context, the role of FL server is crucial. Both centralised and distributed FL servers should be supported, possibly across different network domains (e.g., in a Federated-MEC environment) and with different architectures (hierarchical vs. peer-to-peer, etc.).

5.2.2 Management & Orchestration

As introduced in Section 4.7 (Continuum Orchestration) the introduction of the continuum core-edge-device management and orchestration paradigm is one of the drivers for the B5G/6G architectural transformation. Indeed, the need to include subnetworks of multiple devices in the extreme edge domain, and the associated asynchronous/transient behaviour poses new challenges that will probably have impact on the architecture design.

One of the main reasons for this impact is the need to manage the extreme edge infrastructure, which, as explained in Section 4.7, would require the inclusion of AI/ML techniques applied to the management and orchestration processes (Section 4.1.4) in order to address the high level of complexity associated to this extreme edge infrastructure. Also, another relevant factor with potential impact on the architecture transformation is the envisaged interaction between the MNO and other new stakeholders such as vertical industries, public and private clouds, hyperscalers, etc. (see [D6.1]). This is something that has been already started in previous generations, but it is expected to continue evolving in the coming years. This will include both: automation of orchestration across stakeholders and automating hosting of NFs on external cloud platforms.

The asynchronous management of the infrastructure may require a revision of the Infrastructure Manager concept already in use in previous generations. The new Infrastructure Manager should be sensitive to asynchronous state changes in extreme edge elements to dynamically add or remove them from the pool of available resources. Also, the associated orchestration algorithms should provide the necessary redundancy mechanisms to avoid service losses when extreme-edge devices unexpectedly become unavailable. Conceptually, such asynchronous infrastructure management mechanisms could be seen as analogous to the SON mechanisms for eNB self-configuration [36.902], but with application to the extreme edge resources.

The inclusion of AI/ML techniques in the management and orchestration context is already considered for 5G to certain extend [28.809], but for 6G we expect potentially big amount of heterogeneous data sources and the increased complexity that should be managed to implement the “continuum” management and orchestration concept that we envisage here, i.e., considering data sources not only from the core or the edge networks (as it happens in the current 5G technology), but also from the extreme edge end-devices that, as stated in Section 4.6, are deployed in a sort of “uncontrolled” environment (the end-users scope), so having a highly volatile unpredictable behaviour and being based on a variety of different technologies which leads to a myriad source of data to monitor and orchestrate.

Furthermore, a very high degree of automation is needed to manage far-edge deployments in a self-reliant fashion, as these may not be accessible by maintenance teams for prolonged periods of time or could suffer a loss of connection towards the central orchestration (principle 02). For this reason, these deployments should also feature a standardized form factor and the ability to easily swap out hardware modules to adapt deployment functionality or replace/upgrade existing hardware.

Finally, regarding the interaction with new stakeholders, the architecture should be enabled with the necessary interfaces and security mechanisms to make this feasible. Multi-domain orchestration should be enabled considering the resources from those different stakeholders making it possible the dynamic management of network slices with the aggregation of resources

from the different involved actors. All these aspects will be addressed in WP6, that together with WP1 will try to clarify how this architectural transformation would be enabled during the lifecycle of the Hexa-X project. In relation to this, automation of service provisioning across stakeholders (e.g., for the lifecycle management of network slices involving resources of different stakeholders) can also be stressed

5.3 Research directions for Flexible networks

The 6G architecture for network of networks shall enable efficient integration of every type of (sub)network, see Figure 5-4

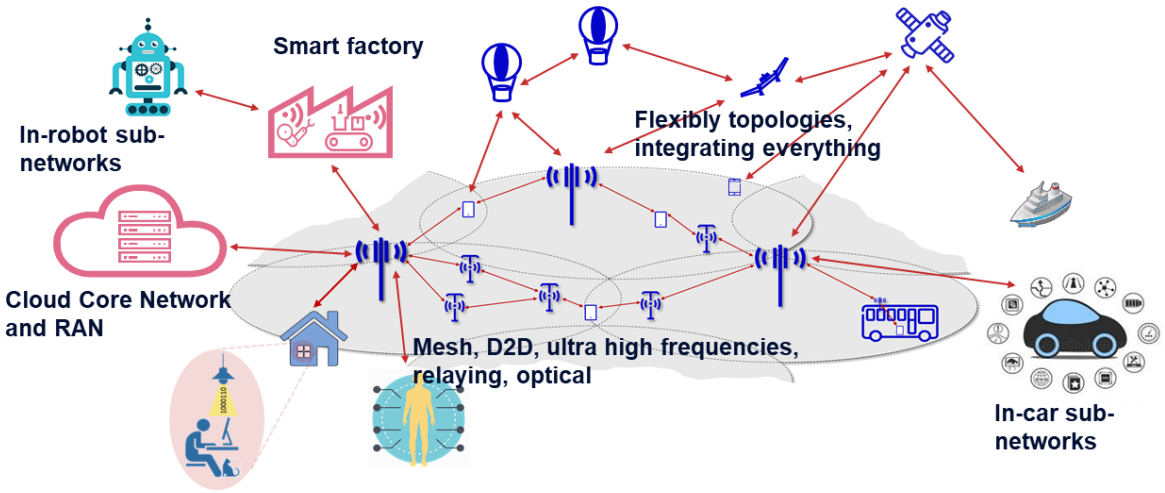

Figure 5-4 The 6G architecture of network of networks should enable efficient integration of different types of (sub)networks.

The current vision of 6G has been enhancing the idea of 'ecosystem' of networks (or network of networks). This brings 6G to embody the original concept of the 'Web of Everything Everywhere' [PSB15]. 6G will finally integrate satellite, aerial and terrestrial networks, in a unique dynamic-adaptive network infrastructure [SBC+20]. 6G will of course not start from scratch but will build upon 4G and 5G legacy of support different types of networks (such as URLLC and IoT networks). However, 6G need to support the flexible deployment beyond what 5G can do. The 6G architecture should be able to support new use cases and deployments that may arise in a more flexible way and avoiding lengthy standardization processes.

A network of networks is defined as a network that can both incorporate different (sub)network solutions as well as a network that easily (flexibly) can adapt to new topologies. For example, different network solutions can be a network using sub-terahertz spectrum, different variants of mesh networks, NTN networks using satellites, HAPS or drones, networks using cell-free L1/2 mobility (e.g., D-MIMO) and local device networks. The network functionality and architecture must then be flexible enough so that it can adapt to these different topologies (principle 03).

The reason why 6G architecture should support network of networks is to meet the requirements of both extreme performance and global service coverage, well beyond what 5G is capable of. Another key enabler for the flexibility paradigm of the 6G era will be the enhancement of scalability using radio access virtualisation methods (principle 04). To enable a new smart connectivity platform across verticals and the associated value chains, the architecture must be able to integrate mission-critical networks with dependability, coverage, and reliability beyond 5G's URLLC. Integrating and leveraging such architectural trends from the start of the 6G concept development is expected to significantly increase flexibility and cost efficiency, while also decreasing the overall complexity for device, RAN, CN, service layer, data analytics and management services. In general, the degree of integration of any sub-network may vary, depending on use-case and cost. We must therefore consider a suitable level of how tight it should

be integrated, based on relevant use cases, and cost so that the network becomes scalable (principle 04). For example, the integration of NTN networks needs to focus the solution on the most important use case and considering the cost aspects.

Another important principle is reliability, availability, and resilience (principle 05). Network of Networks technologies (flexible topologies, the introduction and "connection" of enhanced intelligence, and cost efficiency) are key for enhancing these aspects. Flexible topologies will be alternatives that may increase the availability and reliability of infrastructures. New intelligence will be needed to make decisions "on the fly". Cost efficiency (e.g., limited resource and energy consumption) should underpin all operations. Another argument for the importance of enhanced reliability is that MBB is becoming more and more critical to the society as well as the need to support new demanding vertical services. Therefore, the 6G architecture must enhance its support for reliability, availability, and resilience beyond 5G, in terms of both service and infrastructure provisioning. Related to reliability is digital inclusion and global service coverage. The 6G architecture shall enable coverage of remote places, e.g., in rural areas, transport (e.g., ships and airplanes) over oceans or vast land masses.

Finally, it is important to highlight that NTN will significantly affect the design of 6G Layer 3/Layer 4 new network and transport protocols, which can be capable to exploit the three-dimensional and heterogeneous characteristics of cross-technology communications efficiently and effectively. These new network and transport protocols will also have to ensure the reliability via network erasure correction and/or retransmissions. On the other hand, flexibility will be ensured by the new paradigm of programmable protocol stack (PPS). In fact, the softwarization of the protocol stack will guarantee efficient performance, real-time performance analysis, greater scalability, dynamic protocol upgrades, and protocol stack flexibility according to the various heterogeneous technologies used among communication parties.

5.4 Research directions for Efficient networks

As discussed in Section 4.4, we believe that the 5G functional allocation and procedures prevent full integration of cloud-native network functions across all layers. The 6G architecture shall be able to fully utilize the cloud platform. With a cloud-native network, it should be possible to streamline the RAN and CN network architecture, i.e., to reduce some of the complexity. In today's 5G networks there is too much signalling between different nodes. Typically, the procedures traverse the UE, gNB, the AMF, SMF and the UPF. To take full advantage of the cloud-native approach, enhancements, and extensions to the 3GPP SBA and SBMA [23.501, 28.533] will be investigated. Building on SBA and SMBA will ensure architectural and operational consistency across the RAN-CN and network management while (i) minimizing potential backwards compatibility issues and (ii) facilitating 3rd party service integration. This is expected to contribute to TCO aspects significantly. With a cloud-native implementation of the CN and RAN, it should be possible to dispense with some of this signalling, hence optimizing the overall signalling speed (principle 08). The network functions' responsibility and functionality must be separated more clearly than in today's network. Furthermore, we should avoid duplicating functionalities (principle 07).

As mentioned in Section 3.3 dealing with trends in cloud technology, the cloud environment is expected to evolve towards a heterogeneous and edge-integrated architecture. Such an architecture emphasizes the open nature of cloud platforms and calls for open interfaces between the cloud and its applications, e.g., cloud-native network functions. The cloud infrastructure is distributed across multiple tiers, crossing administrative boundaries, using possibly different virtualization technologies, orchestrators, and cloud federation. However, involving multiple entities - even within a single domain - increases integration complexity, incurs the risk of overbooking available resources, and opens the possibility of concurrent contradicting requests from different competing services that easily compromise the overall system performance. Therefore, the functional modularization and proliferation of the number of standardized open

interfaces needs to be optimized to match the needs of the envisioned use cases, the number of involved actors, and the resources that are offered to the services. Moreover, the exposed interfaces and used protocols should follow a coherent and consistent design methodology - especially for multi-domain boundaries. The interfaces should be configurable to be suitable for micro, macro, and edge clouds as well as to support different redundancy modes and latency needs.

A cloud-native computing infrastructure enables elastic and scalable computing that can accommodate sporadic workloads. In the anticipated edge-centric cloud architecture, the workload elasticity needs to also be extended horizontally across multiple edge clouds and vertically to their upper-tier clouds. Predictive workload provisioning that takes processing and network load fluctuations into account - including provisioning delays - needs to be applied to ensure function colocation to jointly optimize processing and connectivity needs to meet latency requirements.

Cloud-native network functions should be designed to avoid overlapping and redundant functionality whenever possible (principles 7 and 8). Therefore, the underlying cloud platform and infrastructure should offer standardized telco grade services, e.g., message delivery, load balancing, high availability, trusted computing, and HW acceleration - among others. These infrastructure and platform services need to be available in a consistent fashion across multi-cloud deployment and offer predictable performance characteristics. This is needed to optimize network performance and code portability.

The architectural solutions and enablers must consider the requirements of localization and sensing. Localization already exists in today's 5G networks, where the UE, gNB (RAN) and AMF together with the Location Mobility Function (LMF) are involved during a positioning procedure. The UE and the gNB transmit the Positioning Reference Signal (PRS) and do positioning measurements on the PRS depending on the positioning method being employed. The LMF and AMF nodes play a co-ordination role during UE position estimation procedure. For 6G it is expected that the use cases will put higher requirements on the 6G solution for localization, e.g., lower time until a position is derived (latency), and better resolution; requirements that may result in higher demands on a synchronized network. Sensing is a new feature that is expected to be developed for 6G. Sensing, different to localization, refers to determining the position of objects that do not have UE like capability but are located in and around the radio environment. In this regard, sensing may require a new architecture capable of supporting sensing.

5.5 Trustworthiness and sustainability

In Section 4.7 we address the need for 6G architecture shall support the Sustainable Development Goals goal. This can be achieved by significantly improving the 6G system’s energy efficiency compared to earlier network generations but also by a simplified system design, allowing for flexible and specialized network architectures, reduction of CP signalling, etc.

As explained in Section 4.7, the new 6G architecture will most likely need to consider new RED requirements [RED14] as applicable, considering that there are certain implications on privacy and security at the network architecture level. From this perspective, containers may need to be utilised to enhance user privacy and specific protection mechanisms need to be introduced, building for example on “zero-trust” principles. It is critical to ensure AI knowledge-base management by preserving data anonymisation.

Hexa-X proposes to include metadata (e.g., in the form of a tag) in each container indicating the level of privacy of the respective content, along with other metadata/ attributes reflecting the type of data contained (e.g., network or application data). The exact classification requires further study, but, regarding the level of privacy, it may include the following: (i) content not privacy-sensitive, can be shared publicly; (ii) content related to user privacy, may only be used within a

6G protected domain; (iii) content related to highly sensitive user privacy data, must be encrypted, and may only be accessed by specifically authorised functions.

Furthermore, focusing on a given area domain (e.g., a country) where user data privacy may be evaluated differently, user privacy related data need to be suitably protected through: (i) protection of stored, transmitted or otherwise processed data against accidental or unauthorised storage, processing, access or disclosure; (ii) protection of stored, transmitted or otherwise processed data against accidental or unauthorised destruction, loss or alteration or lack of availability; (iii) ensuring that authorised persons, programs or machines are able only to access the personal data, to which their access rights refer; and (iv) accountability and traceability, recording which personal data have been accessed, used or otherwise processed, at what times and by whom.

In addition, further processing and communication of user privacy related data needs to be protected, for example, through the following steps: (i) limit transfer of user privacy related data only over suitably protected communication channels meeting minimum requirements, including a minimum level of encryption and verification of user rights of the recipient; (ii) introduction of an activity-logging tool related to, at a minimum, user authentication, changing system settings and operating the system, and (iii) supporting the users' choice to not disclose certain data, such as location data other than traffic data, relating to users or subscribers of public communications networks or publicly available electronic communications services, or where applicable, to give consent to processing. Some of the above requirements are expected to be mandatory in the future for any radio equipment because of minimum requirements introduced by new articles of the RED Directive. For sustainability purposes, as part of the acceptability process which involves any new mobile generation including 6G, there is the need to consider the minimization of Electro-Magnetic Field (EMF) emissions as a primary network design goal, by considering the most stringent EMF exposure limits adopted in some countries/cities. Moreover, to effectively address the EMF exposure aspects, it is necessary to take them into account from the system design to the network planning, optimization, and operation phases, and throughout the entire equipment life cycle. A fundamental step is the definition of a global, open, and interoperable (i.e., not vendor-specific) EMF exposure assessment procedure, which could be relatively straightforward if the topic is natively considered while defining the 6G network architecture concept.

6 Intelligent Networks

The proposed enablers for Intelligent Networks of Hexa-X are meant to facilitate dynamic adaptability of the network architecture to accommodate new use cases and deployment scenarios beyond what the current cellular networks could offer, while keeping the infrastructure and energy costs at acceptable and sustainable levels. In this chapter, we elaborate on the key enablers identified in Section 5.2 and complement them with several supporting mechanisms and technologies that shall be worked on to realise native AI support in 6G networks.

We start with the need for network programmability, which provides a quick way to validate and experiment with new not-yet-standardized communication features that may commence with limited scope at their early state of deployment but have the potential to become adopted globally with minimal standardisation needs. The view on network programmability extends to network automation by offering mechanisms and interfaces to fast functionality exploration, updates, and reconfiguration (Section 6.1).

The network automation in the Hexa-X architectural framework builds on top of closed-loop automation leveraging data-driven algorithms and in-built analytics (principle 02). Cross-domain and cross data plane data collection and analytics framework (see more in Section 6.2) is part of the network automation that provides input to the AI-agents of the AIaaS framework.

The capabilities of the AIaaS are needed in most of the use case families presented in [D1.2]. Most prominently, AIaaS will be indispensable for implementing the use cases in “he "sustainable

development" use case family and the ones requiring “situational awareness”, such as the “robots to cobots” use case family. AIaaS is based on a set of interconnected AI agents that collect data and make, i.e., using closed-loop network automation, resource allocation decisions on their respective parts (Section 6.3). The compound set of AI-agents are trained and supervised by the AI-driven orchestration and management using training data collected for this purpose. The AIaaS framework exposes its services to various network service. (principle 01). To ensure timely and accurate data collection, new analytic capabilities and related AI protocols shall be explored and introduced. Dynamic instantiation of the AI agents and programmable NFs calls for AI-driven orchestration and capability to place network functions in different processing points across the network fabric (aka dynamic function placement, see more in Section 6.5). The newly instantiated or reconfigured NFs and AI agents need to discover each other, and they need to communicate efficiently via a network service mesh (principle 01). Figure 6-1 shows the studied AI enablers of Intelligent Networks in the context of the different domains of Hexa-X network architecture, see Section 5.1. As can be seen, the Intelligent Networks functions reside in the common platform, the network functions domain, and the management.

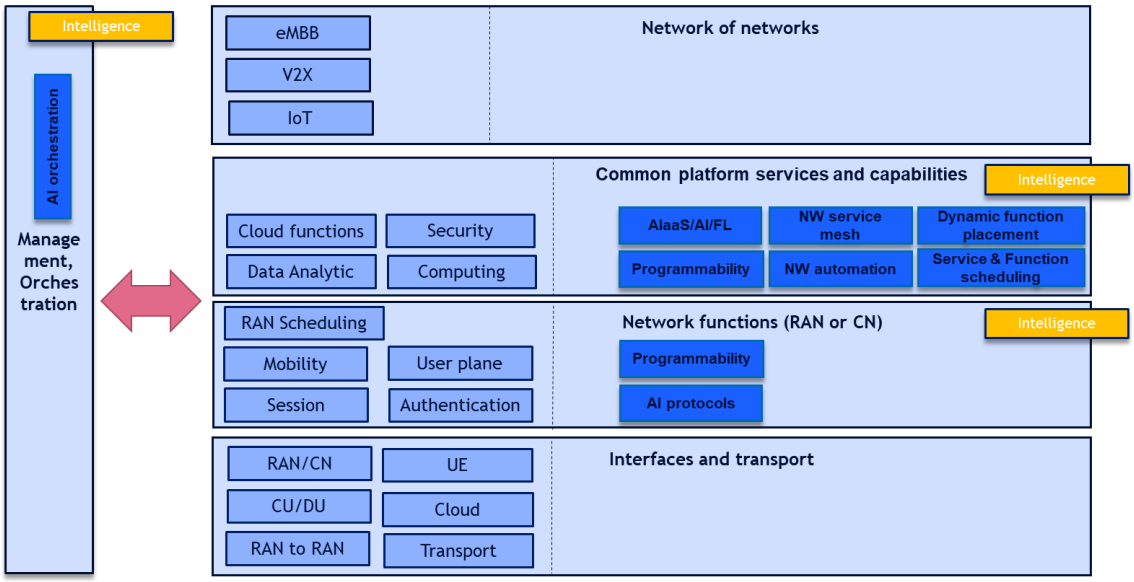

Figure 6-1 Intelligent Networks enablers in dark blue boxes in the context of the different Hexa-X architecture domains.

6.1 UE and Network Programmability

Programmability has been explored mainly in transport networks and various paradigms emerged, such as active networking [BCZ97], [TW07], CP/UP separation [YAG04], [FBR+04], and OpenFlow [MAB+08]. [ATOS] By adopting Software Defined Networking (SDN) and Network Functions Virtualisation (NFV) technologies, the network infrastructure for 5G already offers high flexibility. Recently, with the advent of 5G, programmability is emerging also into mobile networks, as exemplified by the introduction of SBA to the core network (CN) that supports easy addition of new NFs. The use of these approaches is further motivated by the expected diversity of the capabilities of the UEs in new and different contexts (e.g., as tagging and sensing devices, transponders, etc.) in conjunction with the processing capability of edge domains. Programmability can be seen as a complementary approach to standardisation framework on how new features can be introduced and/or improved to enhance network capabilities in an algorithmic way in specific use cases (e.g., enterprise and industrial) and needs where rigorous standardization processes and interoperability is not justified.

In the full softwarization and programmability of UEs and New Radio (NR) will lead to network and protocol stack programmability. Therefore, next generation network would not depend on HW any longer but starts to resemble software versioning. With the help of programmability,

network upgrades will only impact software that is running on general purpose hardware. The next step to potentially capitalize on programmability is to apply it to the air interface protocols (hence, impacting the UE).

Programmability is considered here from multiple viewpoints. Why and what is needed for better programmability of UEs is covered in Subsection 6.1.1. Motivation for better network programmability and what is needed for it is discussed in Subsection 6.1.2. Finally, how we intend to evaluate the suitability of the data plane programming language, P4, for the foreseen needs of 6G is also discussed in Subsection 6.1.2.

6.1.1 Programmability of UEs

The air interface protocols have evolved to be highly configurable with many features for various purposes. However, introducing new features that have an impact on the air interface protocols is time-consuming, as changes should be applied to both UE and gNB via tedious standardization process to achieve consensus between operators, network and device vendors who may all have different priorities. This limitation is even more pressing in dedicated networks, where enterprises call for an integrated networking solution for their operation. UE programmability may be an enabler to help realise the vision of "fit for purpose" promise of dedicated networks for both legacy and upcoming use cases in 6G. For a truly adaptable network able to introduce new changes, a programmable UE is required in dedicated networks enabling faster time to market, faster innovation, and support of verticals to name just a few.

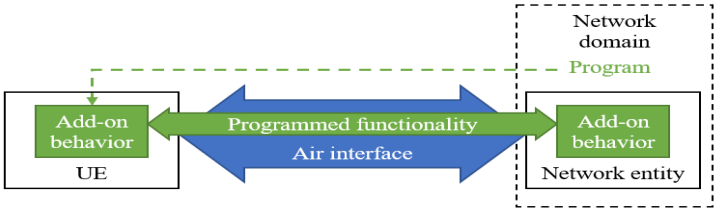

Figure 6-2 Network and UE programmability.

The scope of UE programmability (see Figure 6-2), should be defined with respect to what is intended to accomplish, with it, which is a research question, to strike a balance between pragmatism and vision which functionality is programmable, and by which means considering the needs and requirements from multiple parties.

6.1.2 Programmable Networks

With the advent of B5G/6G networks, the number of UEs will be massive in scale, especially if we consider the devices at the extreme-edge domain. In general terms, it is expected a full digitalization of the real world, which translates in a vast amount of data that must be processed.

A programmable network must enable dynamic changes in devices, functionalities, and parameters to be implemented fast, regardless the number of devices. Scalability and sustainability will drive the design of the 6G, while network programmability will stand as a key piece. Network programmability is expected to cover all available resources from extreme edge to CN, enabling a continuum management, as explained in Section 4.7. Increasing UE programmability will also require improvements in network control. It is expected that new types of softwarized functions will rise within the management, user, and data plane. To extend network programmability from the applications layer down to the network functions and to the data plane it will be necessary to expose new APIs on the wireless devices even to the air interface customisation purposes.

The concept of Wireless Network Operating System (WNOS) [GBD+18] is an anticipation of what will mainly happen to the NFs, operations, and protocol stack. A software network

abstraction will include all the microservices and agents implementing all the services, functionalities, and sub-functionalities of the network. Automated network management can be centralised and/or distributed to deal with issues such as geographically localised management and operational subtasks of specific microservice' or chain of agents. This programmable and intelligent network management and control will apply continuous modifications to all the layers of the programmable stack from the application down to link and physical layers. The Programmable Protocol Stack (PPS) will configure various parameters at each layer. The concept of UE virtualisation will also permit the complete outsourcing of the UE's operations to be executed by microservices or automated agents. This will have the consequence of making the UE fully and adaptively programmable and flexibly adhering to network upgrades. A new challenge can now rise how to provide high reliability in an automated, programmable, and intelligent 6G network.

As mentioned in Section 4.2, to provide ubiquitous connectivity and extremely low latency, deployments at the edge and far-edge locations need to support a wide range of data plane functions and use cases under a limited footprint, i.e., different types of devices each with its own fixed functionality might not all be deployed at the same time in a single location. To this end, P4 [BDG+14, BR18], with its platform-independency feature, can be leveraged to program more general data plane devices, e.g., SmartNICs, NetFPGAs, ASICs or even software, in different locations. P4 compiler needs to abstract different hardware devices and translates P4 constructs into device-specific configurations.

To satisfy the requirement of predictable or even deterministic performance, in particular packet processing latency, it is vital to evaluate and model different components of a P4 program, i.e., constructs, on different platforms. The evaluation results will be analysed considering the required KPIs for different 6G use cases to derive various performance models, which can contribute to an automated decision on which platform to execute a specific NF, given a use case with certain latency requirement and cost constraint.

6.2 Network Automation

The foreseen complexity in operating and managing 5G and beyond networks has propelled the trend toward closed-loop automation of network and service management operations. The ultimate automation target is to enable largely autonomous networks, with no (or minimal) human intervention. Such networks will be equipped with self-x capabilities, including self-configuration, self-monitoring, self-healing, and self-optimisation. This approach requires an end-to-end framework designed for closed-loop automation and optimised for data-driven ML and AI-algorithms. On network automation support by the next generation architecture, the zero touch network, and Service Management (ZSM) framework [EGZ19] together with the existing research work on network automation at different levels of the network, provide a set of technical enablers to achieve automation. The introduced abstractions within the network and between network layers are the starting points to the enhancements and implementation of network automation. Different perspectives of the support of automation as an architectural enabler beyond ZSM will be considered: the abstraction level with respect to the technological background, and how much it should accommodate support for (i) AI, (ii) Internet of Things (IoT), (iii) the presence of Multiple network Operators, (iv) the concept of Zero Touch Provisioning (ZTP), and v) network monitoring.

AI and ML are more than ever considered as "the enablers" for automation, but their integration brings new technological challenges [BT20]. For example, the black box nature of big learning models calls for explainable AI that can both infer and trace the root causes of the output [DVS2020] [KO21]. Special attention needs to be directed to detailing AI and ML concepts for the edge-to-network integration when considering massive number of IoT devices [JDN21] [MCS+19] given its growing presence and dependence on edge computing resources.

Networks and network edges are growing into an ever-growing variety of NFs that, coupled with devices with proprietary functionality, lead to network ossification and difficulty in network management and service provisioning. Abstractions and separation of concerns in a multi-vendor environment brought by NFV for programmability are key in the move towards automation. The new levels of automation in network and service management and orchestration, rely on closed-loop automation [SG21] as also largely discussed in ETSI ISG ZSM [EGZ19]. These management loops usually comprise different steps that are executed one after the other at runtime: for instance, Monitoring, Analysis, Planning, and Execution for self-adaptive networks [LSK19]. This process of network automation is often supported by an automatic model building and model updating processes.

As the network elements are no longer a composition of integrated hardware and software entities, the evolution of both hardware and software becomes increasingly independent of each other. The detachment of software from hardware helps with reassigning and sharing infrastructure resources. This enables network operators to deploy new network services faster on the same physical platform. Therefore, components can be instantiated at any NFV-enabled device with any underlying hardware in the network and their connections can be set up flexibly. Challenges lie in the optimal placement of processing in terms of used resources, latency, dependability and in the way how service requests are routed between Cloud-native NFs. Particularly in multi-domain environments distributed decision making and local optimizations may be needed. This implies that the placement of Cloud-native NFs needs to be adjusted to match the changing requirements of the end users. The solution can be built using the enablers for dynamic function placement (Section 6.5) and network service mesh (Section 6.6) which are made available to network automation and orchestration to support the dynamically changing end user requirements, new network topologies and different service structures (e.g., clusters of VMs interacting in different clouds).

The proposed principles and enablers in Section 5 for a fully softwarized, programmable, and intelligent 6G network architecture, calls for a further step in the automation of the Network of Networks. In fact, in future networks, the automation will not only embrace individual functions or operations (or their sub-modules), but it will also be able to empower the activities of virtual operators. The vision of Autonomic Mobile Virtual Network Operator (AMVNO) [GB18] [GBa18] relies on the dynamic physical network infrastructure, which includes deployable access points with a remote radio head (RRH) (e.g., UAV-based base stations), SDN switches, big/micro/pico datacentres, and their respective satellite-based counterparts (e.g., satellite-based SDN switches, etc.). This dynamicity involves includes the PPS and intelligent resource reservation, management, and operations. The core of the envisioned system is the autonomic manager, which contains the intelligent hypervisor and the intelligent business hypervisor (see Figure 6-3-) [GB18] [GBa18]. The former can use intelligence for all the operations related to network management, while the latter can use AI to adapt economical aspects, such as pricing and expenses for network expansion. The deployment of ML by mobile operators has already been suggested since 2017, for smarter capital spending, automation, and simplification in the back office, predictive analytics in marketing and sales, more efficient customer retention and support.

The intelligent business hypervisor is linked to the intelligent hypervisor so that the exchange of information and their complex interactions can change rules, pricing policies and can also shape the network architecture and characteristics according to the different environmental regulations. Figure 6-3 depicts the logic blocks of an agent based AMVNO considering the SDN-NFV ETSI MANO architecture combined with the softwarized PPS [GB18] [GBa18] and AI/ML for fully automated operations.

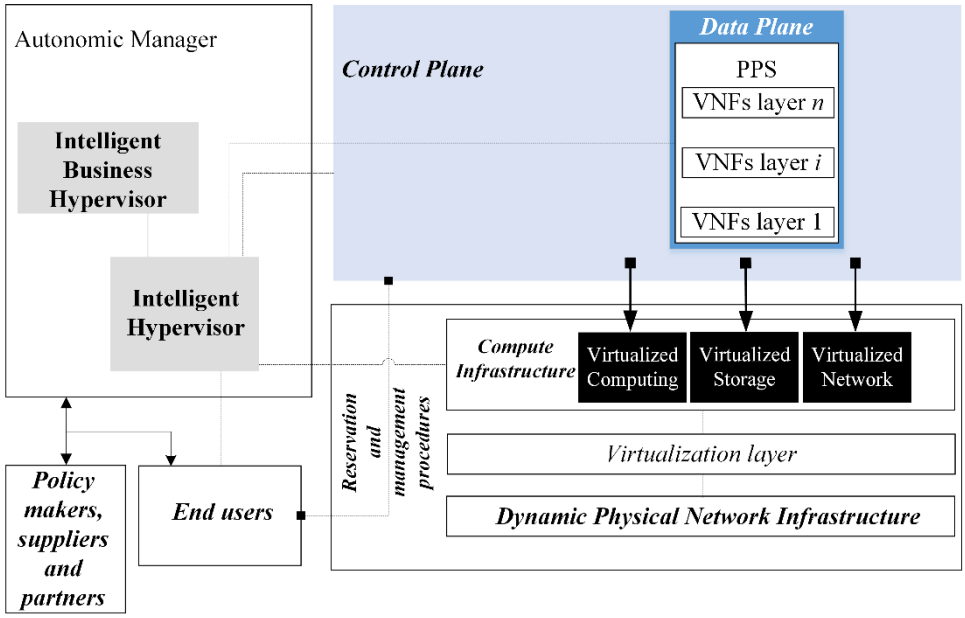

Figure 6-3 Logic blocks of an agent based AMVNO.

Although in 5G the network analytic capabilities are enhancing to be able to use AI/ML techniques to provide more accurate statistics, prediction, and analytics reports, they are considered as a " closed box" and limited to the boundaries of an operator. For instance, core network analytic capabilities are equivalent to a limited list of reports NWDAF can offer. Moreover, there are no methods of analytics exchange between operators. Furthermore, performing those analytics required input data sets coming from pre-specific sources. Due to the limitation of the analytic capability to an operator's boundaries, the 5G analytic entities are not capable of accepting input from 3rd party applications.6G should include analytics capabilities to any application requirements coming from any 3rd party applications. –With the help of AI systems, the analytic capabilities envisioned for 6G can analyse data and uncover hidden trends, patterns, and insights in a more automatic fashion. This can lead to a flexible 6G analytics in the sense that it will be open to input from any data source and be able to provide the requested analytics report to any analytics consumer.

Various distributed entities exist which provide local analytics such as NWDAF (Network Data Analytics Function) in the CP, or MDAS (Management Data Analytics Service) in the management plane. These entities can take advantage of cross-plane/domain knowledge and resources in developing the analytics and training the ML algorithm with input data set. The results of AI analytics stand apart from traditional analytics in various aspects. An AI-based anomaly detection solution learns the pattern of the normal behaviour of the data by constantly monitoring and analysing huge amounts of data and can find an ”outlier" based on the self-observation of the normal behaviour of the data. In addition, an AI-based analytics solution leverages clustering and correlation algorithms constantly and in real-time to detect more accurate anomalies so that any issues can be remediated as soon as possible.

With the help of AI supported by ML and Big Data analytics techniques [BT20], advanced network automation solutions can analyse meta-data (e.g., the descriptive information about a resource, specific characteristics or procedures used to collect the data) and leverage model-driven network programmability to learn network behaviours, deliver predictive analysis, and provide recommendations to network operations teams. These advanced automation solutions can be configured to take remedial action autonomously. They can provide closed-loop corrective actions, to avoid foreseen network issues, at times, even before they even occur. Closed-loop correlation in general is a step-by-step process control for preventing actions or the management of failures of any type. These timely detections are based predictions on collected, analysed, and possibly shared learnings from similar deployments. In doing so, network automation improves

operational efficiency, reduces the potential for human error, increases network service availability, and delivers a better customer experience ideally without human intervention.

6.3 AI as a Service

As detailed in [D1.2], the AIaaS concept can be applied for tasks, such as predictive network resource allocation to mitigate QoS drops or intent classification and prediction in human-to-human and human-to-machine interactions. Such interactions can be based on different criteria/features such as: e.g., gesture, intonation, expressions, surrounding sounds, touching objects. An AI service following this concept can be consumed by either NFs, or instantiated by a user or IoT devices, or by the network infrastructure. The AI service can also be consumed by AFs, which are external to the network (e.g., via NEF) submitting requests for ML-based inferencing decisions to applications in the network. The key difference over state-of-the-art is, indeed, the general availability of AIaaS functionalities to end users and end user devices through well-defined interfaces. Nonetheless, exposure of these interfaces and APIs would need to be accompanied by lightweight security and trust (e.g., authorisation) mechanisms, as it is important to keep network-internal processing (and its complexity) and operator-sensitive information separated from the application layer. An API abstraction layer would be needed for this purpose, aiming to require minimal skills from a service consumer and/or AI application developer. Architecturally, several network entities, and their corresponding (standardised) interfaces, would be needed for efficient offering of AIaaS in an open manner, across systems owned and managed by different network operators and vendors:

- an AI orchestration function responsible for maintaining an overall view of available in-network AI resources and topologies (e.g., learning federations), adding/removing instants of data analytic functions and AI agents;
- an AI repository function for registration of available AI agents and their offered services,
- an AI policy enforcer to implement the recommended learning/inferencing policy, and, optionally,
- an AI monitoring function to monitor the quality, efficiency, and security of the implemented policy.

Like with the above-mentioned AI agent selection procedure, there is a need for standardised network interfaces and services for registering, discovering, and selecting data/model producers. A data/model producer can be a user/IoT device, a network infrastructure entity, an AF external to the network or a combination thereof. Figure 6-4 illustrates the proposed entities to support AIaaS. Note that the entities shown in Figure 6-4 mainly introduce management functionalities and, thus, require access to further networking entities, including a knowledge database, etc. It should be noted that this framework resembles the one of network orchestration and policy enforcement, however, the scope of the proposed functional entities is specific to how a 6G network performs governance of its AI capability.

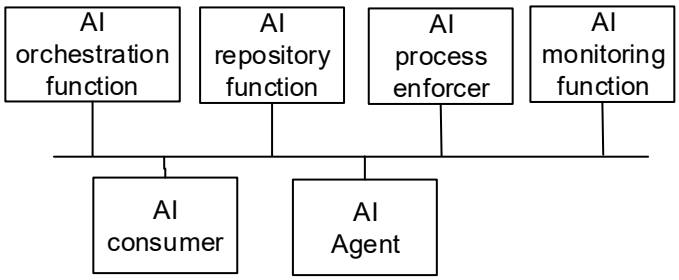

Figure 6-4 Architecture enablers for supporting AIaaS.

The new AI enabled architecture aims to support distributed AI services, needed for supporting AI as close as possible to the application, AI service chaining needed to accomplish a specific AI task, as well as cross-domain AI service consumer and data producers. The proposed in-network AI architectural framework for AI governance also aims to support AI enabled access control considering attributes such as user, data object and environment information. It also aims to facilitate the efficient transfer of large amounts of data, network/ application-specific analytics, and the sharing of AI models, once available and updated. The new architecture supporting AIaaS will be employed for enabling different learning services, such as Federated Learning and feature XAI. Once a consumer requests to exploit the Federated Learning service, mechanisms to allocate resources and to instantiate the required functions should be envisaged. Based on the service requirements and the capabilities of mobile device's capabilities, the AIaaS-supporting 6G network will be able to decide which functions of the service will be instantiated. On one hand, the mobile device may produce data, build local AI models, receive the aggregated model from the Federated Learning server and make decisions based on it. On the other hand, resource-constrained devices (e.g., sensors and actuators) may delegate data aggregation and learning functions to a Digital Twin instantiated at the edge of the network (e.g., a MEC application), while the device only takes care of sensing the environment. Also, the Federated Learning server may be implemented as one or more functions residing in either the cloud or the edge environment, based on both the service requirements and edge nodes' capabilities. In the context of MEC, MEC federation may be employed to extend the scale and coverage of the service, by allocating functions under the domain of several MNOs. Based on how the different Federated Learning and XAI functions are distributed in the device-edge continuum, flexible protocols will be envisaged.

6.3.1 Requirements for AI-enabler protocols

6G network ought to be flexible to dynamically adapt to the variation of mobile service demand and the local upgrade can take place in a few neighbouring cells or even in a single cell in order to flexibly and dynamically implement cutting-edge developments in subnetworks without extensive time-consuming tests. However, large-scale network reconfiguration and ad hoc management can result in high complexity and time-consuming operations. To avoid high complexity and time-consuming operations, Artificial intelligence unsupervised management approaches are proposed. With the help of the local data collected by each subnetwork, AI-enabled closed-loop optimisation techniques may be exploited for the dynamic improvement and management of the networks. For this purpose, recent advances in AI-enabled optimisation techniques such as game theory and learning approaches may improve the efficiency of the networks. Additionally, local upgrade of subnetworks may require a relatively robust CP to support the interaction in the "network of subnetworks" level. Therefore, by using AI-enabled optimisation techniques, (i) the evolution may be evaluated first on the subnetwork level partially, and (ii) network-level optimisation/learning techniques may be implemented later considering changes in local environment and user behaviour.

AI/ML has recently achieved a breakthrough in tackling real-world complicated problems with the advancements in deep learning and hardware technologies. The research community has already shown the feasibility of deploying AI/ML models to implement and assist various network functionalities [ASR+20]. An example where an integrated solution is required to enable the full potential of AI/ML is RAN. Examples of the target use cases include [D4.1]: implementation of PHY-layer functions, channel state predictions, resource allocation, predicting UE trajectory and handover triggers, etc.

In another view, 6G is expected to enable large-scale deployment of AI/ML agents, i.e., providing communication for AI/ML. For example, in a factory automation, AI/ML is not limited to only augment the end-to-end control process, like today, but the actual control process itself will be designed to be AI-driven, moving from regular sensors and actuators to AI-based ones. AI agents with increased intelligence and autonomy will form self-governed sub-systems interacting with other sub-systems. For example, the multitude of intelligent sensors, robots, tools on a factory floor creating different sub-systems and these sub-systems together form a self-governed

production system, which requires constant communication and distributed intelligence across its elements.

Both views, i.e., (i) AI/ML for communication and (ii) communication for AI/ML, require new functional entities and protocol-wise aspects to be efficient, and it is still to be understood which requirements, like signalling, etc., and their realisation over the air interface would require for AI models, involving the transfer of information related to AI such as data, models, and algorithms, between gnB and UEs. New protocols and/or enhancements to legacy protocols are required for training, inference, and maintenance. The training methods and algorithms are evolving very fast, as evident from recent breakthroughs [GoogleAIBlog], thus, the protocol aspects addressing the training of such models, including the model/data distribution techniques, loss minimisation algorithms, compression techniques, etc., should be adaptable to the pace of evolution and innovation. Requirements for inference and actuation need to be well understood before the requirements for the enhancements for the protocols (e.g., signalling) can be defined to implement the input /output operations of the AI/ML agents.

The seamless transfer of insights across domains and planes and the conveyance of data analytics among distributed network microservices and agents result in additional fundamental challenges in designing and realising efficient and effective protocols for managing those operations. "First, to apply ML procedures efficiently, it is necessary to design novel, reliable, and efficient protocols for data mining and traffic (extended to network) engineering. These protocols are pivotal for the virtual intelligent functions/agents/services to get the status of the network (local or metropolitan) and thus to take decisions and act on the actual network infrastructure. Second, multi-agent systems will also perform collaborative-distributed intelligence. These collaborative ML procedures need data distribution protocols that perform the communication through the logical service/agent chain to complete an intelligent task in addition to effective synchronization protocols. It is important to underline that these intelligent network entities will run in different areas of network infrastructure e.g., different servers or different data centres. This implies the need for reliable communication protocols for transferring analytics and their processed results across the multi-agent collaborative system(s). Regarding synchronisation, packet-based protocols like Precision Time Protocol [IEEE19], are very efficient and flexible. Then, the design of more advanced synchronisation protocols for the scenarios mentioned above is a key challenge as well.

6.3.2 Protocol Framework for AIaaS

One of the incurred challenges of realising pervasive use of AI across a 6G system is how to design a protocol allowing a UE carrying a local ML model (obtained by e.g., training a Neural – network - NN) to seamlessly exploit the knowledge of large parts of the network. Such knowledge is useful to its locally undertaken inferencing tasks, by attaching to/ detaching from different learning deployments (e.g., federations) across multiple operator areas. There are several criteria calling for (frequent) UE attachments to/ detachments from a learning structure (e.g., centralised or FL), including user mobility, unavailability of an AI agent due to e.g., a detected security attack, low-quality connectivity, increased model aggregation latency and others. In this section, the aim is to present an overall protocol framework for AIaaS, referring to an exemplary scenario on automotive communications. Protocol details will be documented in forthcoming Hexa-X deliverables.

To tackle this issue, an AI Information Service (AIS) and its corresponding AI Application Programming Interface (AI API) can be introduced to a 6G network, implemented over an open network interface. Such a service and its accompanied API may enable a UE (AIS consumer) to communicate to the AIS information relating to a user/ client application-specific task (e.g., intention to drive a vehicle from location A to location B, starting at time t) calling for an inferencing-based recommendation (e.g., detecting an object/ roadblock) and performance requirements (KPIs) relating to e.g., inferencing accuracy, energy efficiency, end-to-end delay, security, and others. All these criteria are filtering criteria for AI agent selection. The approach

involving an AIS between an AI consumer (e.g., a UE) and an AI capability provider (AI agent), AI operation enablement may follow the AI supporting architecture is illustrated in Figure 6-5.

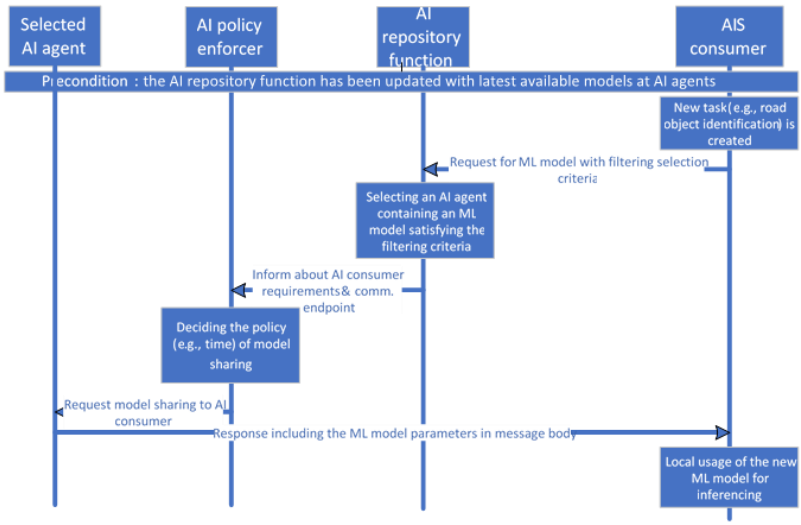

Figure 6-5 Signalling flow for requesting and delivering a ML model satisfying AI agent selection criteria provided by an AI consumer (e.g., UE).

This approach will enable the UE, based on AIS response on available AI agent(s) fulfilling the communicated criteria, to: (i) in case of a commonly supported application layer protocol, subscribe to, unsubscribe from, or update the subscription to one or multiple available AI agents (e.g., FL aggregators) or (ii) in case infrequent/ one-time output is needed to obtain the ML model configuration indirectly from the AIS. As a result, considering each selected AI agent, the UE will be able to share its local model updates to the AI agent(s) it is subscribed to and obtain learning system parameter updates (e.g., aggregated FL model update, transfer of an already trained and tested model) by the subscribed AI agent(s). The advantage of the proposed AI-supporting protocol (at application layer) is to enable an AIS consumer (e.g., end user device) to exploit part or all the knowledge available in a network – without requiring the network owner to expose the knowledge (and thus a key asset) directly. Rather, the AIS consumer receives a trained model (or access to using such a trained model in the network), which is derived through network internal processing of the available knowledge.

6.3.3 Protocols for Federated Learning

To allow mobile devices to exploit FL services provided by the network, the FL algorithms specifically designed for the 6G network (as from WP4) need to be supported by new protocols handling the required interactions among the involved entities. Mobile devices will be able to query a list/registry of available FL services in a given coverage area, each providing at least information about its objective (e.g., optimisation of QoS for V2X) and requirements (e.g., minimum storage required for AI models on the mobile device). Protocols for allowing mobile devices to join a federation, or even trigger the creation of a new one, will also be envisaged. Similarly, mobile devices may leave federations they have joined previously, either voluntarily or triggered by an external event (e.g., the device getting out of the coverage area of such FL service). Then, mobile devices will need to effectively transfer local models and/or aggregated data to the FL server, as well as receive the updated global model from it. Considering the expectable huge number of mobile devices connected to the 6G network, employed protocols for the above operations must be designed to be scalable too. For this reason, different paradigms will be envisaged, i.e., event-based and time-triggered information exchange (or both), and the

6G network will select the most suitable one according to context information, such as federation type and scale, radio access conditions and availability of computing resources.

Moreover, when the FL service is used to collaboratively build XAI models, the 6G network will implement protocols that allow mobile devices to receive explanations about the inferences made by the algorithms themselves, by either explicitly requesting (i.e., using a client-server paradigm) or subscribing to explanation updates (i.e., using a publish-subscribe paradigm), balancing possible trade-offs between effectiveness and overhead of the different approaches.

6.4 AI-driven Orchestration

While the use of AI/ML is already happening to support network management processes for 5G and B5G networks and services, e.g., O-RAN [ORAN21] its seamless and transparent integration with network and service orchestration platforms need to significantly improve to address the complexity and heterogeneity of 6G networks. On the one hand, 6G network and service orchestration can benefit from AI/ML to assist its operations and processes, including service planning and NF placement, service scaling, Service Level Agreement (SLA) management, resource arbitration and sharing, etc. Here, AI/ML techniques and algorithms can output predictions leveraged to support proactive orchestration actions [D4.1]. On the other hand, AI agents and pipelines need to become more seamlessly integrated with the network and service orchestration platforms. Currently, such interaction is mostly happening with ad-hoc and proprietary means by embedding and integrating pre-trained AI/ML algorithms within the network and service orchestration decision logics to assist and drive specific orchestration actions.

Several solutions are already available for managing AI/ML algorithms and pipelines following a unified approach and leveraging on cloud-native technologies to deploy and run AI/ML agents in virtualised environments. Tools like Airflow [Air], Seldon [SEL] and Kubeflow [Kubf] allow to manage the lifecycle of AI pipelines and agents on top of Kubernetes as containerized applications, providing means to cover different aspects from algorithm training to deployment and serving in cloud-native edge-ready virtualised infrastructures. Following this technology trend, AI-driven 6G network and service orchestration platforms could benefit from the availability of and seamless integration with AI/ML orchestration engines offering common services for algorithms and agents lifecycle management (i.e., covering their automated instantiation, configuration, runtime operation, termination), providing and supporting unified APIs, data models and metadata for capabilities discovery, data source requirements description, and invocation of algorithm training (and re-training), execution, serving and evaluation services. Moreover, considering the envisaged highly distributed 6G network and service orchestration approach, aligned with zero touch and ETSI ZSM principles [EGZ19], a deep integration and cooperation of per-domain and per-slice subnet orchestration engines (e.g., taking care of managing services and resources at heterogeneous far-edge, edge, and core domains) will be required. Here, a common and unified approach for the interaction among local AI/ML orchestration engines and network and service orchestration platforms would significantly simplify and ease the end-to-end 6G network management processes covering the UE, far-edge, edge, and core networks.

The integration of AI/ML techniques in the management and orchestration context will be a key enabler to make possible the implementation of the continuum orchestration model (see Section 4.6 also). Specifically, AI/ML techniques will help to manage the high number of devices and heterogeneity of data sources in the extreme-edge domain, as well as provide the capacity of correlating data from different domains (extreme edge, edge, and CN) and provide prediction capabilities based on that data heterogeneity (see Subsection 4.1.4). AI/ML techniques can be used to manage the complexity described above [KGC+20]. Different ML algorithms (e.g., supervised, and unsupervised learning, federated learning or reinforcement learning algorithms) can be used to efficiently improve the management and orchestration processes, implementing orchestration actions such as predictive NF scaling or placement, automated healing, proactive

alerting, hidden patterns discovery, proactive incident analysis, closed-loop automation, or support to network slices profiling, among others [BFS21].

AI/ML techniques should be combined with other technologies, such as virtualisation technologies (already in use in 5G), SDN, and extensive zero touch management techniques. The solution to the problem of intelligent orchestration can be addressed by considering the CP completely distributed, thus actually implementing collaborative and intelligent agent-based systems across the network Multi-agent-based network automation (MANA) [ABG+21] can be a possible solution for transforming the centralised network management and orchestration into a distributed problem, where various intelligent agents collaborate with each other. MANA architecture can in implemented by autonomic distributed agents that implement NFs or some sub-functions of them. These agents are autonomic and 'atomic' units capable of performing orchestration tasks, while interacting with the environment as well. Modularity is a characteristic that can help reducing the complexity of network orchestration. The selection of an appropriate set of autonomic agents allows for building a complete multi-agent system to replace the complex and distributed monolithic orchestrators. Agents share information with each other by passing standard agent communication language (ACL) messages for coordination and reaching consensus on the orchestration policy to be followed. Moreover, the orchestration process consists of different types of agents that can also be organised in hierarchical layers depending on their scope and functionality (e.g., infrastructure, network services and applications). This layered approach can also be used for the multi-agent-based design and realization of the orchestration. In fact, the orchestrator will need to manage virtual-network characteristics (e.g., placement of function and agents in line with the communication network topological aspects), internal network node functionalities and operations. Moreover, it will have to orchestrate sub-functions' organization and placement (in case functional split is used), and protocol-level orchestration (in case the PPS has to be instantiated and adapted). However, such a system can significantly raise security and reliability issues, and the control traffic related to the orchestration procedures.

6.5 Dynamic Function Placement

Network Function Placement refers to the concept of deploying functions to orchestrate differentiated services optimally across multiple sites and clouds based on diverse intents and policy constraints of dynamically changing environments. Network Slicing, which splits the physical network into multiple virtual networks and enables an array of differentiated services, is enabled by the operation of function placement known as the cornerstone of the 5G network. However, with the advent of 6G, the envisaged use cases with evolved KPIs and KVIs and flexible, dynamic, intelligent, and self-evolving architecture required by these use cases are expected to add an automation dimension to the network slicing. The requirement for automation is anticipated to pave the path forward to dynamic network function with AI utilisation enabling the use-case-specific slice formation and intelligent orchestration of use-case-specific slices. The direction of the future network is to enhance the scalability, extensibility, customisation, and continuous integration. For this purpose, specifically, 6G envisages extending the service-based architecture from the core network to end-to-end service-based architecture. In addition to Dynamic Function Placement (DFP) algorithms, the use of AI with NFV or Platform-as-a-Service infrastructure is required. Additionally, expected AI services for future systems are evolving from simple function-based AI services to distributed AI services based on agents and can be deployed as a network service, a NF, a virtual machine, or a software component injected in a VM. There is a need to identify how and where to deploy these distributed AI services composed in different forms, identify diverse use case requirements that trigger different optimisation objectives for DFP, and investigate the implications of deploying and maintaining these distributed AI functions.

The relationship between orchestration and DFP and their roles in Life-Cycle Management (LCM) of NFs is one key item to be studied: What kind of DFP concept is needed to handle NFs in 6G environment? Should DFP be supported over multiple domains, or, instead, be confined

within domains, while multi-domain support is provided by multi-domain orchestration, for instance? Does 6G require updates to the existing DFP features like NF monitoring, and NF scaling? Will these features be enriched with NF mobility and NF offloading and do they need to support multi-domain clouds? Certain NFs like AMF (Access and Mobility management Function) and NRF have close relation to the orchestration (e.g., by selecting NFs based on attributes like service areas and load that are set up by the orchestrator) and they need to work in close synchronization. How much can a synchronized context transfer be used directly between old and newly created NFs and is there a need for different types shared data bases between domains? How NF context sharing works over domain boundaries is an important study item also in terms of backwards compatibility with the existing orchestration and service management "frameworks".

5G NRF is responsible for dynamic NF selection based on requested criteria, such as load of the candidate NF providers. Because NRF (and NWDAF) has the knowledge which NFs are requested to act as providers and what are their workloads it should interact the orchestrator to trigger DFP to instantiate new NF instances as needed. The role of DFP and how it relates to NRF needs to be investigated and defined. Furthermore, NRF and DFP interaction needs to consider strict timing constraints (like further explained in the 'extreme experience'-research challenge in [D1.2]) and use of the limited resources in the connected edge clouds for hosting NFs closer to the end users. Network Slicing is a tool for resource isolation and relating to DFP it adds additional dimensions onto the optimisation problem by combining placement, isolation, and connectivity (see Section 6.6) aspects. 6G shows the need to evolve the concept of NS, which should expand to new cloud domains and maybe be even end-user specific extreme edge domains. The importance of NS is expected to increase and therefore it is one study item.

Figure 6-6 proposes a high-level phasing scheme towards a fully autonomous DFP approach. As we can see, it is divided into three general phases:

- Phase 1 (static), there will be pre-instantiated NF placement, e.g., in a peak-hour, a car parking is usually full, the solution could be to scale up NFs to ensure enough resources are available when needed.
- Phase 2 (adaptive), semiautomatic NF placement in a closed-loop environment based on policies. The orchestrator could choose from various locations in real time.
- Phase 3 (autonomous), dynamic orchestration and re-allocation of NFs. AI/ML can be used to take decisions. Also, placement can be declarative configured using intent-based approaches [IBN].

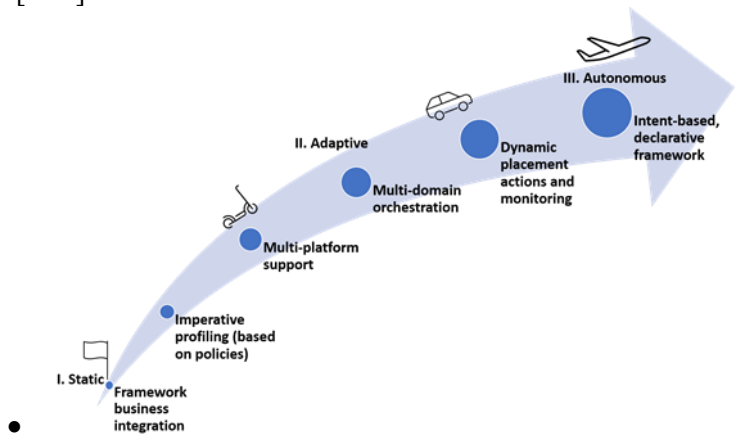

Figure 6-6 Road to fully autonomous NF placement

In the current release of 5G, the 1st (static) and 2nd (dynamic) phases are already in place, mainly focusing on managing the placement of NFs between core and edge domains. However, the target should be to reach phase 3 (autonomous and intent-based) in the upcoming 6G networks. In B5G/6G, core and edge domains are typically located in a supervised and controlled data centres, while devices at the extreme edge are placed in a sort of "uncontrolled" environment (the end-

user's domain) being fault-prone and having a highly dynamic and asynchronous behaviour. In this context, DFP should be highly dynamic, which poses an additional challenge to this concept. Probably event-driven mechanisms should be incorporated to provide fair response times to the state changes of the virtual functions in this environment.

In addition to the asynchronous nature of devices at the extreme edge, another factor to consider is their limited resources (mainly computing and storage) in comparison to infrastructure resources in the core or edge networks. This limits the type and size of NFs, as well as the type of Infrastructure Managers that could be deployed on them. Since migration and placement of virtual functions is ultimately handled from the Infrastructure Manager, it is likely that this component will need to be evolved to deal with this type of infrastructure (in this regard there are already available in the state of the art some components that could be used as an Infrastructure Manager in this context, e.g., KS3 Lightweight Kubernetes [KS3], Minikube [Min] or Microk8s [MK8s]).–Another factor to take in account is the DFP orchestration itself, i.e., the algorithms associated to the triggering of the placement actions on the infrastructure. In this regard, a high automation level should be the target in this context. Also, considering the potentially massive number of devices on the extreme edge and the consequent high heterogeneity of the data sources, AI/ML techniques should be considered as a relevant key enabler.

6.6 Network Service Meshes

Network Service Meshes aim at providing full control and management of L2/L3 connectivity in cloud-native mesh topologies, going beyond the current limitations of available technologies, e.g., supported by Kubernetes, that provide connectivity control for cloud native applications and NFs at the application layer only (i.e., L7). Therefore, the network service mesh is intended to support application-to-application (i.e., UP) and function-to-function (i.e., CP) communications in 6G networks. This can be enabled through the provisioning of dynamic and automated virtual network services, to be allocated on-demand, based on application requirements. Network service meshes will need to provide UP functionalities for L2/L3 connections (to be exploited by higher layer protocols and services), and mechanisms to forward data between virtual network service endpoints, as well as CP functionalities for handling virtual network service discovery, routing, and connection management to create the virtual wires, similarly as in [RFC3985].

The concepts of edge and far-edge computing, already (partially) addressed with 5G, become more and more relevant for 6G architecture and services. Here, cloud-native technologies will be required to create cloudlets at the edge of the network, with application-to-application (i.e., at UP) and function-to-function (i.e., at CP) communication capable to satisfy a large number of interconnected assets with flexible mesh topologies. Current cloud-native solutions like (application) service mesh allow the management of service-to-service communications among microservices, with some limitations. As depicted in Figure 6-7, service meshes have two main components: the UP and the CP. The UP is the actual mesh, where the communications between services happen through the proxies (called "sidecars"), while the CP configures the UP to correctly intercept and route traffic according to the desired mesh topology.

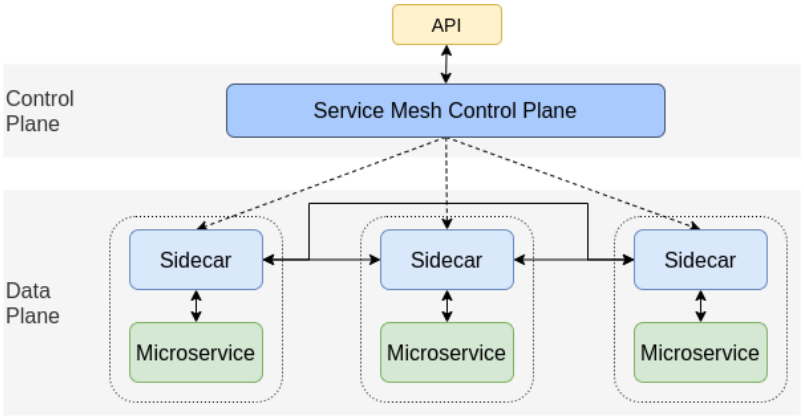

Figure 6-7 Service Mesh concept and components.

This (application) service mesh is a mature technology in its current use cases that can already support different deployment models for different application scenarios for single domain, with several available options for both UP (such as NGINX [NGI], HAProxy [HAP], and Envoy [ENV]) and CP (such as Istio [IST], Linkerd [LINK], and Consul [CONS]). However, they provide full control of communications at the application layer only (i.e., L7), while they offer minimal control at the transport layer, making them unsuitable for very flexible and dynamic mesh topologies in cloud-based virtualised and cloud-native infrastructures which require closer coordination with lower layers.

To fulfil the flexibility and dynamicity requirements imposed by 6G on cloud-native applications, services and infrastructures, the concept of network service meshes needs to be further developed as part of the Hexa-X architecture. Here, flexibility in controlling and managing the network service mesh is required to follow the dynamicity of cloud-native applications and functions that will adapt to 6G service mobility and elasticity. The Hexa-X research objective in this context is to support hybrid scenarios where network service meshes integrate cloud-native and virtual machine-based applications and functions, possibly considering aggregation of different meshes to realise a federation where each mesh exposes virtual network services enabling communication across mesh boundaries. The main goal is to create a more flexible and responsive network (mesh) environment when compared to traditional NFV deployments, which mostly use static forwarding graphs. In this direction, the application and function adaptability can be enhanced by leveraging serverless computing, thus introducing stateless functions executed and integrated in the mesh topology on-demand when needed.

In practice, NSM establishes a communication bus/channel between multiple (operator's managed) domains, enabling seamless networking between them (in ZSM terminology [EGZ19] this is called inter-domain integration fabric). Each domain could have its own virtualisation infrastructure that does not necessarily support inter-domain networking (in ZSM terminology [EGZ19] this is called intra-domain integration fabric). With NSM, IP reachability beyond domain boundaries is added and, therefore, also other functionalities built on top of inter-domain communication could be then supported. Typically, NSM implementations use L2 or L3 tunnelling to interconnect domains, but a higher layer tunnelling, e.g., based on IETF's MASQUE WG [MAS], could also be considered. In other words, NSM makes tunnelling and the related networking context an important functionality.

NSM must support most common communication patterns like Publish-Subscribe, and Request-Response communication, since, for instance, SBA NFs communicate with each other using these two patterns. Additionally, both synchronous and asynchronous communication must be supported. NSM provides registry (or similar) for (de-)registering services that are then exposed via service discovery, which is used to find out that what services are available in NSM and how. This is where similarities with 5G SBA can be observed and where both NSM and SBA provide

similar functionalities but with different scopes and perhaps at different layer, i.e., networking layer vs. service layer.

Networking policies also play important role in NSM, and they must be enforced in distributed manner, i.e., a set of networking policies distributed over the related domains are dealt in the same manner. Depending on how NSM is going to be integrated with SBA, there is a need to verify how SBA's network policies, namely Policy Control Function (PCF), relate to the respective policies in the NSM. Additionally, the transfer of OAM and AI related knowledge information may have different connectivity requirements between the related domains and their internals to harmonise management operations.

Regarding NSM and SBA, the key items to be studied are as follows:

i) Is Service Based Interface (SBI) connectivity built as a flat or hierarchical communication channel over multiple domains and what is the role of Service Communication Proxy (SCP).
ii) How to integrate the selected NSM approach with SBA, since both alternatives (flat and hierarchical) require different things?
iii) What are the roles of the GSMA specified 4 communication models (i.e., direct/indirect via NRF or SCP) [GS5] in the multi-domain scope?
iv) What are the dependencies and interactions between NSM and SBA policy domains?

7 Flexible Networks

Flexible networks intend to enable extreme performance and global service coverage. The network functionality and architecture must then be flexible enough so that it can adapt to these different topologies (please see principle 03 hereafter). 5.1OBJ concerns the Flexible networks:

- Principle 03: Flexibility to different topologies
  - *Adapt to scenarios such as new private networks, autonomous networks, mesh networks, new spectrum, etc.*
- Principle 04: Scalability
  - *Support very small to very large-scale deployments.*
  - *Dynamic scalability by scaling up and down network resources.*
- Principle 05: Resilience and availability
  - *Using multi connectivity and separation of CP and UP.*
  - *Support of local network survivability in case a subnetwork loses connectivity with another network, remove single point of failures.*

First, a 'Network of Networks' (Figure 7-1) enables the principles above with the integration of a) Non Terrestrial Networks (NTN), b) Flexible multi-connectivity (Flexible MC) and/or combined cell/multi-point transmission, c) Sub Terahertz nodes (Sub-THz) and Visible Light Communication (VLC), d) L1/2-mobility, D-MIMO (Multi-TRP/D-MIMO), e) Flexible Device to Device communications (Flexible D2D) in terms of both Distributed D2D (mesh) and Centralized/Operator controlled D2D communications, and f) Campus/non-public networks.

Second, after the 'network of networks' is built, architectural/technology enablers are developed that can manage local ad hoc networks, in coordination with the network (physical) infrastructure, as well as distributing their functionalities between them and at the edge.

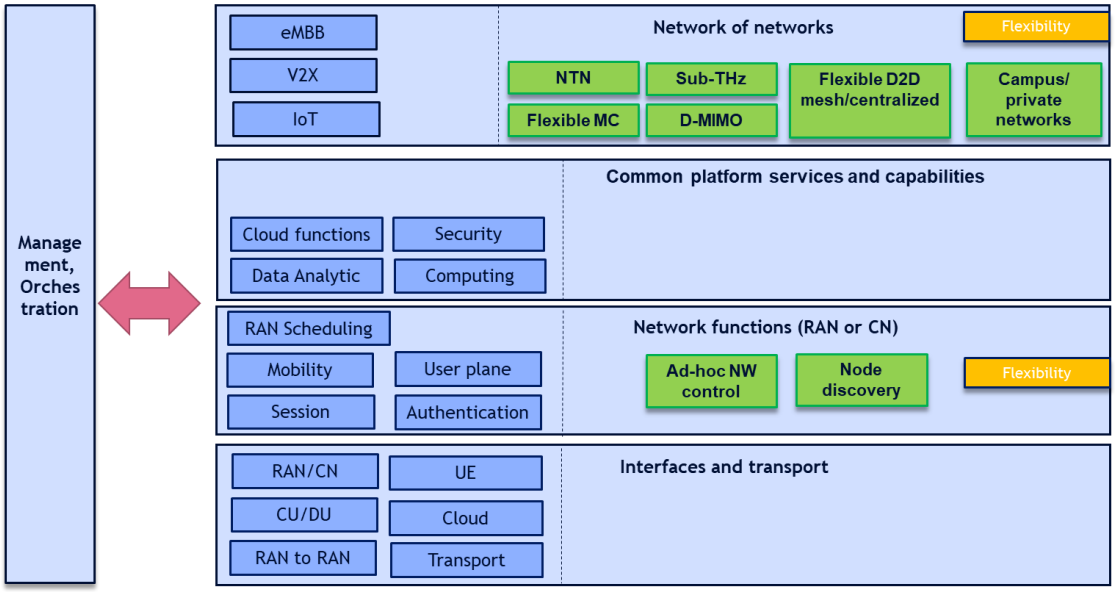

Figure 7-1 Network of Networks (Flexible network) high level functional overview, the enablers, and solutions.

More specifically, the Local Controller enables to change the topology and form the local structure by means of Network/Node discovery and selection function; it selects the nodes that will be admitted in the ad-hoc network formation, it selects the spectrum and Network of Networks technology options to activate, as well as it splits the functionality between the nodes,

participating in the local structure, and finally decommission the structure when there is no need to have it.

In addition, the Local controller comprises the specification of Network/Node control & interfaces to control and interact with far-edge devices for resource advertisements, synchronization, reachability verification, etc.

Finally, the Information Modelling allows a unified modelling of far-edge nodes and devices, in terms of Networks of Networks and computational resource characteristics, capabilities and constraints. This means that the Flexible networks concept enables a Distributed Intelligence approach, where the Orchestration framework (WP6) selects local controllers and assigns specific network-compute requirements and then the Flexible network/Local Controller decides on local structures (nodes with networking, incl. Ad-hoc, and computing resources, terminated at edge node) as coordinated extensions of infrastructure (temporary).

7.1 Network of Networks

As mentioned in section 1.2, 5G enables a dual connectivity (DC) solution with 4G (LTE), a so-called tight interworking (or tight integration). The interworking between LTE and NR is part of the Rel-15 standard for NR, i.e., the EN-DC (E-UTRAN- NR DC) solution [37.340]. The assumption was that the NR FR2 frequency bands are much higher (e.g., 28 GHz) than the LTE frequency bands and therefore the NR cells will have worse coverage than LTE. The EN-DC solution enabled a smooth integration of 5G for operators with 4G coverage. The 5G capacity can be utilized efficiently when a user has 5G coverage, with minimal risk of radio link failure. However, with the advent of 6G it is expected that even higher frequencies will be utilized, the so-called upper mmWave bands (i.e., 100-300 GHz, see [D2.1]). This means that 6G may include frequencies from 450 MHz up to 300 GHz. Due to the nature of the radio propagation the pathloss in free space is increased with 6 dB for each doubling of the carrier frequency[6]; thus, the difference is more than 55 dB between highest and lowest frequencies [Hat80]. If more realistic propagation models (i.e., non-free space models typically used by 3GPP) are utilized the difference can be even higher, see [D2.1].

Figure 7-2 shows an example of the expected 6G Network of networks, with a wide range of different cell types and frequencies as well as different type of networks interworking with each other.

[6] The increase in pathloss is actually due to the decrease of the amount of power an antenna can receive, since the antenna size is proportional to the square of the wavelength

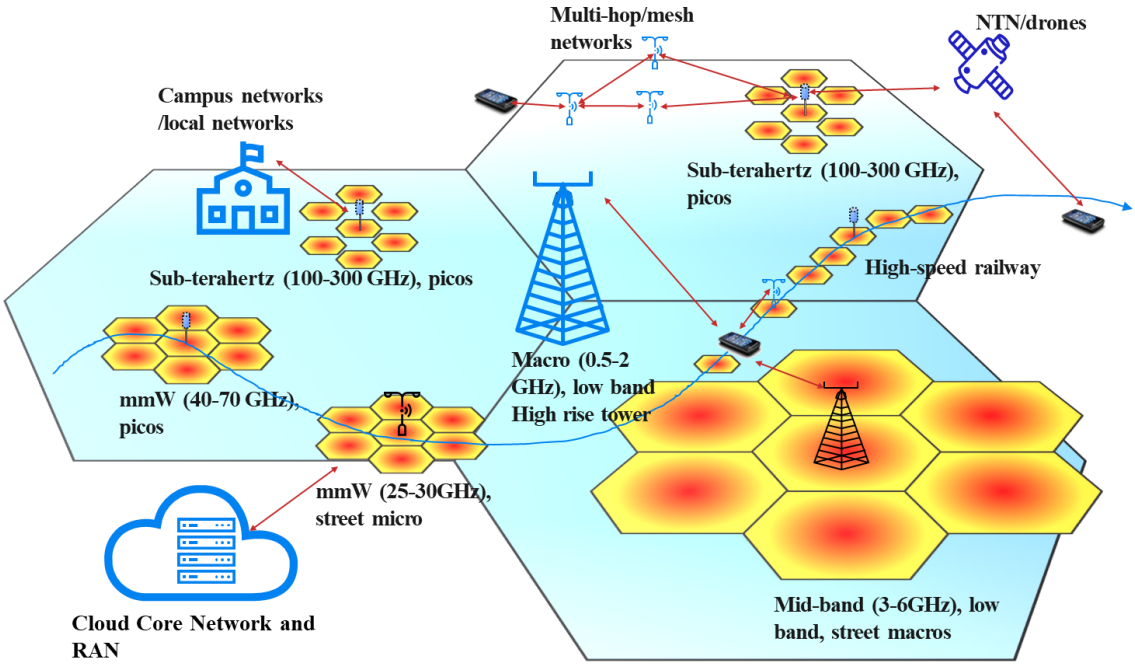

Figure 7-2 The 6G Network of networks will include wide range of cell types, frequencies, and deployments.

In the Figure 7-2 example, the macro cells are using low frequencies around 1-2 GHz with very wide coverage. These macro cells typically are mounted high on masts or on top of buildings, which further enhances the coverage. The 6G networks will also include smaller cells, likely using higher frequencies such as mid-band (3-6 GHz) or mmWave bands (around 30 GHz), with more spotty coverage. Finally, we will have very spotty coverage of upper mmWave nodes, around 10-100 m [D2.1]. The new mobility required to support the network of networks is described in section 7.1.1. Thereafter we describe methods to support mesh networks with D2D in section 7.1.2 and finally, we describe the Campus concept in section 7.1.3.

7.1.1 New mobility solutions for 6G

7.1.1.1 Non-terrestrial networks

Non-terrestrial networks (NTNs) can provide coverage to exceptionally large and isolated areas. NTNs can likely support a low capacity per square km, which suits rural areas. For urban areas, there will always be a need for terrestrial Networks. NTN is not limited to satellites; drones and High-Altitude Platform Station (HAPS) are also included. However, in this section we focus on the satellites. There are two types of architecture options for NTN: Transparent and Regenerative payload. Transparent is the simplest type, where the NTN basically serves as a simple relay of the signal between the UE and the base station on ground, see Figure 7-3. Regenerative payload is equivalent to having the base station (RAN) functions onboard the satellite, see Figure 7-3. For both cases there is a need for a gateway on the surface to connect to the terrestrial network.

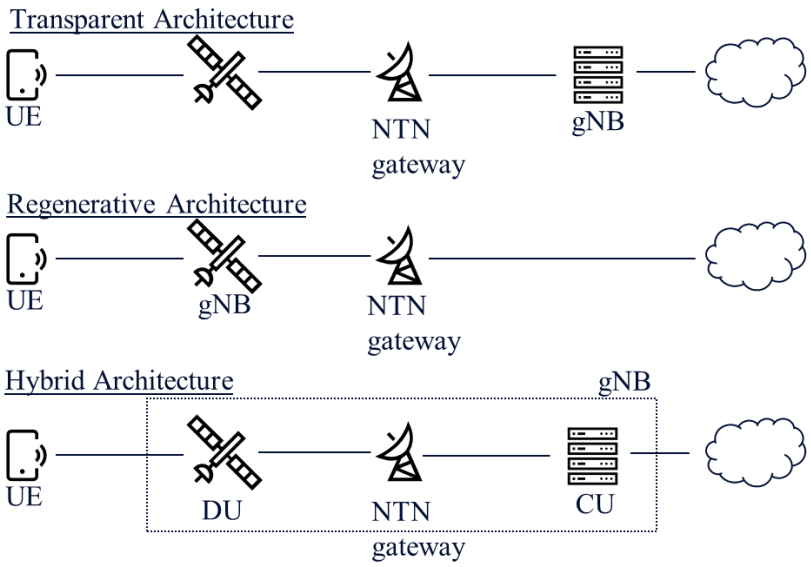

Figure 7-3 Possible Satellite architectures in 5G. Using the transparent payload, the basically acts as a relay since the gNB is on the earth surface. With the regenerative architecture, is equivalent to having base station functions onboard the satellite

The main advantage with transparent architecture is that there is no need for hardware that supports a full gNB, which means that the weight of the satellite can be lower when compared to the case when the gNB functionality is onboard the satellite as for the regenerative architecture. The main advantage with regenerative architecture is that it is more capable and can be upgraded when new functionality is available. Also, it may be easier to cover areas where a ground station cannot be built by using multi-hop between the satellites to find satellites connected to a gateway on the ground. This can then for example be done by reusing e.g., Integrated Access Backhaul (IAB) or D2D type functionality for multi-hop. There is also an option for a RAN-split hybrid solution, where part of the gNB functionality (protocol stack) is boarded on the satellite, e.g., DU (Distributed Unit) in the satellite and the CU (Central unit) on the ground. Note that in both cases there is an NTN gateway on the ground which performs routing of the connections to the correct core network nodes. Since there is a limited number of gateways on the ground, the satellite needs to make a gateway switch now and then as it moves.

In 3GPP for rel-17/18, there has been a moderate update of specification enhancements to enable the support for, e.g., improved synchronization due to doppler fading, improved Hybrid Automatic Repeat Request (HARQ) processes due to long delay etc. It is expected that more advanced NR features will be standardized at a later stage, or possibly in the 6G timeframe. Current NTN solutions in 3GPP aim at normal devices for low frequencies or devices with external high-gain antennas at high frequencies.

The main research question for 6G is how the NTN and terrestrial network mobility will be solved. Since the satellites are moving, it may be needed to find solutions to minimize the number of handovers and the signalling needs for mobility robustness. Another important topic for a 6G NTN is the actual architecture solution, e.g., if regenerative or transparent or a hybrid split shall be used.

7.1.1.2 Multi-connectivity and sub-terahertz mobility

For 6G, it is expected that a flexible multi-connectivity and/or combined cell/multi-point transmission will be important to integrate higher frequencies and to ensure high reliability mobility also for higher frequencies. However, a disadvantage with the EN-DC solution is the specification complexity [38.331], both in terms of standardization time and specification impact (of both 4G and 5G). Other ways to ensure mobility may be mesh networks, e.g., via gNB

integrated access backhaul nodes that can create a very dense network without the need for wireline and the use of NTN and drones.

Other interworking solutions that have been discussed and standardized in 3GPP include Carrier Aggregation (CA), NR-DC (dual connectivity between NR nodes), Supplementary Uplink (SUL) and Multi-connectivity (MC). The SUL is when the UE can make use of up to two different carriers in the UL, at different frequencies but only one carrier in downlink. The supplementary (SUL) carrier is deployed typically at a lower frequency than the uplink carrier and can therefore be expected to provide better coverage. Thus, if the uplink carrier does not have good enough coverage during random access, the UE will try to access via the SUL carrier.

MC can be seen as a natural extension of DC, but with more than two cells involved, e.g., one master cell and two secondary cells. For complexity reasons, this has not been agreed in 3GPP for 5G. Both NR-DC and SUL are however part of the standard for rel-15. In addition to the discussions above, interworking solutions have been also considered in 5G-CLARITY project [CGG+20] and also in in 3GPP with focus on integration of 3GPP and non-3GPP RATs [23.501] and their combined use, leveraging on Access Traffic Steering Switching and Splitting (ATSSS, see [24.193]). Potentially, for 6G, it might be relevant to evaluate the feasibility of similar frameworks in supporting not only RAT-integration but also MC solutions. The research questions here are to find cost efficient solutions for high reliability mobility, avoiding radio link failure and data transmission interruption. It is likely, as discussed above, that there is not one single solution, but several solutions need to be combined. In particular, we will investigate several different architectural options for mobility.

7.1.1.3 Visible Light Communication

Visible Light Communication (VLC) is part of the optical wireless communication technologies. [D2.1] gives an overview of the different variants but here we focus on Visible Light Communication (VLC). As said in section 3.8, VLC is perhaps the most promising of the different optical wireless communication technologies and is considered one of the candidate solutions for 6G, mainly, for indoor short-range deployments. In VLC, communication is achieved using e.g., LEDs (Light Emitting Diodes) together with intensity modulation techniques in the optical bands between 410 THz – 790 THz (380 nm – 730 nm)

With VLC a very high cell densification may be possible due to limited interference. Also, security and privacy can be considered inherently since there is no wall penetration, if all the signal paths toward any windows are obstructed. Some of the main challenges for VLC are the backhauling in dense deployments, the signal blockage (since near line of sight is required), and the UL coverage.

IEEE is currently specifying 802.11bb [802.11bb], which is an extension of normal 802.11 but with a Light Communications Amendment. The uplink and downlink operations are in 380 nm to 5000 nm band, i.e., it also includes infrared spectrum.

If 6G chooses to include the VLC spectrum, there are several options on how to integrate it, e.g.:

Use 6G and IEEE 802.11bb interworking protocols e.g., using LTE-WLAN aggregation (LWA) or NR traffic steering (ATSSS, [24.193]).

- Define a whole new 3GPP protocol for VLC, including PHY, MAC and RLC.
- Define a new protocol for VLC as “License-assisted access” (LAA), like NR-U.

LWA currently enables the device to use both LTE and WLAN at once [36.300]. The PDCP is split over the Xw interface, see Figure 7-4. The eNB decides which bearers should be aggregated over WLAN.

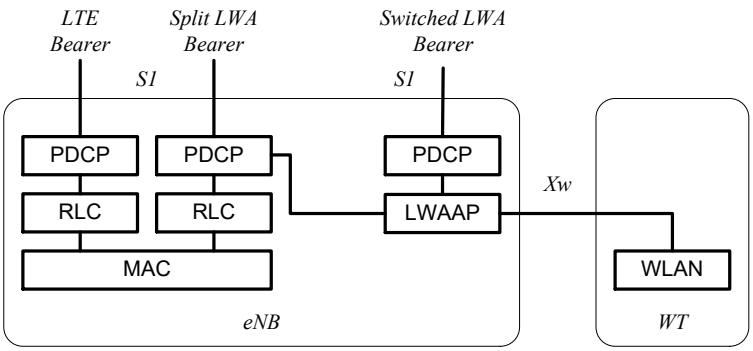

Figure 7-4 LTE WLAN Aggregation (LWA) [36.300] see section 22A.

There is also an option to offload traffic bearers using a so-called fast switch over Xw (switched LWA Bearer in the Figure 7-4, which is known as RAN controlled LTE WLAN Interworking [36.300]. The main research question here is which of the different integration options are most suitable for 6G, in terms of efficiency and cost aspects.

7.1.1.4 L1/2 and D-MIMO mobility

L1/2 mobility techniques such as D-MIMO and Multi-TRP., assume that the mobility is handled at the PHY layer. This means that UE does not need to update the Radio Resource Control (RRC) reconfiguration in this area, the UE continues to use the same configuration as before.

D-MIMO may be a component for 6G systems, due to the potential improvements in system capacity. This may especially be true for the user with worse channel gains, e.g., the users further away from transmitting node or blocked by an object. The L1-mobility system relies on a system with several access points (APs) connected to a central unit (CU) via high-capacity fronthaul transport network. In a given region, all the APs connected to the particular CU typically utilize same resources but without fixed cell borders, which is referred to as a MIMO cluster area. Ideally, the UEs in the MIMO cluster area are connected to all APs. However, for complexity reasons and resource utilization, it may be beneficial that the UEs only connect to a subset of the APs, which is referred to as the UE's AP cluster area [DIB21]. This means that the UE must still select one or multiple APs in the area best suited for transmission and reception. Each time the UE updates its AP cluster it may have to update:

- the fronthaul paths if necessary,
- centralized and/or distributed precoders,
- the link adaptation.

If all UEs can use resources from all APs simultaneously, there may be no need to change the AP fronthaul paths (depending on fronthaul deployment). When the UE's AP cluster is updated, the precoder weights need to be updated due to fast fading per AP. Thus, the UE's AP cluster does not need to be updated as frequently as precoder since the main driver affecting AP-UE association is the slow fading, which mainly depends on the distance to the APs.

The main research questions here are how the D-MIMO and L1 mobility impact on 6G architecture / 6G mobility procedures and how to design the necessary measurements and signalling for this.

7.1.2 D2D and mesh

The concept of device-to-device (D2D) communications is not new per se as it has been discussed since 4G networks [DGK+13] to enable devices to communicate directly in an infrastructure-less manner, offering significant gains. First, the proximity of users can allow even higher QoS levels and lower transmission powers. Moreover, radio resources can be reused more efficiently. The same D2D link can be utilized for both downlink and uplink [FDM+12]. In addition, D2D

communications among UEs can enhance the coverage and capacity of cellular networks through traffic relay [VFS12]. Therefore, "Network of Networks" aspect envisages the usage of technologies for supporting flexible topologies, the introduction and "connection" of enhanced intelligence, and cost efficiency. Flexible topologies will essentially be alternatives that will increase the availability and reliability of infrastructures. New intelligence will be needed to make decisions "on the fly". Cost efficiency (e.g., limited resource and energy consumption) should underpin all operations.

Two main approaches can be considered as described below, in order to achieve D2D communications:

- Distributed approach, where devices use mesh protocols to find neighbours and connect autonomously: In [CDY+12] D2D communications are exploited in intelligent and cognitive networks. Specifically, users can either transmit to the BS, or, if they can, communicate directly with other users via D2D and can form a D2D group. Also, a distributed protocol is proposed for resource allocation and transmission mode selection.
- Centralized approach (or operator controlled) in which devices will connect to each other as designated by a central management entity (e.g., an operator): For instance, in [LZL+12] a classification of operator controlled D2D communications is provided, according to the level of operator control.

The following technologies can be considered for D2D and mesh networking:

- WiFi Direct [WIDIR] /WFA EasyMesh [WIMES] can be utilized as the enabler technology for D2D communications in urban environments to offload traffic of cellular networks. Protocol 802.11s [802.11s] can be considered for mesh connectivity between nodes.
- Proximity Services (ProSe) [23.303] standard has been already introduced in LTE late releases for direct-mode communications when network connectivity is not yet available. It can work today, but its range is limited because of the low power levels of devices.
- Low-power wide-area network protocols for IoT devices (e.g., LoRaWAN), in which Long Range LoRa [LORA] network is used to implement a mesh topology. Each device can be equipped with LoRa Transceiver and some of the devices can have both LoRa and cellular network-based connectivity. All the data will hop to LoRa devices until a device with cellular network-based connectivity is reached to transfer the packet to the cloud. When the data reaches the cloud, the server is responsible for dropping duplicated packets.

The following challenges can be observed for the researching related to D2D and mesh networking:

- How much "trusted" is a device in order to be part of the D2D/mesh network.
- Unified modelling of far-edge nodes and devices, in terms of network and computational resource characteristics, capabilities and constraints.
- Definition of interfaces to control and interact with far-edge devices for resource advertisements, synchronization, reachability verification, etc.
- Design algorithms for selecting best possible nodes and far-edge devices depending on specific parameters (e.g., position, signal quality, battery level, availability, reachability, available computational resources etc.).
- Integration with network and service orchestration for seamless management, control, and enforcement of D2D/mesh network communications to satisfy end-users application constraints and requirements. It includes proper abstraction of D2D mesh network topologies towards orchestration layers.
- Methods and procedures for discovery of nodes and far-edge devices (including synchronization aspects for capabilities advertisement).
- Best possible routing for multi-hop D2D communications (creation of routing tables), minimizing latency or increasing resilience and cost efficiency.
- Select technology/technologies to use for D2D communications.

7.1.3 Campus Networks

Campus Networks are used by industry, municipalities, or educational institutions. Campus Networks are networks of LANs in specific areas, which can use heterogeneous access technologies. These networks have limited geographical coverage, as they are designed to serve only factories or small towns. Given the requirements of very low latency (even below 1 ms), ultra-high reliability and flexibility, Campus Networks will be a pillar of 6G architecture. They can also ensure the high-level of security and trustworthiness, lower energy usage and sustainability, that is necessary for various critical and highly sensitive use cases of 6G. Moreover, Campus Networks can be used for closed user groups, resource guarantees, local network management, and for the availability of exclusive radio bands. In particular, closed user groups are not only attractive from a security viewpoint, but also from a business model and business offering perspective.

6G Campus Networks will be a pivotal RAN-edge paradigm of the network architecture since they will be important to deploy nano edge data centres for RAN softwarization and the supporting of low-latency verticals. They will mainly support mobility within their domain without necessarily interconnect to other external networks. For example, 6G Campus solutions can be deployed by network operators to provide customised and effective solutions to their industrial customers. This will allow 6G operators to ensure the quality and the performance of the specific local Campus Networks that are deployed. Side by side, small operators can also benefit from Campus solutions.

Furthermore, the 6G scenario of three-dimensional network of networks opens the way to three-dimensional Campus Networks, based on aerial platforms and UAVs. 6G Campus Networks can provide an on-demand, low-latency, and secure and resilient RAN and MEC by exploiting the mobility and flexibility of aerial platforms. The network connectivity can be provided via balloons (or other HAP platforms) and several hovering drones. Both HAPS and UAVs are equipped with communication interfaces and computing/storage. In this way, in-HAP MEC can also provide computing resources and virtualization means to define end-to-end slices of the available resources to meet the KPIs and enable the deployment of 6G critical and sensitive services, especially in rural and remote areas. Additionally, it can also host value-adding service for network and users' applications.

In respect of terrestrial Campus Networks, in 3D Campus Networks, a proper resource allocation is even more critical. The HAPS must adjust their altitude and positioning based on environmental and atmospheric conditions. This is very important to ensure the constant service availability and the satisfaction of KPIs requested by the verticals hosted. The HAPS have longer battery life than mobile BSs (UAVs) and can potentially carry heavier weight and cover a relatively larger area. This means that HAPS will realise the Campus Network interconnecting for example wireless LANs served by UAVs. The research community has proposed mmWaves and free-space optics (FSO), however some attenuation problems may arise. Next, LTE have also been proposed as potential radio technologies for this scope. As already mentioned, the RAN can be implemented in a distributed way through a swarm of UAVs that are positioned in local areas where the sensitive verticals require services. Information regarding traffic load, radio quality, and so on of end users, UAVs, and HAPS are therefore available through mobile far-edge platforms and can be used to better orchestrate the limited resources on the UAVs. Edge applications running at the UAVs or HAPS can send their bandwidth requirements to their respective mobile edge platforms.

7.2 Edge-to-Network-Cloud integration enabler

Edge cloud is used to decentralize resources (compute, network, storage) to the edge of the network; in that respect, it is formed mostly by resource-constrained user equipment and devices working on-demand for the edge cloud, having a very different nature from the cellular networks and their radio access networks. Many open issues remain to integrate the edge cloud into the

same orchestration platform as other 5G network segments. This would have to include the management of the mobile small cells, hosting the edge cloud and the resource slicing operations within the mobile small cell itself.

As network functions can be dynamically provided in nodes of an edge network or infrastructure, the first challenge is to analyse network function requirements to guarantee the network resource allocation and performance for the services. Moreover, the second challenge is to adapt the network to complex, changing environments and enabling smart services, by means of AI and ML enablers, making these powerful tools to realize an edge to network integration. All these challenges need a systematic approach to provide an integrated solution with a primary goal to integrate the edge cloud infrastructures into operators' network architectures through integrated network control and management solutions for the compute, network and storage resources of edge nodes, functions and edge-enabled network instantiation considering a unified or collaborative-distributed orchestration framework.

Therefore, we propose for the Hexa-X architecture to develop and validate the Edge-to-network integration concept as an extension to operators' networks with a unified CP, integrated network orchestration methods, and intelligent resource and NFP (see Section 6.5) and management to support the wide variety of 6G use cases and services for society, industry, entertainment and public health and safety applications.

One of the major proposals to feed Hexa-X architecture enablers is to develop flexible strategies for the deployment of edge-enabled network functions within the mobile small cells to support different local and time-sensitive services as well as coverage extension.

Here we have examined how to intelligently deploy network-integrated edge infrastructures capable of providing hyper-local, low latency services. The different devices and computing/storage resources in the local area will communicate with each other and with the rest of the network to provide seamless services in a highly dynamic environment with varying traffic and changing channel and mobility conditions. The network-integrated edge using the mobile small cells will also be deployed for coverage extension, with moving network elements or terminals acting as relay nodes acting as intelligent network-integrated Edge with network function deployment on demand. After the definition and development of the relevant network-integrated Edge functions, the time taken to deploy/instantiate these in appropriate hardware will be assessed against targeted times. The network-integrated edge will enable intelligent decisions on computing and storage resource use, including user device cooperation and cooperation with MEC to ensure low-latency. The determination of appropriate storage and computing resources will be shown to offer latency, energy savings through the offloading of data traffic from the core network or operators' cloud, and the avoidance of repetitive data transfer.

The edge-to-network integration enables support of multi-domain end-to-end network slicing and network management. Indeed, network slicing is necessary to extend the slicing paradigm onto all edge resources so that the slice deployment is supported in an optimized way over all possible resources with changing environments in an end-to-end manner especially the Edge-enabled ones allowing to meet the requirements of URLLC-like services. AI and ML techniques are also be used for integrated Edge-enabled service slice definition and selection. The performance of the proposed solutions will be measured through service-level KPIs for end-to-end latency and information loss. Targets will be typically less than 10 ms latency for time-sensitive applications and 99.999% reliability of information transfer [URLL]. The targeted KPIs will be measured over periods of time during which the edge network has to adapt to changing environments, e.g., using different link connectivity to sustain the connection/slice. Slice management will also validate slice isolation by measuring KPIs when more than one demanding slice is set-up.

From the RAN and physical layer perspectives, it is an important enabler to have enhanced PHY, MAC and network techniques for network-integrated edge, and rely on AI and ML techniques to boost throughput, reduce latency, increase reliability, and enhance energy efficiency.

Here, a target is to rely on 5G and 6G radio technologies for network-integrated edge through some necessary enhancements. Using the 5G NR RAT, in addition to other RATs like IoT ones (LoRa and NB-IoT), will facilitate the integration of the edge resources with the mobile small cells under the same network infrastructure as domains. Investigate security functions in edge-to-network integration and orchestration is also one of the enablers for edge-to-network integration. The main security concerns are related to the level of trustworthiness of the integrated edge components or resources in an accurate way with regards to the end-to-end security in which all network components are managed by assumably trusted entities. In such a fully virtualized architecture, all the hardware in data centres and in edge nodes are owned by known and authenticated infrastructure operators. However, in the more distributed edge network, things become far more complicated. The extreme edge paradigm extends the computing and storage towards the edge of the network including even the user devices. Thus, security constraints in a network-integrated or "integratable" edge are mainly related to the authentication of edge nodes, the trust model, and the distributed user data storage.

8 Efficient Networks

As introduced in section 5.4, areas which we will attempt to enhance comprise: *interfaces*, which need to be prepared for cloud operation (principle 06 in section 5.1); *complexity*, by reducing signalling between different nodes and fewer external interfaces, (principle 08)); and *function separation*, to clarify each network functions' responsibility and functionality (principle 07). All the above-mentioned functions need to be handled while at the same time considering *sustainability* [D1.2], including circle economy principles, e.g., ensuring that use of material for producing a network is optimized and that losses at end-of-life are minimized Already in current networks some functions are being prepared for a more dynamic future, for instance, some functions are being virtualized. By making sure that unnecessary interactions are avoided in a cloud-native implementation of the CN and RAN, it should be possible to dispense with some of this signalling, thus, reducing power consumption. A network that makes the most out of the possibilities with virtualised functions, e.g., only virtualizing functions that benefit from this, is an Efficient network.

The work in Efficient Networks follows principles such as:

- Network functions are designed to be prepared for cloud operation, considering that some functions benefit from cloud deployment while, e.g., functions requiring HW acceleration might not benefit as much. (principle 06)
- Exposed interfaces are service based, and it should be easy to add new services to the network (principle 06).
- Aim for NF separation in order to maximize the potential for function reuse (principle 07).
- Ensure separation of concerns of network functions since minimal dependency with other network functions results in network functions that can be developed and replaced independently from each other (principle 07).
- Introduce network simplification in comparison to previous generations, e.g., by reducing complexity and the number of parameters needed to configure cloud-native RAN and CN functions (principle 08).

To support the Key value indicators for 6G [D1.2] the capability of the network needs to be sufficiently equipped to do so. For instance, future networks will most likely need to handle more diverse devices, such as the reduced capability (RedCap) [RP-211574] devices being standardized right now. Diverse devices in future networks may be modular further reducing capability (and device cost), where each module provides a set of capabilities and sometimes by combining modules it is possible to solve tasks. Further, technology developments in the form of drones and satellites, together with the availability of higher frequencies, will also lead to more diverse networks. Just handling mobility efficiently in all cases may require updated network capabilities. Also, the network shall be able to provide functions that support all anticipated AI operations.

The resulting architecture needs to support all different types of traffic. However, one particularly demanding type of traffic that can be mentioned is that generated by Digital Twins, i.e. a digital representation or model of real-world things, places, or processes [Ohl21]. There will be large amounts of data that may be processed by both human and AI operators. This data processing can be viewing, modifying, and scheduling various activities. The model can be generated with data from multiple sensors (IoT), collecting data (data ingestion), analysing data (data management and analysis), visualizing, making decision and then simulating the digital representation. For this data collecting and processing to be possible the network, supporting twinning functions, needs to provide low latency and high reliability as well as high capacity. Finally, two very important responsibilities of the architecture are to provide trustworthiness and ensure sustainability.

Figure 8-1 depicts the enablers for the Efficient network task and in which domain the enablers belong to, namely the Interfaces and transport domain and the Network functions domain.

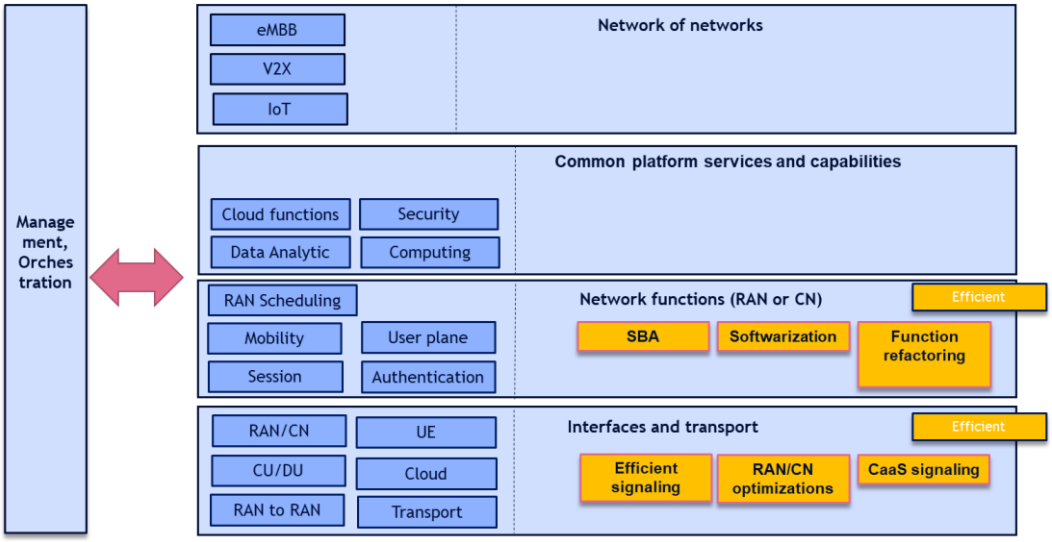

Figure 8-1 Efficient Networks enablers in orange boxes in the context of Hexa-X layered functional architecture.

The enablers in Efficient Networks comprise:

- Efficient signalling and RAN/CN optimizations: Streamlined signalling interactions across RAN and CN interface;
- Service based access (SBA) and softwarization
- Network function refactoring, run-time scheduling of network functions;
- Compute as a Service;
- And finally, reducing cost (Total Cost of Ownership (TCO)) and complexity utilizing a cloud-native service-based RAN and CN.

8.1 Architecture transformation with cloud and SBA

A key factor in the 6G architectural transformation is the impact of cloud technology and SBA to the RAN and possibly to the UE. Harmonizing the CP technology across RAN and core boundaries by adopting a cloud-native SBA approach is expected to provide operational efficiencies through a shared Service Management and Orchestration (SMO) layer, a common security- and protocol framework, common data collection and shared data layer for functionality that operates on feasible near-real timescales. A consistent SBA approach will enable more rapid innovation in the RAN and edge-based services. Even more importantly, TCO benefits are expected from a common policy framework between RAN and CN that would localize the processing of signalling messages to those NFs only that need to process the information. Thanks to virtualization and dynamic use case specific function placement of CN and RAN CP functions the signalling latency and response times can be optimized for a given use case (not all services require ultra-low latency). The evolutionary approach to co-locate CN (v)NFs with RAN (v)NFs on the same cloud infrastructure would provide a fast way forward but would keep the current interfaces and dependencies intact leading to unnecessary processing and protocol overheads. Refactoring the functionality into cloud native network functions across RAN-CN and removing duplications (e.g., unnecessary tunnelling, addressing, policies), and harmonizing the CP protocol framework (e.g., replacing SCTP with HTTP/2, etc.) will result in reduced UP and CP traffic between RAN and CN, and therefore leads to a simpler architecture with TCO and operational benefits. At same time, the functionality refactoring should consider a relevant and efficient split of RAN functionality, including split options and new functionality proposed in Open Radio Access Network (O-RAN), Multi-access Edge Computing (MEC), Service Based Management Architecture (SBMA) and core to yield an efficient overall architecture with built-in security. A consistently refactored SBA-based functionality across RAN-CN will enable new groupings of

CU-CP with relevant core functionality enabling new deployment scenarios for NPN, network sharing, changing topologies, and slicing. We will study refactoring of RAN-CN functionality in conjunction with O-RAN and MEC functionality and based on that we will further investigate how a cloud optimized functionality distribution can be adapted to various scenarios such as NPN, autonomous networks, mesh networks, etc., without loss of performance or need for over provisioning and that enables easy deployment that can scale into very local as well as into global use cases. Viable migration paths from the current RAN splits (CU-CP, CU-UP, DU, DU/RU split) and SBA based core from 3GPP will be identified and their benefits assessed.

SBA as introduced in 5G CN is composed of services using REST APIs based on HTTP. The development of REST APIs is supported by a set of powerful tools (commonly used in the industry) supporting automatic API generation and verification. In SBA each service is provided by a service producer and can be consumed by one or more service consumers. Supporting SBA on interactions between RAN-core and among RAN nodes may simplify subsequent network enhancements and bring corresponding benefits as it is done in 5GC. 5G RAN uses peer-to-peer (p2p) signalling interface protocols (e.g., NGAP and XnAP) which are more difficult to enhance. For example, when adding new functionality in a system using p2p interfaces the existing network functions need to be enhanced and new p2p interfaces need to be introduced to support communication between existing and new network functions. SBA also simplifies load distribution by design. According to [23.501] clause 6.3 the dynamic load of the candidate NF instances can be used in the process of NF selection. We will study the feasibility to replace the existing NGAP protocol (N2 interface) and XnAP protocol (Xn-C interface) with a Service Based Interface (SBI) solution for enabling native cloud CU implementations.

One objective with a cloud-native RAN application architecture is to take full advantage of the flexibility of being cloud deployed and of the technology supporting it. Before cloudification an application may be resource constrained, however, in the cloud resources are “limitless” (but still not free) [AH21]. The penalty for this relieved resource constraint is latency. Externalizing states cost latency but is a prerequisite for elastic scaling, service specialization, and in-service upgrades without experiencing service degradation. Network injected latency is a consequence of using microservices to run the network applications.

In an optimized architecture, there is a balance with the network application utilizing the elasticity support of the cloud to provide the right functionality so that the latency is as low as possible, or at least as low as the network applications needs.

We see that, for instance, bundling together different layers or functionalities has driven several optimizations for service request procedures, e.g., low-latency handling. Separating these functionalities could potentially make the architecture cleaner, which also is cloud friendly, but it could also lead to more latency due to more handshakes among separate functionalities. So, in some cases functionality separation could lead to increased latency.

Some options for reduced latency for service requests handling could be to:

- Bundle more functionalities together, e.g., have a more single controller approach. It is not clear though that this always solves the problem since if we for instance bundle RRC / NAS we then instead separate the RRC from MAC,
- Another option for the service request would instead be to try to rely on the RRC_INACTIVE state and store the full RAN/CN configuration of the UE. In this way it would be possible to resume the connection faster even if functionality is more clearly separate.

An area where cloud friendliness and latency could possibly go together is when we signal via the CP for pure UP related modifications. Having the possibility to use “inband” or UP parallel CP could potentially speed up procedures. Dependencies between NFs may create long and complicated procedures, with more error cases, race conditions, flavours, etc. Further, dependencies that split responsibilities across different NFs can result in unnecessary complexity and make the system less future proof.

For this topic, we will study how we can streamline the 6G UP architecture, e.g., to which extent can we leverage virtualization in CP and UP. We want to identify and understand current dependencies between functions to make signalling more efficient (and reduce processing as well as the number of interactions and consequently save latency). Possible outcomes of this part of the task are reduced footprint by processing packets in fewer NF instances, benefits from common cloud deployment, possibilities for single vendor optimization and reduced complexity. Part of this process involves studying if all functions benefit from being in the cloud, and if not, how we characterize functions as suitable for cloud.

The following example demonstrates one way to make functionality more efficient, namely by removing the hierarchical dependencies. In 5G, signalling of a typical procedure involves the UE, gNB, and usually several NFs in the CN. The example in the signalling diagram in Figure 8-2 shows an Xn handover. The hierarchical dependencies RAN->AMF-SMF are clear, with many interactions between AMF/SMF and UPF.

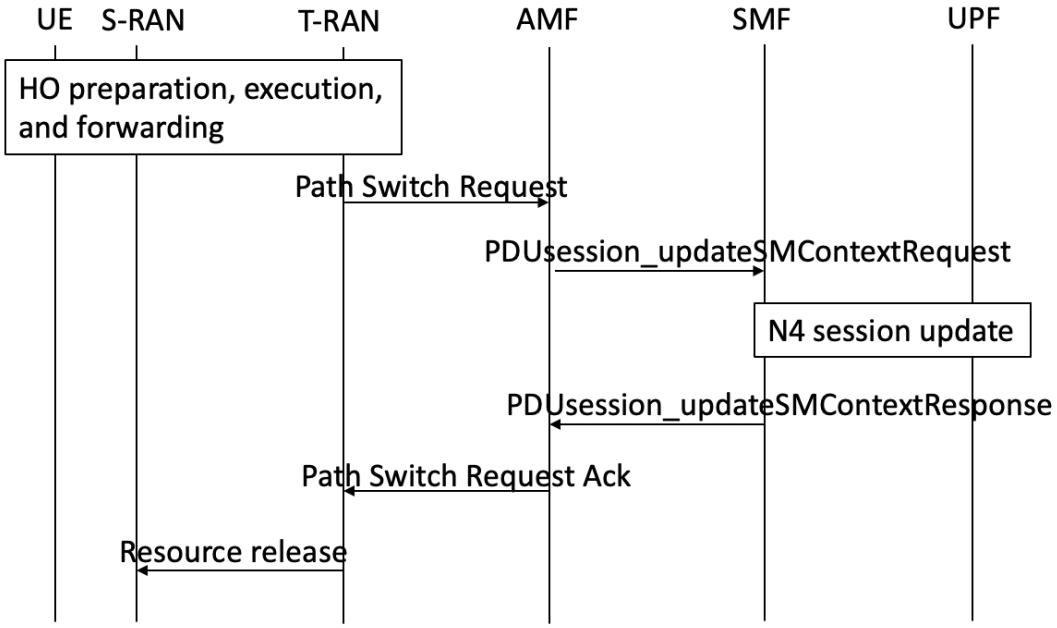

Figure 8-2 Rel-15 Xn handover- signalling flow

Figure 8-3 shows an example on how the current mobility process can be simplified leading to more efficient use of signalling resources. The idea here is that RAN directly notifies the SMF about mobility, since it is likely that there is RAN-SMF specific functionality/content. The AMF notifies the NFs requiring information on the event using URLs for AMF and SMF. These URLs are transferred from the source RAN (S-RAN) to target RAN (T-RAN) in step 1. T-RAN can then use the URLs when contacting AMF and SMF, and T-RAN can now directly request a path switch to SMF which then interacts with the UPF to update the N4 session.

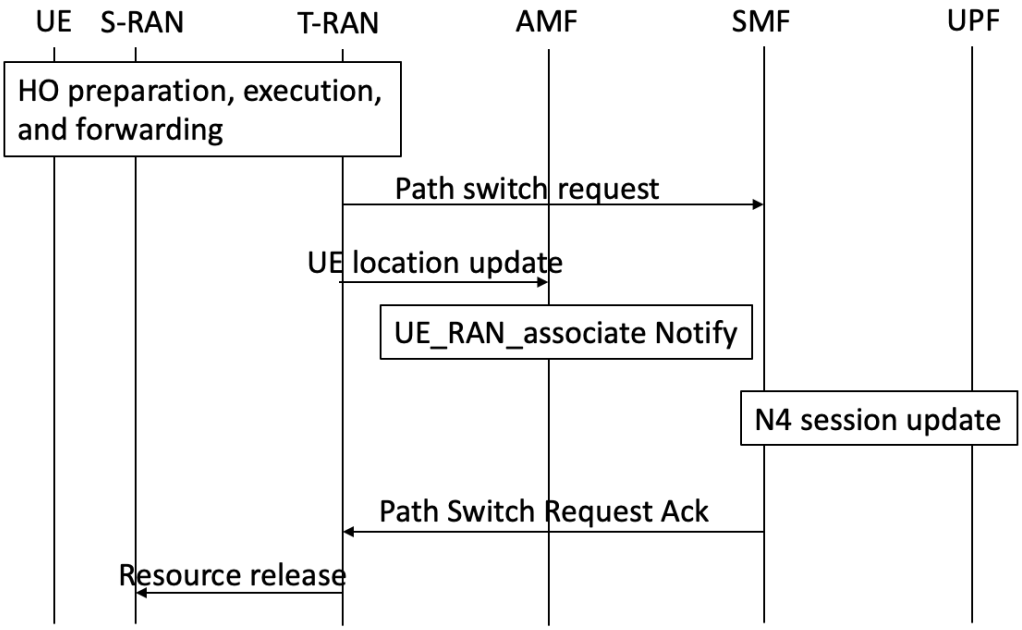

Figure 8-3 Simplified signalling for Xn mobility

There may be other ways to make signalling more efficient. Given that signalling directly or indirectly affects end-to-end latency this is an area that should be scrutinized. We aim to study to which extent L4-L7 mechanisms can be used to improve, e.g., session continuity for dual connectivity (DC).

The future networks will be more complex from a deployment perspective. This refers to that base stations can be anything from LEDs (for visual light communication), through micro and macro bases all the way to satellites. There may also be direct communication between devices. There may be services that move in patterns so that mobility is necessary between all kinds of base stations and devices. Although, it is more likely that future services will require very specific characteristics from the network and thus be limited to certain base stations. Voice, which will probably continue to be an important service, could be an example of the first type of service; one that can be handed off to any type of base station to ensure the possibility to remain in call while moving around. A service that will be limited to very few base stations may be the envisioned information shower, where a connected device can down/up-load with extremely high bitrates in certain locations. Whatever the service, future networks need to provide reliable and fast mobility to fulfil whatever requirements that a certain service has.

The research questions here are to find cost efficient solutions for high reliability mobility. It is likely, as discussed above, that there is not one single solution, but several solutions combined that are needed.

8.2 Initial TCO considerations for 6G

The key values of 6G comprise trustworthiness of the system, sustainability, limitless connectivity, and inclusiveness [D1.2]. At least two of them are closely related to cost, namely parts of sustainability and limitless connectivity, which may be a function of cost. Important building blocks needed for 6G are, e.g., large chunks of spectrum, methods to use this spectrum efficiently such as carrier aggregation, methods to handle sub-THz spectrum, (more than today) functionality at the edge of the network, etc. In this section we describe some of the hurdles, specifically those related to cost, that need to be overcome to deliver a 6G that delivers upon the key values.

A mobile operator's Total Cost of Ownership (TCO) for the introduction of a brand-new mobile system includes both capital expenses (one-time costs) and operating expenses (recurring costs), i.e., CapEx and OpEx, respectively [EFA+19]. Different operators might have significant local variations in cost structure, however the same relation in the relative cost structures is usually kept.

In a typical mobile network today, CapEx is 30% and OpEx is 70% of the TCO over a period of ten years; the RAN is the biggest cost in CapEx (approximately 50%) and, in decreasing order in terms of cost, there are transport, core network and operations & support. Site construction, radio equipment and spectrum acquisition costs account for most of the RAN CapEx. The RAN is also the biggest component in OpEx (approximately 65%), which is mainly due to site rental, the power consumed to maintain radio sites fully operational as well as operations & maintenance (including vendor support) [Gha20].

Several means and innovations can be exploited to improve the cost structure of an operator's network, hence positively impacting both CapEx and OpEx; for the RAN, the CapEx impact can be reduced as follows [Gha20] [DEL21]:

- by reducing site construction costs (representing ~30% RAN CapEx), e.g., via site sharing among several operators or by leasing from tower companies,
- by introducing innovations in radio equipment (accounting for ~40% RAN CapEx), e.g., innovative RAN deployment options.

On the other hand, for the RAN OpEx, its impact can be minimized as follows [Gha20]:

- by lowering the site rental costs (representing ~30% RAN OpEx) – which typically include the backhaul costs – via e.g., reduced equipment footprint (i.e., smaller size and weight) or by exploiting initiatives from the municipalities which could allow providing antennas on lamp posts;
- by optimizing the Operations (accounting for ~30% RAN OpEx), which typically include laboratory testing, troubleshooting, optimization, etc., by means of automation, AI and ML, virtualization and softwarization; also, the Continuous Integration/Continuous Delivery (CI/CD) approach could also help in keeping costs down;
- by reducing the network power consumption (representing ~15% RAN OpEx) [GSMAEE], e.g., via intelligent switch-off of radio equipment during non-business hours or by developing/selecting innovative power-efficient technologies for 6G networks, considering a trade-off between flexibility and cost/bit;
- by minimizing the manpower (accounting for ~10% RAN OpEx) – which typically includes field testing, drive testing and site maintenance, etc. – via e.g., enhanced AI-based SON mechanisms.

As mentioned above, 6G is expected to be an enabler in ensuring sustainability (for communication) on a global society level. For these reasons, sustainability is investigated under topics relevant to ICT and 6G, e.g., the Sustainable Development Goals by the United Nations.

Increased traffic volumes (in future cellular networks) are likely to lead to higher energy consumption. Some of this increase can be mitigated by more efficient hardware. However, even if more efficient hardware can mitigate the increased energy usage needed for wider bandwidth, the use of higher frequencies will demand a denser network where each gNB consumes energy, which will affect the total energy consumption in the network. A key challenge for 6G will be to break the energy curve, i.e., to suppress energy consumption with increasing traffic. In addition to KPIs for energy consumption per data bit there will be KPIs for circular resource handling.

With 6G we will attempt to fulfil the sustainability targets primarily in 3 different areas: 1) Industry, Innovation, and Infrastructure, 2) Quality education and Decent Work as well as 3) Economic Growth. The targets will be satisfied by key 6G features such as comprehensive coverage, affordable infrastructure, more powerful smart devices, and many AI-driven functions ensuring high quality, sustainable, and resilient infrastructure.

In order to define a TCO for a 6G network, a baseline architecture needs to be identified, with respect to which the most promising network features/elements characterizing the "new" architecture are evaluated in terms of cost benefits when deployed. In this sense, the TCO is to be evaluated in relative terms (i.e., *x*% cost savings) with respect to the baseline architecture. As in Hexa-X the focus is on 6G and considering that in 5G there are multiple architecture flavours, the baseline architecture could be the 5G NR SA architecture, where both 5GC and NG-RAN are considered and corresponding TCO studies are already available, e.g. [Gha20a]. Then, based on the outcome of the WP5 tasks about the identification of the most significant and disruptive architectural enablers, how such enablers affect the (relative) TCO of the 6G network can be evaluated and results of such analysis will be provided in a future WP5 deliverable.

Examples of features that are likely to be investigated to reduce TCO for 6G are O-RAN, and novel antenna systems, such as massive MIMO or the use of a single antenna handling multiple frequency bands and radio access technologies. Also, automation, e.g., using AI, is by many seen to have very large potential for reducing OpEx [DEL21].

8.3 Methods for enabling SBA in 6G

The evolution of cloudification is likely to continue for 6G. A key challenge is to design a future 6G architecture, which is able to fully utilize the cloud platform with regards to speed of development, reuse of common cloud components, balancing the need to standardize critical business interfaces with the fast evolution of IT tools, such as the ones described below. Similar to the development of SBA and clouds, the whole architecture and its functions are designed based on reusable and composable microservices that will enable dynamic workload scheduling to optimal execution points in a hierarchy of data centres across the network, matching the latency and scalability needs.

The objective is to study the software and service-based features and their relevance to be integrated in an "as-a-Service" network, an Edge architectural model for the associated VNFs. Firstly, a landscape of component-based models issued from software development technologies is presented. Secondly, the strengths and shortcomings of each model are analysed. The analysis also looks at associated management approaches to show the applicability to Edge and distributed network architectures.

In Component-Based Software Engineering (CBSE), component models offer a structured programming paradigm that allows developers to reuse software components. A component is a software module offering predefined services, capable of communicating with other components. By enforcing a strict separation between interface and implementation and by making software architecture explicit, component-based programming can facilitate the implementation and maintenance of complex software systems.

The functionality that a component provides is defined by its dependencies via interfaces in a structured, usually hierarchical, form. This structure makes component models good candidates for the architecture design of an NFV network service and consequently the composed VNFs. Network as a Service (NaaS) strongly relies on user-centric service models. Web and Cloud service models and applications for example are based on CBSE models.

Let us then analyse component models from different initiatives, focusing on models that achieve user-centric services:

1. Fractal Component Model
2. Grid Component Model (GCM)
3. Service-Oriented Architecture (SOA)
4. Self-Controlled service Component Model (SCC)

These models are explained in more detailed below.

1) **Fractal Component Model**: Fractal is a modular and extensible component model that can be used to design, implement, deploy, configure, and manage complex software systems and applications. Fractal objects have a regular invariant structure even at different scales. This scalability and elasticity feature of Fractal is important for distributed systems. In Fractal, dynamicity aspects are presented through the possibility to add or remove components to or from an already deployed application thanks to reconfiguration of the bindings between components (at run-time scaling) and to the introspection of composite components.

Fractal defines three types of interfaces:

- Usage interfaces: client and server interfaces for functional bindings, where the component defines what it needs and provides,
- Control interfaces: controllers or membrane, useful to control and manage the component life cycle and provides methods to start or stop the component,
- Management interfaces: also, through the membrane for component configuration.

In addition to that, Fractal defines an Architecture Description Language (ADL). The ADL uses an XML syntax and is a way to describe a component-based system without having to worry about the implementation code.

2) **Grid Component Model (GCM)** was proposed to extend Fractal towards autonomous distributed systems. Specified by ETSI [102828], it defines well-structured non-functional aspects defined with all necessary management elements and interfaces in the membrane. Thus, non-functional bindings or compositions are possible similarly to the functional bindings in Fractal. GCM architecture strength resides in this separation of concerns in designing separately:

- Business Content: functional aspects of a primitive component responsible for business logic, and
- Membrane: responsible for non-functional aspects.

In GCM, the membrane structure is a strong asset for component management and control. For the functional aspects, communication is performed on interfaces and follows component bindings. GCM components management approach uses an autonomic approach in hierarchical architectures implementing Monitoring-Analysis-Planning-Execution (MAPE) loop to manage this hierarchy where Execution is used to change the bindings. However, the main shortcoming of GCM is this hierarchical nature of the bindings between components which adds functional coupling and makes primitive components hard to be composed. Bindings need to be less tight to have a dynamic composition. SOA has addressed this rigidity in bindings as we present hereafter.

3) **Service-Oriented Architecture**: Quite differently, SOA [ZSP+13] is an architectural model, where a component is a service and is the unit of work done by a service provider to achieve a desired service for a service consumer. SOA provides a composition model with loose coupling and reduced functional coupling among interacting service components. These features are a must for building customized services. The composition consists in building a global service composed of a set of elementary service components. Flexible composition is customizable by adding, replacing, removing service components based on user needs. SOA proposes a structure called Service Component Architecture (SCA), and a Web Services (WS). The SCA [Edw11] is a component model adapted to SOA features and enables creation of service components and modelling of service compositions. The strong features of SCA are re-usability, reference points and loose coupling. The WS implement SOA requirements and thus support flexible composition and methods for service integration in Cloud applications. The strong asset of WS is the Client-Provider relationship based on service publication and description in catalogues by the service provider, and service discovery and invocation by the client. To achieve that, web services rely on the WS Description Language (WSDL), Universal Description Discovery and Integration (uDDI), EXtensible Markup Language (XML) and APIs such as Simple Object Access Protocol (SOAP) and Representational State Transfer (REST).

However, in practice the SCA model is not dynamic, and it is not possible to add, replace or remove service components or bindings dynamically (at run-time). Also, web services do not

integrate the notion of contract or Service Level Agreement (SLA) and do not support QoS management as GCM does.

4) **Self-Controlled service Component (SCC):** One model that may cope with the shortcomings of SOA implementations exposed above is the SC model, which combines the features of GCM and SOA. SCC has adopted GCM and to a certain extent the MAPE loop. SCC has designed components based on initial SOA features to inherit the ability to achieve dynamics at run-time service composition. Further, the SCC model has defined new service component features beyond the ones advocated by existing models. SCC introduced dynamicity in handling components through auto-control approach for the management of non-functional aspects (QoS) in flat architectures enabled by the membrane structure.

8.4 Compute as a-Service

Compute-as-a-Service (CaaS) is a use case enabling service approach, as described in [D1.2], "*is aimed to be used by any devices (static or mobile, IoT, handhelds, etc.) or network infrastructure equipment that choose to delegate demanding, resource-intensive processing tasks to other parts of the network providing more powerful compute nodes, which are also of higher availability at the time of workload generation. These service-offering compute nodes can be either onboard other devices or, for example, edge cloud servers at the infrastructure side*". In the CaaS case, external compute resources can be made available to a specific entity or user device through a well-defined open interface. The basic principle is indicated in the signal flow chart of Figure 8-4 below.

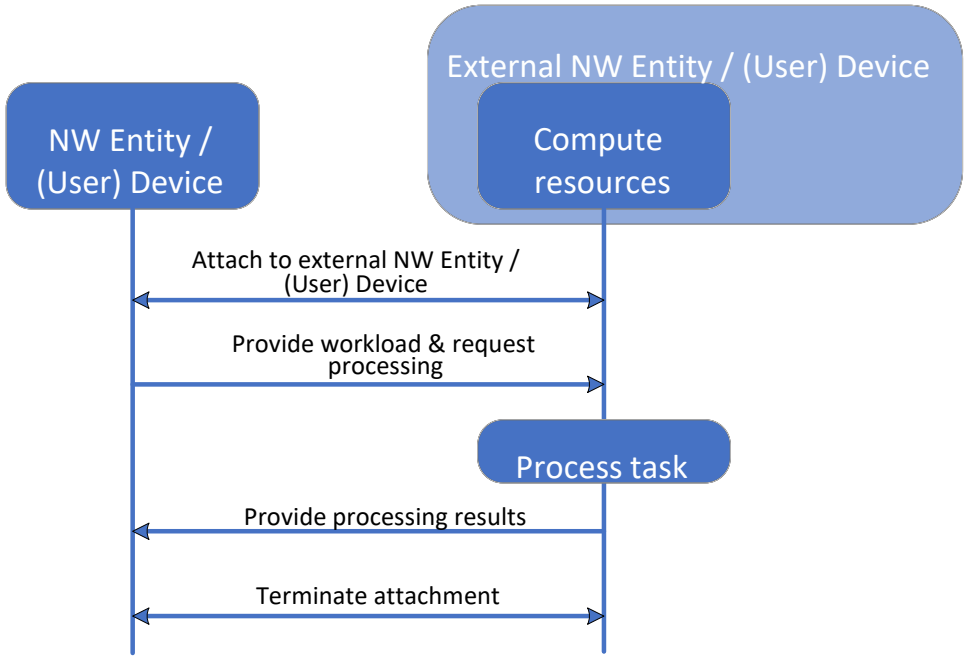

Figure 8-4: Signal flow chart illustrating CaaS case - it is assumed that discovery and selection of a network entity containing the needed computational resources has been performed.

The basic principles relate to an offload of processing tasks to external compute resources. In this context, some of the needed features to be defined, as part of a 6G network architecture design, are the following: (i) a general interface providing access to external computational resources; (ii) mechanisms for discovery/ detection of available compute resources (e.g. via a general register reachable by the CaaS provider); (iii) a functional entity (e.g., central controller/ workload orchestrator) that will make the decision upon to offload (fully or partly) a processing workload, based on the available resources of network nodes and taking into account requirements, such as

the incurred latency from task generation to output acquisition by the CaaS consumer, the energy footprint of the task delegation and the trustworthiness of the workload hosting entity. Of course, it could also be the case that workload delegation is performed on an opportunistic and ad-hoc fashion. The latter option may be specifically challenging in a dense network environment with distributed computing capabilities.

A key challenge in the CaaS concept is to balance the following -often contradicting- objectives: on the one hand, the user should be enabled to be given access to computational resources, which often consist of heterogeneous elements and architectures (e.g., CPUs, FPGAs, etc.) and which would, therefore, need to be utilized at the maximum possible level of efficiency; on the other hand, the user should ideally rely on generic and simple interfaces that are abstracted from the underlying hardware and software (such as Operating Systems, etc.) to the extent possible. As a way forward, it is proposed to offer various trade-offs between flexibility, simplicity, and efficiency to users to choose from, depending on the underlying scenario. Three CaaS approaches -which will be studied and assessed in forthcoming deliverables- can be thought of; these are the following:

1) Applications customised to assist with the processing of a specific workload (e.g., signal processing, object recognition etc.) can be pre-installed by a service provider. In the simplest case, a user accesses a pre-installed and pre-configured application offered by a service provider. The involved trade-off consists in the high level of simplicity and efficiency, versus a possible low level of flexibility.
2) Full access to “raw” hardware and software resources is provided to the service calling entity (e.g., device). An expert user is expected to develop code, which is compiled and optimised for a specific target platform. The user is required to have deep knowledge on the platform architecture, expert software development skills and will need to re-design any applications if they should operate on different platform architectures. The key direction beyond 5G lies in the definition of a related general interface, which is available to consumers internal as well as external of the network. This interface should be able to support various programming languages and platform architectures, hence, the incurred trade-off lies in the high level of flexibility and efficiency offered, at the cost of low level of simplicity.
3) Compute Virtual Machine (CVM) based approach: as a compromise between the previous two approaches, the novel proposal is to use a “Compute Virtual Machine (CVM)” approach building on an extension of the “Radio Virtual Machine (RVM)”, which is defined as an Abstract Machine capable of executing Configcodes and it is independent of the hardware [EN303146-4]. The basic virtual platform is illustrated in Figure 8-5. Any application code is first processed by a front-end compilation step, which creates so-called *Configcodes* for RVMs. In a second step, a back-end compilation is performed to map the application to a specific (heterogeneous) target platform consisting of a specific number of resources (such as CPUs, FPGAs, memory, etc.). The process is supported by a library, which provides optimized functionalities to the developers. The Virtual Machine approach mainly consists of Data Objects (DO), which are connected to Abstract Processing Elements (APEs) through an Abstract Switch Fabric. A Control Unit interconnects the building blocks, as required.

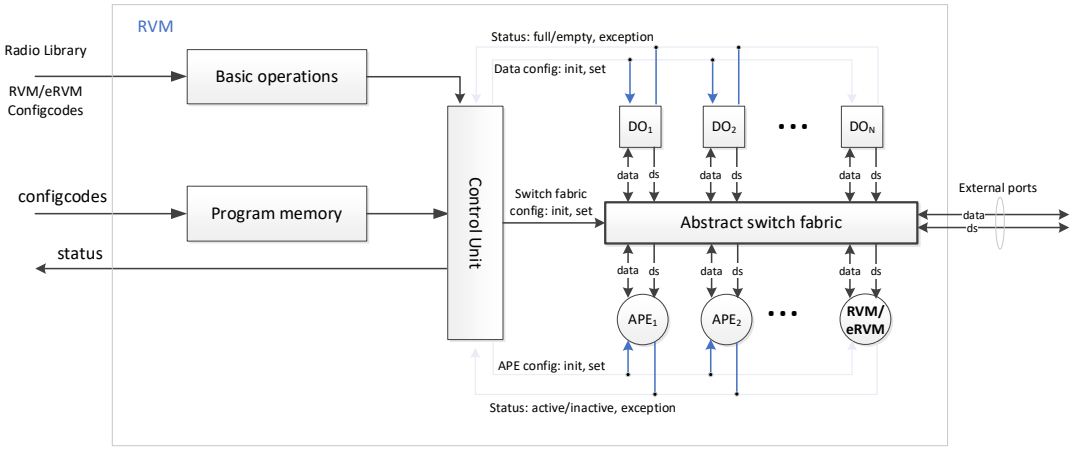

Figure 8-5: Radio Virtual Machine (RVM) [EN303146-4].

The proposed approach enables a developer to create code optimized for a single (virtual) platform, which is then back-end compiled to be executed on any target platform. Thus, a maximum level of flexibility and code portability is offered, while maintaining a high level of efficiency. The incurred trade-off consists in mid/high level of flexibility and efficiency, versus a moderate level of simplicity. It should be noted that the CaaS approach may be considered to include the AIaaS concept as a special case, where the computational task to be addressed is a learning (e.g., updating an ML model) or inferencing one, relating to either automated network operation (e.g., network security enhancement based on anomaly detection) or to any end user application (e.g., image classification). On the other hand, AI/ML can be used to perform workload delegation, as e.g., ML models may be proven useful to provide in-advance predictions about the availability of compute nodes, their level of trust, incurred latencies, and the expected level of network energy efficiency when a given offloading decision is taken (see [D4.1] for further details from an algorithm/ methodology perspective).

9 KPIs for the 6G architecture

The KPIs for the 6G architecture are derived based on the:

- D1.2 Use cases (see Section 2.1)
- D1.2 KPIs (see Section 2.2)
- Architectural transformation (see Section 5)
- KPIs from each WP5 Task (WP5 (i.e. the Intelligent networks in Chapter 6, Flexible networks in Chapter 7 and Efficient networks in Chapter 8)

Note that the target requirements for the Architecture KPIs will be defined in the next deliverable, D5.2.

The proposed KPI from Task 5.2 (Chapter 6) are the following:

- **Convergence time to adopt the network to changes**: This is combined KPI of the change of coverage, scaling up and down the number of sessions for the extreme experience services.
- **AI communication and computing overhead**: The amount of additional computing and communication resources allocated to optimize end to end QoS in comparison to acceptable but static resource allocation.

The proposed KPI from Task 5.3 (Chapter 7) are the following:

- **Reliability/robustness for network of networks**: How well the network connects to the best sub-network to minimize the radio link failure time, while maximizing QoS (even if some QoS requirement may not be reached).
- **Network flexibility**: The ability that the network architecture must perform well over a wide range of deployments and network states, i.e., an average of selected KPIs over several different deployments.

The proposed KPI from Task 5.4 (Chapter 8) are the following:

- **Separation of concerns of network functions** has to do with how many dependencies that a new function has. A function with a lot of dependencies is generally a less good function. When adding new functions, effort should be put in the process to try to separate the functions as much as possible. There must be clear division of responsibility, especially in multi-vendor networks. For example, situations where many nodes need access to UE context like the multiple solutions for handling IoT where some are CN and some RAN. In such situations it is not clear who "owns" the responsibility.
- **Ease of adding new functions in future:** can be measured as number of specifications that need to be updated. Another way to measure this is to see how many other NFs need to be changed when adding a new NF. However, this is also related to the amount of separation mentioned in the previous measure
- **TCO reduction,** is the TCO for additional use case specific features relative to the baseline architecture.

10 Proof of Concepts

This Chapter describes the proof of concepts (PoCs) in WP5.

10.1 Flexible topologies (FLEX-TOP) for efficient network expansion

The demo will deliver insights on the efficiency of the flexible topologies concept. The key benefits will be coverage extensions (in line with challenge: Global Service Coverage), service provision with lower latencies (since local structure is terminated at the infrastructure edge), security (engagement of selected devices, in line with challenge Trustworthiness), and lower energy consumption (at infrastructure, challenge: Sustainability). There will be leverage on mesh/ad-hoc/D2D networking, disaggregated devices with the ability to flexibly allocate functionality (management of computing resources), ultra-high spectrum, and on coordination with the infrastructure, e.g., in terms of resources to use. Architecture aspects and data flow issues will also be studied and validated.

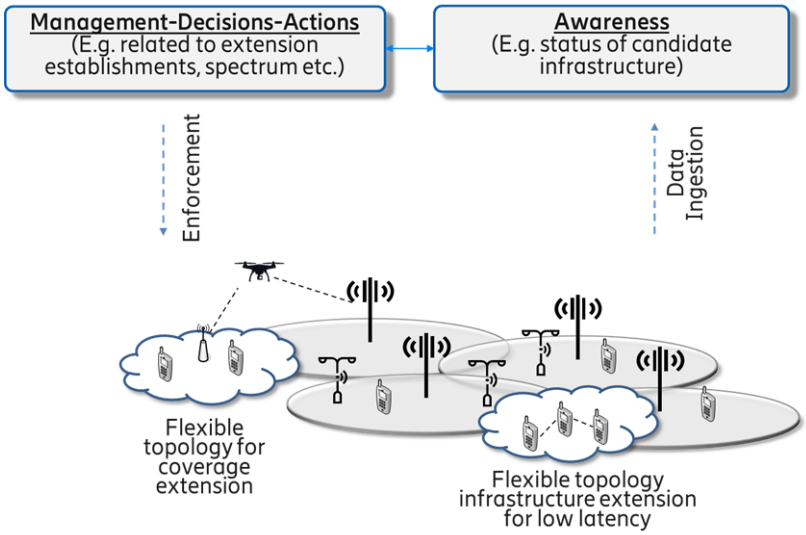

Figure 10-1. Flexible topologies for efficient infrastructure extensions demonstration

10.2 FED-XAI - FEDerated XAI demo

The objective of this Proof-of-Concept (jointly carried out with WP4 by UPI, INT and TIM) is to demonstrate a framework for the FEDerated learning of eXplainable AI (FED-XAI) models, including both algorithms and the related signalling. The FED-XAI demo will focus on a general V2X use case, where collecting a high volume of contextual and sensor data from traffic participants and road infrastructure will be common practice.

This demo will consider an application where UEs (i.e., vehicles) connected to a base station receive a video stream the quality of which plays a decisive role to the safety of remote or assisted driving. With reference to Figure 10-2, this can be mapped to a see-through use case, where the receiving car uses live feed from another car's camera (e.g., to make overtaking safer in the presence of visual impairments for the driver). The operation may be also supported by a digital twin residing at the network edge. The objective is to employ XAI models, learnt (and updated) in a federated fashion exploiting aggregated QoS/QoE data shared by the UEs, to predict the QoE perceived by UEs in the near future.

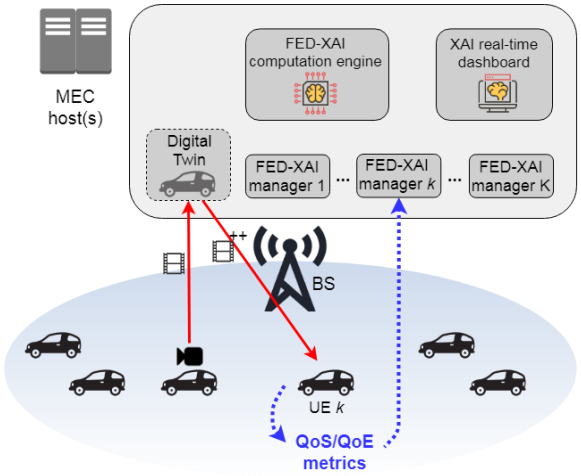

Figure 10-2. UEs collecting QoS/QoE metrics

This is achieved by allowing the UE to send collected QoS/QoE metrics to a FED-XAI Manager (FM). The latter can be either a function residing at the UE or in the cloud/edge system of the network. Each UE communicates with its dedicated FM, and data isolation between FMs of different UEs must be ensured. FMs will then feed a FED-XAI Computation Engine (FCE) with aggregated, privacy-preserving versions of QoS/QoE metrics (Figure 10-3, step 1). The FCE is a function instantiated within an edge/cloud node and leverages aggregated data from FMs to build or update the QoE prediction model, which are then provided back to the FMs (Figure 10-3, step 2) that will use it to perform the QoE prediction for their corresponding UE. The results of the prediction will feed a dashboard (Figure 10-3, step 3) that displays them in real time and explains how they were obtained.

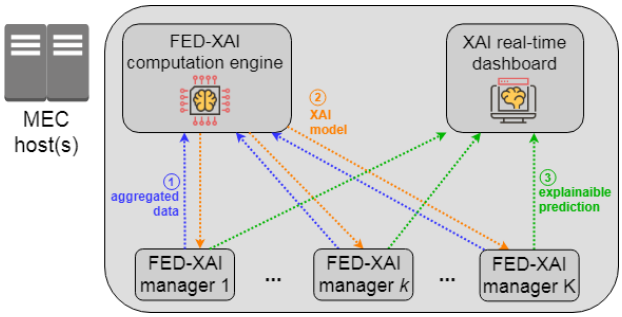

Figure 10-3. FED-XAI managers sharing data aggregates in order to build/ update the XAI model; prediction explanations are shown on a dashboard in real time.

The above scenario will be implemented in a real-time distributed testbed. The communication between real end-devices will be realized by Simu5G, a modular simulator of 3GPP-compliant New Radio based on OMNeT++ [NSS+20] that can work in real time and interfaces with external, real devices and applications [NSV+20], allowing real network packets generated by the latter to be transported through the simulated mobile network. The simulator will be fed with data coming from a MNO's live radio access network in order to build more realistic simulation scenarios. QoS/QoE data will be taken in real time from Simu5G and transmitted to the FMs implemented in the edge subsystem, which hosts the AI-related aspects of the testbed. More details about the implementation of the testbed will be provided in D5.2, possibly including some preliminary results.

11 Summary and Conclusions

The first objective is to perform a gap analysis of current architecture and thereafter establish the general architectural direction for a possible 6G architecture. This includes identifying the necessary enablers (e.g., functions, algorithms, and enhancements) to support the new 6G architecture. The second objective is related to the novel architectural enablers and how these enablers shall support the future use cases and the requirements. Thereafter, we describe the initial scope of the novel architectural enablers. These enablers allow an intelligent distributed network, new flexible network topologies and enables an efficient deployment of future networks.

For the general direction of the 6G architecture, we have developed an initial layered architecture of the different architecture functions. The layered architecture is used to group where different architecture functionalities will be placed and understand how they are related to each other. In addition to this we have also defined eight different architectural principles. These principles are used to guide the work in the different Hexa-X tasks. The principles are:

1. **Exposure of capabilities:** Expose new and existing network capabilities to end-to-end applications, by exposure of capabilities to the applications we may enhance the by providing enhanced capabilities for features.
2. **Designed for (closed loop) automation (and AI):** support full automation, utilizing distributed AI/ML agents to manage and optimize the network without human interaction.
3. **Flexibility to different topologies:** The ability of the network to adapt to various scenarios and deployments without loss of performance and easy deployment.
4. **Scalability:** support of very small to very large-scale deployments, by scaling up and down network resources based on needs
5. **Resilience and dependability:** resilient in terms of service and infrastructure provisioning using multi connectivity, and separation of CP and UP, support of local network survivability.
6. **Exposed interfaces are service based:** Network interfaces should be designed to be cloud native, utilizing state-of-the-art cloud platforms and IT tools in a coherent and consistent manner.
7. **Separation of concerns of network functions:** The network functions have bounded context and all minimal dependency with other network functions, so that network functions can be developed and replaced independently from each other.
8. **Network simplification in comparison to previous generations:** reduce complexity utilising cloud-native RAN and CN functions with fewer (well-motivated) parameters to configure and fewer external interfaces.

The novel architectural enablers are addressed in three tasks, i.e., Task 5.2 Intelligent networks, Task 5.3 Flexible networks, and Task 5.4 Efficient networks. The proposed enablers for Intelligent Networks of Hexa-X are meant to facilitate dynamic adaptability of the network architecture to accommodate new use cases and deployment scenarios beyond what the current cellular networks could offer, while keeping the infrastructure and energy costs at acceptable and sustainable levels. Flexible networks intend to enable extreme performance and global service coverage. This is achieved by developing enablers that can manage local ad hoc networks and distributing their functionalities between them and at the edge. The Intelligent Networks will streamline the interfaces assuming a cloud-native RAN and CN, by function separation and clarify each network functions' responsibility and functionality. Finally, to lay a foundation for the coming deliverables, the research scope and problems of the architectural enablers are defined.

12 References

[23.288] 3GPP TS 23.288, "Architecture enhancements for 5G System (5GS) to support network data analytics services (Release 17)", v17.2.0, Sep. 2020.

[23.303] 3GPP, 23.303, “Proximity-based services (ProSe); Stage 2”, Version 17.0.0

[23.501] 3GPP TS 23.501, "System architecture for the 5G System (5GS); Stage 2 (Release 17)", v17.2.0, Sep. 2021.

[23.502] 3GPP TS 23.502, "Procedures for the 5G System (5GS)”, v17.2.1, Sep. 2021.

[23.503] 3GPP TS 23.503, "Policy and charging control framework for the 5G System (5GS); ”Stage 2", v17.2.0, Sep. 2021.

[23.700-91] 3GPP TR 23.700-91, "Study on enablers for network automation for the 5G System (5GS); Phase 2 (Release 17)", v17.0.0, Dec. 2020.

[24.193] 3GPP TS 24.193 Access Traffic Steering, Switching and Splitting (ATSSS)

[28.533] 3GPP TS 28.533, "Management and orchestration; Architecture framework (Release 17)", v17.0.0, Sep. 2021.

[28.552] 3GPP TS 28.554, "Management and orchestration; 5G performance measurements (Release 17)", v17.4.0, Sep. 2021.

[28.554] 3GPP TS 28.554, "Management and orchestration; 5G end to end Key Performance Indicators (KPI) (Release 17)", v17.4.0, Sep. 2021.

[28.809] 3GPP TR Study on enhancement of Management Data Analytics (MDA) (Release 17), March 2021.

[36.300] 3GPP TS 36.300, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)”, V16.6.0 (2021-06)

[36.902] 3GPP, “LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (3GPP TR 36.902 version 9.3.1 Release 9)”, ETSI TR 136 902 V9.3.1 (2011-05), Online available: https://www.etsi.org/deliver/etsi_tr/136900_136999/136902/09.03.01_60/tr_136902v090301p.pdf

[37.340] 3GPP TS 37.340 “Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity”, V16.2.0 (2020-07)

[38.300] 3GPP TS 38.300 “NR and NG-RAN Overall description; Stage-2”, § 9.2.3.1, Rel-16

[38.331] 3GPP TS 38.331 “NR; Radio Resource Control (RRC); Protocol specification”, Rel-16, Version 16.6.0, Sep. 2021

[38.401] 3GPP TS 38.401 “NG-RAN; Architecture description”, Rel-15

[38.807] 3GPP Rel-16 TR–38.807 – “Study on requirements for NR beyond 52.6 GHz”

[38.801] 3GPP Rel-14 TR 38.801 “Study on new radio access technology: Radio access architecture and interfaces”

[802.11bb] “IEEE 802.11bb Standard,” [Online]. Available: http://www.ieee802.org/11/Reports/tgbb_update.htm.

[802.11s] IEEE 802.11s, Overview of the Amendment for Wireless Local Area Mesh Networking, Online: https://www.ieee802.org/802_tutorials/06-November/802.11s_Tutorial_r5.pdf

[102828] ETSI TS 102 828 GRID; Grid Component Model (GCM); GCM Application Description

[3g4garch] https://blog.3g4g.co.uk/2021/07/different-types-of-ran-architectures.html

[3g4ghist] https://blog.3g4g.co.uk/2018/02/tutorial-service-based-architecture-sba.html

[5G6GNTT] 5G Evolution and 6G, NTT DOCOMO, [link]

[5GSA] 5G Standalone 2021 – Summary, March 23, 202. https://cell-corner.com/5g-standalone-2021-summary/

[5GSMA] Road to 5G: Introduction and Migration, GSMA, Monday 23 Apr 2018. https://www.gsma.com/futurenetworks/resources/road-to-5g-introduction-and-migration-whitepaper/

[6GSam20] 6G The Next Hyper-Connected Experience for All, Samsung, on July 14, 2020. https://news.samsung.com/global/samsungs-6g-white-paper-lays-out-the-companys-vision-for-the-next-generation-of-communications-technology

[AB18] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," in IEEE Access, vol. 6, pp. 52138-52160, 2018.

[ABG+21] S. T. Arzo, R. Bassoli, F. Granelli and F. H. P. Fitzek, "Multi-Agent Based Autonomic Network Management Architecture," in IEEE Transactions on Network and Service Management, doi: 10.1109/TNSM.2021.3059752.

[AH21] Anderson, T. (2021). VC's paper claims cost of cloud is twice as much as running on-premises. Let's have a look at that. https://www.theregister.com/2021/06/02/andressen_horowitz_paper/

[Air] Airflow, https://airflow.apache.org/

[AK19] I. F. Akyildiz and A. Kak, "The Internet of Space Things/CubeSats," in IEEE Network, vol. 33, no. 5, pp. 212-218, Sept.-Oct. 2019, doi: 10.1109/MNET.2019.1800445.

[AMA19] Alsamhi, S.H., Ma, O. and Ansari, M.S. Survey on artificial intelligence based techniques for emerging robotic communication. Telecommun Syst 72, 483–503 (2019). https://doi.org/10.1007/s11235-019-00561-z

[ApOp] Apache OpenWhisk, https://openwhisk.apache.org/

[ADS+20] A. B. Arrieta, , N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera,"Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion, Vol. 58 (2020): 82-115.

[ARS16] M. Agiwal, A. Roy, and N. Saxena, “Next Generation 5G Wireless Networks: A Comprehensive Survey,” IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1617–1655, 2016.

[ASR+20] S. Ali, W. Saad, N. Rajatheva, K. Chang, D. Steinbach, B. Sliwa, C. Wietfeld, K. Mei, H. Shiri, H-J. Zepernick, T. M. Chinh Chu, I. Ahmad, J. Huusko, J. Suutala, S. Bhadauria, V. Bhatia, R. Mitra, S. Amuru, R. Abbas, B. Shao, M. Capobianco, G. Yu, M. Claes, T. Karvonen, M. Chen, M. Girnyk, and H. Malik. 6G White

Paper on Machine Learning in Wireless Communication Networks. 2020, arXiv:2004.13875v1

[AWS] AWS Lambda, https://aws.amazon.com/lambda

[BBG+20] S. Bonafini, R. Bassoli, F. Granelli, F. H. P. Fitzek and C. Sacchi, "Virtual Baseband Unit Splitting Exploiting Small Satellite Platforms," 2020 IEEE Aerospace Conference, Big Sky, MT, USA, 2020, pp. 1-14, doi: 10.1109/AERO47225.2020.9172316.

[BCH+17] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini and J. Widmer, "A Survey of Anticipatory Mobile Networking: Context-Based Classification, Prediction Methodologies, and Optimization Techniques," in IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1790-1821, third quarter 2017, doi: 10.1109/COMST.2017.2694140.

[BCZ97] Bhattacharjee S., Calvert K.L., Zegura E.W. (1997) An Architecture for Active Networking. In: Tantawy A. (eds) High Performance Networking VII. HPN 1997. IFIP — The International Federation for Information Processing. Springer, Boston.

[BDG+14] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese, and D. Walker. "P4: Programming protocol-independent packet processors." ACM SIGCOMM Computer Communication Review 44, no. 3 (2014): 87-95.

[BFS21] R. Bassoli, F. H.P. Fitzek, and E. Calvanese Strinati, “Why do we need 6G?”, ITU Journal on Future and Evolving Technologies, 2 (6), 2021.

[BGS+20] R. Bassoli, F. Granelli, C. Sacchi, S. Bonafini and F. H. P. Fitzek, "CubeSat-Based 5G Cloud Radio Access Networks: A Novel Paradigm for On-Demand Anytime/Anywhere Connectivity," in IEEE Vehicular Technology Magazine, vol. 15, no. 2, pp. 39-47, June 2020, doi: 10.1109/MVT.2020.2979056.

[BR18] R. Bifulco and G. Rétvári, "A survey on the programmable data plane: Abstractions, architectures, and open problems." 2018 IEEE 19th International Conference on High Performance Switching and Routing (HPSR). IEEE, 2018.

[BT20] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions." IEEE Network 34.2 (2020): 186-194.

[CC14] W. Cerroni and F. Callegati, "Live migration of virtual network functions in cloud-based edge networks," 2014 IEEE International Conference on Communications (ICC), 2014, pp. 2963-2968, doi: 10.1109/ICC.2014.6883775

[CCD08] Cunningham P., Cord M., Delany S.J. (2008) Supervised Learning. In: Cord M., Cunningham P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2.

[CDY+12] P. Cheng; L. Deng, H. Yu, Y. Xu, and H. Wang “Resource Allocation for Cognitive Networks with D2D Communication: An Evolutionary Approach,” IEEE Wireless Commun. and Net. Conf. 2012, Shanghai, China, Apr. 2012, pp. 2671–76

[CGG+20] D. Camps-Mur, M. Ghoraishi, J. Gutierrez, J. Ordonez-Lucena, T. Cogalan, H. Haas, A. Garcia, V. Sark, E. Aumayr, S. van der Meer, S. Yan, A. Mourad, O. Adamuz-Hinojosa, J. Pe´rez-Romero, M. Granda, and R. Bian.,“5G-CLARITY: Integrating 5GNR, WiFi and LiFi in Private 5G Networks with Slicing Support”, June, 2020: European COnference on Networks and Communications, EuCNC

2020, https://www.5gclarity.com/wp-content/uploads/2021/03/5G-CLARITY_EuCNC20_Final.pdf

[CHI+18] Chowdhury M. Z., Hossan M. T., Islam A., and Jang Y. M., "A comparative survey of optical wireless technologies: Architectures and applications," IEEE Access, vol. 6, pp. 9819–9840, Jan. 2018, https://arxiv.org/ftp/arxiv/papers/1810/1810.02594.pdf

[CNCF-21] Cloud Native Computing Foundation, https://www.cncf.io/

[CONS] Consul https://www.consul.io/docs/connect

[D1.1] Hexa-X Deliverable D1.1, "6G Vision, use cases and key societal values", Online: https://hexa-x.eu/wp-content/uploads/2021/02/Hexa-X-D1.1_v1.0.pdf

[D1.2] Hexa-X Deliverable D1.2, "Expanded 6G vision, use cases and societal values", Online: https://hexa-x.eu/wp-content/uploads/2021/05/Hexa-X-D1.2_v1.0.pdf

[D2.1] Hexa-X Deliverable D2.1 "Towards Tbps Communications in 6G: Use Cases and Gap Analysis", https://hexa-x.eu/wp-content/uploads/2021/06/Hexa-X_D2.1.pdf

[D4.1] Hexa-X Deliverable D4.1, "AI-driven communication & computation co-design: Gap analysis and blueprint". Online: https://hexa-x.eu/wp-content/uploads/2021/09/Hexa-X-D4.1_v1.0.pdf

[D6.1] Hexa-X Deliverable D6.1, "Gaps, features and enablers for B5G/6G service management and orchestration" Online: https://hexa-x.eu/wp-content/uploads/2021/06/Hexa-X-D6.1_v1.0.pdf

[D7.1] Hexa-X Deliverable D7.1, "Gap analysis and technical work plan for special-purpose functionality", Online: https://hexa-x.eu/wp-content/uploads/2021/06/Hexa-X-D7.1_v1.0.pdf

[DEL21] https://www2.deloitte.com/xe/en/insights/industry/technology/technology-media-and-telecom-predictions/2021/radio-access-networks.html

[DGK+13] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao , "5G on the Horizon: Key Challenges for the Radio Access Network," IEEE Vehic. Tech. Mag., vol. 8, no. 3, 2013, pp. 47–53.

[DIB21] C. D'Andrea, G. Interdonato and S. Buzzi, "User-centric Handover in mmWave Cell-Free Massive MIMO with User Mobility" 2021 European Signal Processing COnference (EUSIPCO), 23-27 August 2021, Dublin, Ireland

[DKV18] I. Devi, G.R. Karpagam and B. Vinoth Kumar. (2018). "A survey of machine learning techniques", in IJCSE, vol.9, no.4, pp 203-212, doi: 10.1504/IJCSYSE.2017.089191

[DVS20] A. Diamanti, J. M. S. Vilchez, and S. Secci. "LSTM-based radiography foranomaly detection in softwarized infrastructures". In:2020 32nd International Teletraffic Congress (ITC 32). IEEE. 2020, pp. 28–36.

[E16] ETSI, GANA - Generic Autonomic Networking Architecture Reference Model for Autonomic Networking, Cognitive Networking and Self-Management of Networks and Services, October 2016.

[E19] ETSI, Experiential Networked Intelligence (ENI); System Architecture, September 2019.

[EB5G] AI and ML – Enablers for Beyond 5G Networks https://5g-ppp.eu/wp-content/uploads/2021/05/AI-MLforNetworks-v1-0.pdf

[EC99] Official Journal of the European Communities, L 91/10, "Directive 1999/5/EC of the European Parliament and of the Council of 9 March 1999 on Radio Equipment and Telecommunications Terminal Equipment and the Mutual Recognition of Their Conformity," Mar. 1999.

[Edw11] Edwards, Mike, "Service Component Architecture". OASIS. Retrieved 7 April 2011

[EE17] ETSI. (2017, Oct.) Improved operator experience through Experiential Networked Intelligence (ENI). [Online]. Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp22_ENI_FINAL.pdf

[EFA+19] A.C. Eriksson, M. Forsman, H.R. Ainen, P. Willars, and C. Östberg, "5G New Radio RAN and transport choices that minimize TCO ", 7 November 2019, Ericsson Technology Review, https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/5g-nr-ran-and-transport-choices-that-minimize-tco

[EGZ19] ETSI GS ZSM 002, "Zero-touch network and Service Management (ZSM); Reference Architecture", V1.1.1 (2019-08). [Online] Available: https://www.etsi.org/deliver/etsi_gs/ZSM/001_099/002/01.01.01_60/gs_ZSM002v010101p.pdf, Accessed May 2020.

[Ekk20] Ekkono Solutions, Short White Paper, May-20, SWP-openfika7-2005-01 "Federated Learning". Link: https://ekkono.ai/wp-content/uploads/2020/06/SWP_Federated_Learning_Ekkono_Solutions_May_2020.pdf

[EN303146-4] ETSI EN 303 146-4 V1.1.2, "Radio Virtual Machine (RVM)" (2017-04) developed by the European Telecommunications Standards Institute

[ENV] Envoy https://www.envoyproxy.io/

[ENFV04] ETSI NFV plan for NFV Release 4. https://nfvwiki.etsi.org/images/NFVIFA%2820%29000163_NFV_release_4_FEAT17_CNF_management_concepts.pdf

[ESES01] ETSI TSI TR 101 866 V1.1.1 (2001-07), Satellite Earth Stations and Systems (SES); Satellite Component of UMTS/IMT-2000; Analysis and definition of the Packet Mode, https://www.etsi.org/deliver/etsi_tr/101800_101899/101866/01.01.01_60/tr_101866v010101p.pdf

[ETSI10] ETSI. (2010, Jan.) ETSI TR 102 643 – Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services. [Online]: https://www.etsi.org/deliver/etsi_tr/102600_102699/102643/01.00.02_60/tr_102643v010002p.pdf

[ETSI20] ETSI Harmonizing standards for edge computing - A synergized architecture leveraging ETSI ISG MEC and 3GPP specifications, Jun. 2020, Available on https://www.etsi.org/images/files/ETSIWhitePapers/ETSI_wp36_Harmonizing-standards-for-edge-computing.pdf.

[ETSI21] ETSI GR MEC 035 v3.1.1, "Multi-access Edge Computing (MEC); Study on Inter-MEC systems and MEC-Cloud systems coordination", June 2021

[EU14] Official Journal of the EU, L 153/62, "DIRECTIVE 2014/53/ EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 April 2014 on the Harmonisation of the Laws of the Member States Relating to the Making

Available on the Market of Radio Equipment and Repealing Directive 1999/5/EC," May 2014.

[FBR+04] N. Feamster, H. Balakrishnan, J. Rexford, A. Shaikh, and K. van der Merwe, "The case for separating routing from routers." In ACM SIGCOMM Workshop on Future Directions in Network Architecture, Portland, OR, Sept. 2004.

[FDM+12] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Turányi, "Design Aspects of Network Assisted Device-to-Device Communications," IEEE Commun. Mag., vol. 50, no. 3, March 2012, pp. 170–77

[FGS20] F. H. P. Fitzek, F. Granelli, and P. Seeling, "Computing in Communication Networks – From Theory to Practice Book", 1st Ed., Elsevier, 2020, ISBN: 9780128204887.

[FLS+21] F. H.P. Fitzek, S.-C. Li, S. Speidel, T. Strufe, M. Simsek, and M. Reisslein (Ed.), "Tactile Internet with Human-in-the-Loop", Academic Press, 2021.

[GB18] F. Granelli and R. Bassoli, "Autonomic Mobile Virtual Network Operators for Future Generation Networks," in IEEE Network, vol. 32, no. 5, pp. 76-84, September/October 2018, doi: 10.1109/MNET.2018.1700455.

[GBa18] F. Granelli and R. Bassoli, "Towards Autonomic Mobile Network Operators," 2018 IEEE 7th International Conference on Cloud Networking (CloudNet), 2018, pp. 1-4, doi: 10.1109/CloudNet.2018.8549552.

[GBD+18] Z. Guan, L. Bertizzolo, E. Demirors, and T. Melodia, "WNOS: An Optimization-based Wireless Network Operating System", Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, June 2018.

[GDPR] https://gdpr.eu/what-is-gdpr/

[GDPR-PD] https://www.gdpreu.org/the-regulation/key-concepts/personal-data/

[Gha20] Z. Ghadialy, "Understanding the TCO of a Mobile Network", 26 October 2020, The 3G4G Blog, https://blog.3g4g.co.uk/2020/10/understanding-tco-of-mobile-network.html

[Gha20a] Z. Ghadialy, "Samsung Talks about TCO Optimization to Accelerate 5G Network Evolution", 3 December 2020, Telecoms Infrastructure Blog, https://www.telecomsinfrastructure.com/2020/12/samsung-talks-about-tco-optimization-to.html

[GoClo] Google Cloud Functions, https://cloud.google.com/functions

[GoogleAIBlog] https://ai.googleblog.com/

[GS5] GSMA, 5G Implementation Guidelines: SA Option 2, June 2020.

[GSMA18] GSMA. (2018, April) Network Slicing: Use Case Requirements. [Online]. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2020/01/2.0_Network-Slicing-Use-Case-Requirements-1.pdf.

[GSMA14] GSMA. (2014, Dec.) Understanding 5G: Perspectives on future technological advancements in mobile. [Online]. Available:https://www.gsma.com/futurenetworks/wp-content/uploads/2015/01/2014-12-08-c88a32b3c59a11944a9c4e544fee7770.pdf

[GSMA20] "Operator Platform Telco Edge Proposal" White paper, Version 1.0, 22 October 2020

[GSMAEE] Energy Efficiency: An Overview (2019)

[H2020] H2020-ICT-52 AI@Edge, https://aiatedge.eu/

[Hat80] M. Hata “Empirical formula for propagation loss in land mobile radio services”, IEEE Trans. Veh. Tech., vol. 29, no. 3, pp.317 -325 1980

[HAP] HAProxy. http://www.haproxy.org/

[HHA+20] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. Di Renzo, and M. Debbahet, "Holographic MIMO Surfaces for 6G Wireless Networks: Opportunities, Challenges, and Trends," in IEEE Wireless Communications, vol. 27, no. 5, pp. 118-125, October 2020, doi: 10.1109/MWC.001.1900534

[HLE19] High Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, 2019.

[IBN] Intent-Based Networking (IBN) - from https://www.cisco.com/c/en/us/solutions/intent-based-networking.html

[IEEE19] IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems, 2020, IEEEStd 1588-2019 (Revision of IEEE Std 1588-2008).

[ISO] https://standards.iso.org/ittf/PubliclyAvailableStandards/index.html

[IST] Istio. https://istio.io/

[JCD+21] Peter Jonsson, Stephen Carson, Steven Davis, Peter Linder, Per Lindberg, Juan Ramiro, Jose Outes, Amit Bhardwaj, Claudia Muñiz Garcia, Harald Baur, Jake Alger, Todd Krautkremer, Rohit Chandra, Tomas Lundborg, Brahim Belaoucha, Fredrik Burstedt, Courtney Latta, Robert McCrorey Karri Kuoppamaki. Ericsson Mobility Report, June 2021. Online available: https://www.ericsson.com/49f7c7/assets/local/mobility-report/documents/2021/june-2021-ericsson-mobility-report.pdf

[JDN21] Nan Jiang, Yansha Deng, and Arumugam Nallanathan. “Traffic Prediction and Random AccessControl Optimization: Learning and Non-Learning-Based Approaches”. In:IEEE CommunicationsMagazine59.3 (2021), pp. 16–22

[KGC+20] Kaloxylos, Alexandros, Gavras, Anastasius, Camps Mur, Daniel, Ghoraishi, Mir, & Hrasnica, Halid. (2020). AI and ML – Enablers for Beyond 5G Networks. Zenodo. https://doi.org/10.5281/zenodo.4299895

[Knat] Knative, https://knative.dev/

[KKI+21] J. Kaur, M. A. Khan, M. Iftikhar, M. Imran and Q. Emad Ul Haq, "Machine Learning Techniques for 5G and Beyond," in IEEE Access, vol. 9, pp. 23472-23488, 2021, doi: 10.1109/ACCESS.2021.3051557

[KO21] Andreas Krichel, Marie-Paule Odini. “The Challenge of Zero Touch and Explainable AI”. In: Journal of ICT Standardization (2021), pp. 147–158.

[KS3] Lightweight Kubernetes. https://k3s.io

[KTY20] T. Kanaya, N. Tabata and S. Yamaguchi, "A Study on Performance of CUBIC TCP and TCP BBR in 5G Environment," 2020 IEEE 3rd 5G World Forum (5GWF), Bangalore, India, 2020, pp. 508-513, doi: 10.1109/5GWF49715.2020.9221188.

[Kubf] Kubeflow, https://www.kubeflow.org/

[LCS+19] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A.Zhang, “The Roadmap to 6G: AI Empowered Wireless Networks, ”IEEE Communications Magazine, vol. 57, no. 8, pp. 84–90, 2019.

[Lem20] Max Lemke, Next generation of IoT, Opportunities for Europe, https://aioti.eu/wp-content/uploads/2020/09/200929-Next-Generation-IoT-M-Lemke.pdf, September 2020.

[LINK] Linkerd https://linkerd.io/

[LLH+20] W. Y. B. Lim, N. C. Luong, D.T. Hoang et al. “Federated Learning in Mobile Edge Networks: A Comprehensive Survey”, IEEE Communications Surveys and Tutorials, Vol 22, No. 3, 2031-2063), 2020

[LORA] https://lora-alliance.org/

[LSK19]. Euijong Lee, Young-Duk Seo, and Young-Gab Kim. “Self-adaptive framework based on MAPE loopfor Internet of things”. In: sensors19.13 (2019), p. 2996.

[LZL+12] L. Lei, Z. Zhong, C, Lin and X. Shen, “Operator Controlled Device-to-Device Communications in LTE-Advanced Networks,” IEEE Wireless Commun., vol. 19, no. 3, June 2012, pp. 96–104

[MAB+08] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner. OpenFlow: Enabling innovation in campus networks. ACM SIGCOMM Computer Communications Review, Apr. 2008.

[MAS] https://datatracker.ietf.org/wg/masque/about/

[MCS+19] Mavromatis A., Colman-Meixner C., Silva A., Vasilakos X., Nejabati R., Simeonidou D., “A software-defined IoT device management framework for edge and cloud computing”. In IEEE Internet of Things Journal7.3 (2019), pp. 1718–1735

[MFW19] P. J. Mateo, C. Fiandrino and J. Widmer, "Analysis of TCP Performance in 5G mm-Wave Mobile Networks," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1-7, doi: 10.1109/ICC.2019.8761718.

[MII17-D24] METIS-II, Deliverable D2.4, “Final Overall 5G RAN Design”, https://metis-ii.5g-ppp.eu/wp-content/uploads/deliverables/METIS-II_D2.4_V1.0.pdf

[Min] Minikube. https://minikube.sigs.k8s.io/docs/start/

[mk8s] Microk8s . https://microk8s.io/

[MMR+17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas. "Communication-efficient learning of deep networks from decentralized data," in Artificial Intelligence and Statistics, pp. 1273-1282. PMLR, 2017.

[MS21] Malhotra I. and Singh G., “Terahertz Integrated Circuit Design”, Chapter 11 of “Terahertz Antenna Technology for Imaging and Sensing Applications”, May 2021, Springer, https://doi.org/10.1007/978-3-030-68960-5_11

[NGI] NGINX https://nginx.org/en/

[NKE+04] K. Narenthiran, M. Karaliopoulos, B. G. Evans, W. De-Win, M. Dieudonne, P. Henrio, M. Mazzella, E. Angelou, I. Andrikopoulos, P. I. Philippopoulos, D. I. Axiotis, N. Dimitriou, A. Polydoros, G. E. Corazzaand A. Vanelli-Coralli, “S-UMTS access network for broadcast and multicast service delivery: the SATIN approach,” in International Journal of Satellite Communications and Networking, 2004

[NSS+20] G. Nardini, D. Sabella, G. Stea, P. Thakkar, A. Virdis "Simu5G – An OMNeT++ library for end-to-end performance evaluation of 5G networks", IEEE Access, 2020, DOI: 10.1109/ACCESS.2020.3028550

[NSV+20] G. Nardini, G. Stea, A. Virdis, D. Sabella, P. Thakkar, "Using Simu5G as a Realtime Network Emulator to Test MEC Apps in an End-To-End 5G Testbed", PiMRC 2020, London, UK, 1-3 September 2020

[ORAN21] Operator Defined Open and Intelligent Radio Access Networks, https://www.o-ran.org/

[O21] ONF, The Next Generation Architecture of ONOS, https://gonorthforge.com/the-next-generation-architecture-of-onos/

[O6GF] Oulu 6G Flagship. [online] https://www.oulu.fi/6gflagship/

[Ohl21] P. Öhlen. "The future of digital twins: what will they mean for mobile networks?" https://www.ericsson.com/en/blog/2021/7/future-digital-twins-in-mobile-networks, 2021

[PML+19] J. Portilla, G. Mujica, J. -S. Lee and T. Riesgo, "The Extreme Edge at the Bottom of the Internet of Things: A Review," in IEEE Sensors Journal, vol. 19, no. 9, pp. 3179-3190, 1 May1, 2019, doi: 10.1109/JSEN.2019.2891911.

[PSB15] V. S. Pendyala, S. S. Y. Shim, and C. Bussler, "The web that extends beyond the world," Computer, vol. 48, no. 5, pp. 18–25, 2015.

[RAD21] Mobile Statistics Report, 2021-2025. https://www.radicati.com/?p=17218

[RED14] The Radio Equipment Directive, European Commission, https://ec.europa.eu/growth/sectors/electrical-engineering/red-directive_en

[RFC3986] Prayson P., Stewart B., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", Request for Comments, 2005. DOI 10.17487/RFC3985

[RP-211574] Revised WID on support of reduced capability NR devices. 3GPP TSG RAN Meeting #92e. Electronic Meeting, 14th – 18th June 2021

[SAP+21] A. Shahraki, M. Abbasi, M. J. Piran, M. Chen and S. Cui "A Comprehensive Survey on 6G Networks: Applications, Core Services, Enabling Technologies, and Future Challenges.", January 2021 Computer Science - Networking and Internet Architecture. arXiv:2101.12475

[Sau21] Martin Sauter, "From GSM to LTE-Advanced Pro and 5G: An Introduction to Mobile Networks and Mobile Broadband" 4th Edition. Wiley. ISBN-13: 978-1119714675, 2011

[SB] Serverless Benefits, https://www.cloudflare.com/it-it/learning/serverless/what-is-serverless/

[SBC+20] E. Calvanese Strinati, S. Barbarossa, T. Choi, A. Pietrabissa, A. Giuseppi, E. De Santis, J. Vidal, Z. Becvar, T. Haustein, C. Nicolas, F. Costanzo, J. Kim, and I. Kim, "6G in the sky: On-demand intelligence at the edge of 3D networks," ETRI Journal, vol. 10.4218/etrij.2020-0205, 2020.

[SBC20] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," IEEE Network, vol. 34, no. 3, pp. 134–142, 2020.

[SBJ+19] E. Calvanese Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, "6G: The Next Frontier: From Holographic Messaging to Artificial Intelligence Using Subterahertz and Visible Light Communication, "IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 42–202150, 2019.

[SEL] Seldon, https://www.seldon.io/

[SF21] P. Seeling and F. H. P. Fitzek, "Anticipatory Networking: Negative Latency for Ubiquitous Computing," 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 2021, pp. 1-4, doi: 10.1109/CCNC49032.2021.9369624.

[SG21] Ivan Šimunić and Ivan Grgurević. "Automation of Network Device Configuration Using Zero-TouchProvisioning-A Case Study". In: International Conference on Future Access Enablers of Ubiquitous and Intelligent Infrastructures. Springer. 2021, pp. 105–119.

[SSA+20] H. Sarieddeen, N. Saeed, T. Y. Al-Naffouri and M. Alouini, "Next Generation Terahertz Communications: A Rendezvous of Sensing, Imaging, and Localization," in IEEE Communications Magazine, vol. 58, no. 5, pp. 69-75, May 2020

[SUS] https://sdg-tracker.org/sustainable-consumption-production.

[GPM+20] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan and M. Zorzi, "Toward 6G Networks: Use Cases and Technologies," in IEEE Communications Magazine, vol. 58, no. 3, pp. 55-61, March 2020, doi: 10.1109/MCOM.001.1900411.

[TW07] David L. Tennenhouse and David J. Wetherall. 2007. "Towards an active network architecture." SIGCOMM Comput. Commun.

[URLL] https://www.ngmn.org/wp-content/uploads/200210-Verticals-URLLC-Requirements-v2.5.4.pdf

[VFS12] K. Vanganuru, S. Ferrante, and G. Sternberg, "System Capacity and Coverage of a Cellular Network with D2D mobile Relays," MILCOM 2012, Orlando, FL, Oct. 2012

[VM20] H. Viswanathan and P. E. Mogensen, "Communications in the 6G Era," IEEE Access, vol. 8, pp. 57063-57074, Mar. 2020.

[WG20] A. Willner and V. Gowtham, "Toward a Reference Architecture Model for Industrial Edge Computing," in IEEE Communications Standards Magazine, vol. 4, no. 4, pp. 42-48, December 2020, doi: 10.1109/MCOMSTD.001.2000007.

[WIDIR] https://www.wi-fi.org/discover-wi-fi/wi-fi-direct

[WIMES] https://www.wi-fi.org/discover-wi-fi/wi-fi-easymesh

[YAG04] L. Yang, R. Dantu, T. Anderson, and R. Gopal. Forwarding and Control Element Separation (ForCES) Framework. Internet Engineering Task Force, Apr. 2004. RFC 3746.

[YLC+19] Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., & Yu, H. (2019). Federated Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 13(3), 1-207. https://doi.org/10.2200/S00960ED2V01Y201910AIM043.

[YZZ+20] Y. Yuan, Y. Zhao, B. Zong, S. Parolari, "Potential Key Technologies for 6G Mobile Communications", Science China Information Sciences, May 2020, vol. 63, issue 8, https://doi.org/10.1007/s11432-019-2789-y

[ZFW+19] B. Zong, C. Fan, X. Wang, X. Duan, B. Wang and J. Wang, "6G Technologies: Key Drivers, Core Requirements, System Architectures, and Enabling Technologies," in IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 18-27, Sept. 2019, doi: 10.1109/MVT.2019.2921398.

[ZMF+16] M. Zhang, M. Mezzavilla, R. Ford, S. Rangan, S. Panwar, E. Mellios, D. Kong, A. Nix, and M. Zorzi, "Transport layer performance in 5G mmWave cellular," 2016 IEEE Conference on Computer Communications Workshops (INFOCOM

WKSHPS), San Francisco, CA, USA, 2016, pp. 730-735, doi: 10.1109/INFCOMW.2016.7562173.

[ZSP+13] A- Zimmermann, K. Sandkuhl, M. Pretz, et. Al., "Towards an integrated service-oriented reference enterprise architecture". in Proceedings of the 2013 International Workshop on Ecosystem Architectures. ACM, New York, NY, USA, 26-30. DOI=10.1145/2501585.2501591 http://doi.acm.org/10.1145/2501585.2501591

[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 Magzine, vol. 14, no. 3, pp. 28–41, 2019.

Annex A: Additional information

A.1 Terminology

Table A-1 Terminology used in D5.1

Term	Abbreviations	Term description	Ref. deliverable
Service Based Architecture	SBA	A modular architecture introduced for 5G for the first time in which the control plane functionality and common data repositories of a 5G network are delivered by way of a set of interconnected Network Functions (NFs), each with authorization to access each other's services.	D5.1
Access and Mobility management Function	AMF	A Core Network function/node that handles user's access and mobility.	D5.1
Dynamic Function Placement	DPF	DPF the act of dynamically place network functions. This is done by deploying intelligent algorithms to orchestrate differentiated services optimally across multiple sites and clouds, based on diverse intents and policy constraints of dynamically changing environments.	D5.1
Flexibility to different topologies	Not Applicable (N/A)	The ability of the network to adapt to various scenarios such as new non-public networks, autonomous networks, mesh networks, new spectrum, etc., without loss of performance and easy deployment. Addition of service capabilities and new services endpoints require no changes to existing end-to-end services.	D5.1
Network of networks	N/A	Defined as a network that can both incorporate different network solutions as well as a network that easily (flexibly) can adapt to new topologies (same thing as Flexibility to different topologies also)	D5.1
Network Service Meshes	N/A	Network service mesh is intended to support application-to-application and function-to-function communications in 6G networks and scenarios through dynamic and automated virtual network services, to be allocated on-demand, based on application requirements (similar to DPF).	D5.1
Full Network Automation	N/A	Full Network Automation is driven by high-level policies and rules without minimal human intervention. Networks will be capable of self-configuration, self-monitoring, self-healing, and self-optimisation	D5.1
Non-Terrestrial Network	NTN	Satellites and other flying objects such as HAPS and UAVs.	D5.1
Programmability	N/A	UE and network programmability, a framework that gives the possibility to update the program for specific features in a network entity	D5.1
Scalability	N/A	The network architecture needs to be scalable both in terms of supporting very small to very large-scale deployments, by scaling up and down network resources based on needs, e.g., varying traffic, utilizing underlying shared cloud platform	D5.1

Resilience and availability	N/A	This means that the network (architecture) shall be resilient in terms of service and infrastructure provisioning using multi connectivity, and separation of CP and UP, support of local network survivability if a subnetwork loses connectivity with another network, removing single point of failures	D5.1
Dependability	N/A	Dependability is the "ability to perform as and when required". Dependability consists of the attributes: availability, reliability, safety, integrity, and maintainability [Avi04]. End-to-end dependability refers to dependability from the application perspective, encompassing multiple services (c.f. Productivity)	D7.1
Reliability	N/A	Reliability is the probability to perform as required for a given time interval, under given conditions	D7.1

A.2 KPIs

During the design of the architecture of a system, its targets are the fundamental drivers. Specifically, services and use cases to be hosted define the performance and quality indicators to state the metrics for the design, implementation, and operations. By definition, KPIs are "[...] items of information collected at regular intervals to track the performance of a system. [...]" [FiG90]. The following Table A-2 summarises and compares the technology driven KPIs targeted within 5G [GSMA14, ARS16] and the ones envisioned in 6G [LCS+19], [ZXM+19].

Table A-2 Possible 6G KPIs, based on the 5G KPIs

5G	6G
Throughput/data rate up to 1−10 Gbit/s	Throughput/data rate up to 1 Tbit/s; user-experienced data rate of 1 Gbit/s (ten times the one targeted by 5G).
End-to-end latency down to 1−10 ms; Mobility up to 350 km/h.	End-to-end latency less than 1 ms; An 'over-the-air' latency of 10−100 µs with mobility up to 1000 km/h.
1000 times increase in bandwidth per unit of area, compared to the previous generation.	Very broad bandwidth with frequencies reaching 1−3 THz.
99.999% perceived availability and 100% terrestrial geographical coverage.	"Always-ON" terrestrial-aerial-satellite network.
Frame error rate equal to 10^{-5}	Frame error rate (reliability) equal to 10−9
90% reduction in network energy usage, compared to previous generation.	Supporting" battery-free IoT devices" (10-100 times the one of 5G) and energy efficiency down to 1 pJ/bit
Localisation precision equal to 10 cm in two dimensions.	Localisation precision equal to 1 cm in three dimensions.
Spectrum efficiency three-five times greater than the one of 4G	Spectrum efficiency greater than three times the one of 5G
Density of connected devices 10^6 km^{-2}	Density of connected devices greater than 10^6 km^2.Connectivity density ten times the one provided by 5G, with an area traffic capacity of up to 1 Gbit/s/m^3 (10 Gbit/s in 3D).
Receiver sensitivity about−120 dBm	Receiver sensitivity less than −130 dBm

Next to KPIs, another important metric of communication networks is *quality*. In the literature, three main quality metrics were defined: Quality of Service (QoS), Quality-of-Perception (QoP)

- also user-perceived QoS – and Quality-of-Experience (QoE). *Network QoS* is the "[...] degree of conformance of the service delivered to a user by a provider with an agreement between them [...]" [ETSI10]. Next, the QoS is the "[...] totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service [...]" [ETSI10]. The QoS relies on technical metrics so it is technology centred. The QoP "[...] is primarily concerned with the detectability of a change in quality or the acceptability of a quality level. [...]" [ETSI10]. As an example, the use of Mean Opinion Score (MOS) measures the perception of quality according to a subjective rating. Next, QoE is a "[...] measure of user performance based on both objective and subjective psychological measures of using an ICT service or product. [...]" [ETSI10]. For this definition, two notes are mentioned: "[...] It considers technical parameters (e.g., QoS) and usage context variables (e.g., communication task) and measures both the process and outcomes of communication (e.g., user effectiveness, efficiency, satisfaction and enjoyment). [...]" [ETSI10]; the second states "[...] The appropriate psychological measures will be dependent on the communication context. Objective psychological measures do not rely on the opinion of the user (e.g., task completion time measured in seconds, task accuracy measured in number of errors). Subjective psychological measures are based on the opinion of the user (e.g., perceived quality of medium, satisfaction with a service). [...]" [ETSI10]. Next, the second definition states that the QoE is the "[...] overall acceptability of an application or service, as perceived subjectively by the end-user [...]" [ETSI10].

Additionally, other definitions of concept of quality have been introduced in the context of 6G networks. The concept of Quality-of-Physical-Experience (QoPE) [SBC20] is an attempt to complete and to unify the evaluation separately given by QoS and QoE, by combining them with other physical aspects of humans such as brain cognition, body characteristics, and gestures.

A.3 Service-centric functional model for 6G system architecture proposed by Oulu 6G Flagship

As an early actor in 6G research the Oulu 6G Flagship [O6GF] has pondered 6G system architecture aspects as well. Here is a short, high-level overview of the current modelling approach. The adopted service-centric functional model enables one to approach the 6G system architecture by defining the services that the 6G system offers to realize the envisioned new use-cases. It provides only the functional description of the services and their components, leaving more freedom for the design of a logical network architecture at a later phase.

The 6G system is expected to serve both human and machine *users*. A user could also consist of a group of people or machines, or a mix thereof. Regardless of the user type, a certain *application* is involved using 6G system services for its implementation in a potentially challenging (high-speed, dynamic, etc.) operation environment. The 6G system domain model is shown in Figure A-1.

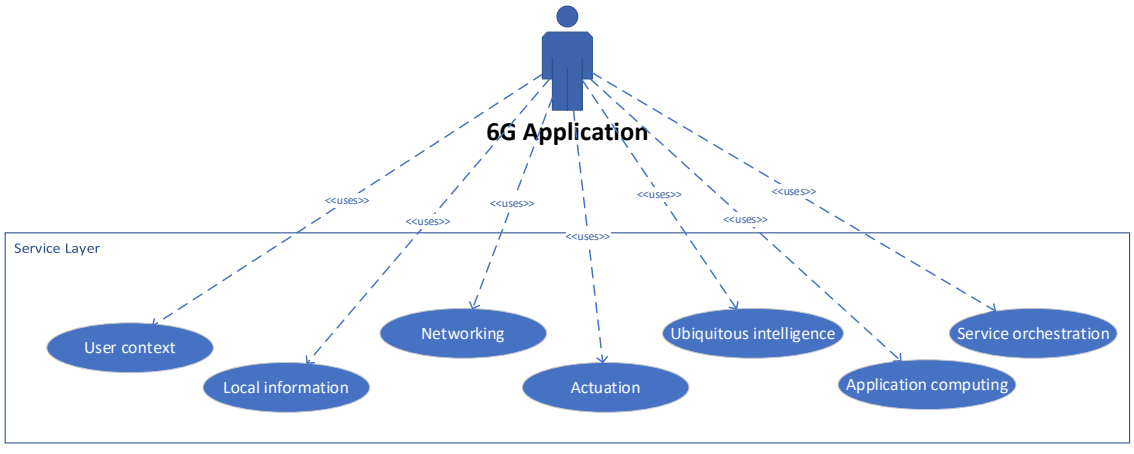

Figure A-1. Oulu 6G Flagship 6G system domain model.

The uppermost layer of the 6G system functional model is named *Service Layer*. It provides a set of services to fulfil the multi-faceted needs of challenging new 6G applications of 6G system users, such as eXtended Reality (XR), holographic telepresence [VM20], digital twins, collaborative mobile robots, drone control, etc. A combination of services is usually needed to fulfil the 6G user and application needs.

The 6G system service set consists of six main services and one additional supporting service:

- ***Networking*** service: provides basic connectivity functionality to the 6G application (as seen in Figure A-1) and ensures service continuity during the session.
- ***User context*** service: manages all relevant information that is needed to characterize the user situation, especially the user position and contextual data management.
- ***Local information*** service: provides an information bus and means to utilize sensing and radar capabilities of 6G wireless system to create 3D maps of the environment and characterize the environment conditions.
- ***Actuation*** service: provides means to control actuators and initiate changes in the physical environment. This mechanism can be used to change the internal states of the physical entities as well.
- ***Ubiquitous intelligence*** service: facilitates information delivery, processing and reasoning (learning and inference) for applications that operate in the ubiquitous environment.
- ***Application computing*** service: provides support for application-related data processing from 6G system side, supporting application mobility.
- ***Service orchestration*** service: takes care of selection, use and information sharing among other 6G system services to provide holistic service combination to satisfy the 6G user and application needs.

Each service in the *Service Layer* can be further divided into service components, i.e., functional components, to obtain a more detailed functional model of the 6G system services. Each service can utilize its own service components, as well as service components owned by another service, as illustrated in Figure A-2 via <<includes>> or <<uses>> relations.

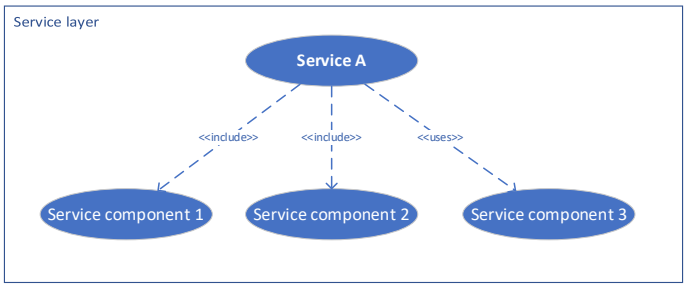

Figure A-2. Oulu 6G Flagship Conceptual view of a service and service components.

In addition to the *Service Layer,* the 6G service-centric functional model contains functions placed in four other layers, i.e.:

Network Layer: provides the necessary functions to accomplish data transmission requested by the Service Layer.

Resource Access Layer: provides functions to select and manage resources (e.g., transmission, reception, spectrum) to implement higher-layer functionalities.

Algorithm Layer and *Hardware (HW) Layer:* provide the actual lower-layer resources, i.e., functions, such as MAC and PHY layer algorithms and antenna and RF device type hardware.